nervous system and also modulate both motor and secretory functions. The parasympathetic nerve supply conveys both visceral sensory as well as excitatory pathways to the motor components of the colon. Parasympathetic fibers via the vagus nerve reach the small intestine and proximal colon along the branches of the superior mesenteric artery. The distal colon is supplied by sacral parasympathetic nerves (S_{2-4}) via the pelvic plexus; these fibers course through the wall of the colon as ascending intracolonic fibers as far as, and in some instances including, the proximal colon. The chief excitatory neurotransmitters controlling motor function are acetylcholine and the tachykinins, such as substance P. The sympathetic nerve supply modulates motor functions and reaches the small intestine and colon alongside the arterial arcades of the superior and inferior mesenteric vessels. Sympathetic input to the gut is generally excitatory to sphincters and inhibitory to nonsphincteric muscle. Visceral afferents convey sensation from the gut to the central nervous system; initially, they course along sympathetic fibers, but as they approach the spinal cord they separate, have cell bodies in the dorsal root ganglion, and enter the dorsal horn of the spinal cord. Afferent signals are conveyed to the brain along the lateral spinothalamic tract and the nociceptive dorsal column pathway and are then perceived. Other afferent fibers synapse in the prevertebral ganglia and reflexly modulate intestinal motility.

INTESTINAL FLUID ABSORPTION AND SECRETION

On an average day, 9 L of fluid enters the gastrointestinal tract; approximately 1 L of residual fluid reaches the colon; the stool excretion of fluid constitutes about 0.2 L/d. The colon has a large capacitance and functional reserve and may recover up to four times its usual volume of 0.8 L/d, provided the rate of flow permits reabsorption to occur. Thus, the colon can partially compensate for intestinal absorptive or secretory disorders.

In the colon, sodium absorption is predominantly electrogenic, and uptake takes place at the apical membrane; it is also compensated by the pumping out functions of the basolateral sodium pump. A variety of neural and non-neural mediators regulate colonic fluid and electrolyte balance, including cholinergic, adrenergic and serotonergic mediators. Angiotensin and aldosterone also influence colonic absorption, reflecting the common embryologic development of the distal colonic epithelium and the renal tubules.

ILEOCOLONIC STORAGE AND SALVAGE

The distal ileum acts as a reservoir, emptying intermittently by bolus movements. This action allows time for salvage of fluids, electrolytes, and nutrients. Segmentation by haustra compartmentalizes the colon and facilitates mixing, retention of residue, and formation of solid stools. In health, the ascending and transverse regions of colon function as reservoirs (average transit, 15 h), and the descending colon acts as a conduit (average transit, 3 h). The colon is efficient at conserving sodium and water, a function that is particularly important in sodium-depleted patients in whom the small intestine alone is unable to maintain sodium balance. Diarrhea or constipation may result from alteration in the reservoir function of the proximal colon, or the propulsive function of the left colon. Constipation may also result from disturbances of the rectal or sigmoid reservoir, typically as a result of dysfunction of the pelvic floor or the coordination of defecation.

SMALL INTESTINAL MOTILITY

During fasting, the motility of the small intestine is characterized by a cyclical event called the migrating motor complex (MMC), which serves to clear nondigestible residue from the small intestine. This organized, propagated series of contractions lasts on average 4 min, occurs every 60 to 90 min, and usually involves the entire small intestine. After food ingestion, the small intestine produces irregular, mixing contractions of relatively low amplitude, except in the distal ileum where more powerful contractions occur intermittently and empty the ileum by bolus transfers.

COLONIC MOTILITY AND TONE

The small intestinalMMC only rarely continues into the colon. However, short duration or phasic contractions mix colonic contents, and high amplitude propagated contractions (HAPCs) are sometimes associated with mass movements through the colon and occur approximately five times per day, usually on awakening in the morning and postprandially. Increased frequency of HAPCs may result in diarrhea. The predominant phasic contractions are irregular and nonpropagated and serve as a "mixing" function.

Colonic tone refers to the background contractility upon which phasic contractile activity (typically contractions lasting less than 15 s) is superimposed. It is an important cofactor in the colon's capacitance (volume accommodation) and sensation.

COLONIC MOTILITY AFTER MEAL INGESTION

After meal ingestion, colonic phasic and tonic contractility increase for a period of approximately 2 h. The initial phase (about 10 min) is mediated by the vagus nerve in response to mechanical distention of the stomach. The subsequent response of the colon requires caloric stimulation and is at least in part mediated by hormones, e.g., gastrin and serotonin.

DEFECATION

Tonic contraction of the puborectalis muscle, which forms a sling around the rectoanal junction, is important to maintain continence; during defecation, sacral parasympathetic nerves relax this muscle, facilitating the straightening of the rectoanal angle (Fig. 42-1). Distention of the rectum results in transient relaxation of the internal anal sphincter via intrinsic and reflex sympathetic innervation. As sigmoid and rectal contractions increase the pressure within the rectum, the rectosigmoid angle opens by more than 15°. Voluntary relaxation of the external anal sphincter (striated muscle innervated by the pudendal nerve) permits the evacuation of feces; this evacuation process can be augmented by an increase in intraabdominal pressure created by the Valsalva maneuver.

DIARRHEA

DEFINITION

Diarrhea is loosely defined as passage of abnormally liquid or unformed stools at an increased frequency. For adults on a typical Western diet, stool weight exceeding 200 g/d can generally be considered diarrheal. Because of the fundamental importance of duration to diagnostic considerations, diarrhea may be further defined as *acute* if <2 weeks, *persistent* if 2 to 4 weeks, and *chronic* if>4 weeks in duration.

Two common conditions, usually associated with the passage of stool totaling <200 g/d, must be distinguished from diarrhea, as diagnostic and therapeutic algorithms differ. *Pseudodiarrhea*, or the frequent passage of small volumes of stool, often is associated with rectal urgency and accompanies the irritable bowel syndrome or anorectal disorders like proctitis. *Fecal incontinence* is the involuntary discharge of rectal contents and is most often caused by neuromuscular disorders or structural anorectal problems. Diarrhea and urgency, especially if severe, may aggravate or cause incontinence. Pseudodiarrhea and fecal incontinence occur at prevalence rates comparable to or higher than that of chronic diarrhea and should always be considered in patients complaining of "diarrhea." A careful history and physical examination generally allow these conditions to be discriminated from true diarrhea.

ACUTE DIARRHEA

More than 90% of cases of acute diarrhea are caused by infectious agents; these cases are often accompanied by vomiting, fever, and abdominal pain. The remaining 10% or so are caused by medications, toxic ingestions, ischemia, and other conditions.

Infectious Agents Most infectious diarrheas are acquired by fecal-oral transmission via direct personal contact or, more commonly, via ingestion of food or water contaminated with pathogens from human or animal feces. In the immunologically competent person, the resident fecal microflora, containing more than 500 taxonomically distinct species, are rarely the source of diarrhea and may actually play a role in suppressing the growth of ingested pathogens. Acute infection or injury occurs when the ingested agent overwhelms the host's mucosal immune and nonimmune (gastric acid, digestive enzymes, mucus secretion, peristalsis, and suppressive resident flora) defenses. Established clinical associations with specific enteropathogens may offer diagnostic clues.

In the United States, high risk groups are recognized:

1. *Travelers*. Nearly 40% of tourists to endemic regions of Latin America, Africa, and Asia develop so-called traveler's diarrhea, most commonly due to enterotoxigenic *Escherichia coli* as well as to *Campylobacter*, *Shigella*, and *Salmonella*. Visitors to Russia (especially St. Petersburg) may have increased risk of *Giardia*-associated diarrhea; visitors to Nepal may acquire *Cyclospora*. Campers, backpackers, and swimmers in wilderness areas may become infected with *Giardia*.

2. *Consumers of certain foods*. Diarrhea closely following food consumption at a picnic, banquet, or restaurant may suggest infection with *Salmonella*, *Campylobacter*, or *Shigella* from chicken; enterohemorrhagic *E. coli* (O157:H7) from undercooked hamburger; *Bacillus aureus* from fried rice; *Staphylococcus aureus* or *Salmonella* from mayonnaise or creams; *Salmonella* from eggs; and *Vibrio* species, *Salmonella*, or acute hepatitis A or B from seafood, especially if raw.

3. *Immunodeficient persons*. Individuals at risk for diarrhea include those with either primary immunodeficiency (e.g., IgA deficiency, common variable hypogammaglobulinemia, chronic granulomatous disease) or the much more common secondary immunodeficiency states (e.g., AIDS, senescence, pharmacologic suppression). Common enteropathogens often cause a more severe and protracted diarrheal illness; and, particularly in persons with AIDS, opportunistic infections, such as by *Mycobacterium* species, certain viruses (cytomegalovirus, adenovirus, and herpes simplex), and protozoa (*Cryptosporidium*, *Isospora belli*, Microsporidia, and *Blastocystis hominis*) may also play a role (Chap. 309). In patients with AIDS, agents transmitted venereally per rectum (e.g., *Neisseria gonorrhoeae*, *Treponema pallidum*, *Chlamydia*) may contribute to proctocolitis.

4. *Daycare participants and their family members*. Infections with *Shigella*, *Giardia*, *Cryptosporidium*, rotavirus, and other agents are very common and should be considered.

5. *Institutionalized persons*. Infectious diarrhea is one of the most frequent categories of nosocomial infections in many hospitals and long-term care facilities; the causes are a variety of microorganisms but most commonly *Clostridium difficile*.

The pathophysiology underlying acute diarrhea by infectious agents produces specific clinical features that may also be helpful in diagnosis (Table 42-2). Profuse watery diarrhea secondary to small bowel hypersecretion occurs with ingestion of preformed bacterial toxins, enterotoxin-producing bacteria, and enteroadherent pathogens. Diarrhea associated with marked vomiting and minimal or no fever may occur abruptly within a few hours after ingestion of the former two types; vomiting is usually less, and abdominal cramping or bloating is greater; fever is higher with the latter. Cytotoxin-producing and invasive microorganisms all cause high fever and abdominal pain. Invasive bacteria and *Entamoeba histolytica* often cause bloody diarrhea (referred to as *dysentery*). *Yersinia* invades the terminal ileal and proximal colon mucosa and may cause especially severe abdominal pain with tenderness mimicking acute appendicitis.

Finally, infectious diarrhea may be associated with systemic manifestations. Reiter's syndrome (arthritis, urethritis, and conjunctivitis) may accompany or follow infections by *Salmonella*, *Campylobacter*, *Shigella*, and *Yersinia*. Yersiniosis may also lead to an autoimmune-type thyroiditis, pericarditis, and glomerulonephritis. Both enterohemorrhagic *E. coli* (O157:H7) and *Shigella* can lead to the *hemolytic-uremic syndrome* with an attendant high mortality rate. Acute diarrhea can also be a major symptom of several systemic infections including *viral hepatitis*, *listeriosis*, *legionellosis*, and *toxic shock syndrome*.

Other Causes Side effects from medications are probably the most common noninfectious cause of acute diarrhea, and etiology may be suggested by a temporal association between use and symptom onset. Although innumerable medications may produce diarrhea, some of the more frequently incriminated include antibiotics, cardiac antidysrhythmics, antihypertensives, nonsteroidal anti-inflammatory drugs, certain antidepressants, chemotherapeutic agents, bronchodilators, antacids, and laxatives.

Occlusive or nonocclusive *ischemic collitis* typically occurs in persons older than 50 years of age, often presents as acute lower abdominal pain preceding watery, then bloody diarrhea, and generally results in acute inflammatory changes in the sigmoid or left colon while sparing the rectum. Acute diarrhea may accompany colonic *diverticulitis* and *graft-versus-host disease*. Acute diarrhea, often associated with systemic compromise, can follow ingestion of toxins including organophosphate insecticides, amanita and other mushrooms, arsenic, and preformed environmental toxins in seafoods, like ciguatera and scombroid. The conditions causing chronic diarrhea can also be confused with acute diarrhea early in their course. This confusion may occur with inflammatory bowel disease and some of the other inflammatory chronic diarrheas that may have an abrupt rather than insidious onset and exhibit features that mimic infection.

Approach to the Patient

The decision to evalute acute diarrhea depends on its severity and duration and on various host factors (Fig. 42-2). Most episodes of acute diarrhea are mild and self-limited, and they do not justify the cost and potential morbidity of diagnostic or pharmacologic interventions. Indications for evaluation include profuse diarrhea with dehydration, grossly bloody stools, fever³38.5° C, duration>48 h without improvement, new community outbreaks, associated severe abdominal pain in patients older than 50 years of age, and elderly (³70 years) or immunocompromised patients. In some patients with moderately severe febrile diarrhea with fecal leukocytes (or increased fecal levels of the leukocyte proteins lactoferrin or calprotectin) present or with dysentery, a diagnostic evaluation might be eschewed in favor of an empiric antibiotic trial (see below).

The cornerstone of diagnosis in those suspected of severe acute infectious diarrhea is microbiologic analysis of the stool. Workup includes cultures for bacterial and viral pathogens, direct inspection for ova and parasites, and immunoassays for certain bacterial toxins (*C. difficile*), viral antigens (rotavirus), and protozoal antigens (*Giardia*, *E. histolytica*). The aforementioned clinical and epidemiologic associations may assist in focusing the evaluation. If a particular pathogen or set of possible pathogens is so implicated, then either the whole panel of routine studies may not be necessary or, in some instances, special cultures may be appropriate as for enterohemorrhagic and other types of *E. coli*, *Vibrio* species, and *Yersinia*. Molecular diagnosis of pathogens in stool can be made by identification of unique DNA sequences; and evolving microarray technologies could lead to a more rapid, sensitive, specific, and cost-effective diagnostic approach in the future.

Persistent diarrhea is commonly due to *Giardia*, but additional causative organisms that should be considered include *C. difficile* (especially if antibiotics had been administered), *E. histolytica*, *Cryptosporidium*, *Campylobacter*, and others. If stool studies are unrevealing, then flexible sigmoidoscopy with biopsies and upper endoscopy with duodenal aspirates and biopsies may be indicated.

Structural examination by sigmoidoscopy, colonoscopy, or abdominal CT scanning (or other imaging approaches) may be appropriate in patients with uncharacterized persistent diarrhea to exclude inflammatory bowel disease, or as an initial approach in patients with suspected noninfectious acute diarrhea such as might be caused by ischemic colitis, diverticulitis, or partial bowel obstruction.

TREATMENT

Fluid and electrolyte replacement are of central importance to all forms of acute diarrhea. Fluid replacement alone may suffice for mild cases. Oral sugar-electrolyte solutions (sport drinks or designed formulations) should be instituted promptly with severe diarrhea to limit dehydration, which is the major cause of death. Profoundly dehydrated patients, especially infants and the elderly, require intravenous rehydration.

In moderately severe nonfebrile and nonbloody diarrhea, antimotility antisecretory agents like loperamide can be useful adjuncts to control symptoms. Such agents should be avoided with febrile dysentery, which may be exacerbated or prolonged by them. Bismuth subsalicylate may reduce symptoms of vomiting and diarrhea but should not be used to treat immunocompromised patients because of the risk of bismuth encephalopathy.

Judicious use of antibiotics is appropriate in selected instances of acute diarrhea and may reduce its severity and duration (Fig. 42-2). Many physicians treat moderately to severely ill patients with febrile dysentery empirically without diagnostic evaluation using a quinolone, such as ciprofloxacin (500 mg bid for 3 to 5 d). Empiric treatment can also be considered for suspected giardiasis with metronidazole (250 mg qid for 7d). Selection of antibiotics and dosage regimens is otherwise dictated by specific pathogens and conditions found (Chaps. 131,153,156-162). Antibiotic coverage is indicated whether or not a causative organism is discovered in patients that are immunocompromised, have mechanical heart valves or recent vascular grafts, or are elderly. Antibiotic prophylaxis is indicated for certain patients traveling to high-risk countries in whom the likelihood or seriousness of acquired diarrhea would be especially high, including those with immunocompromise, inflammatory bowel disease, or gastric achlorhydria. Use of trimethoprim/sulfamethoxazole or ciprofloxacin may reduce bacterial diarrhea in such travelers by 90%.

CHRONIC DIARRHEA

Diarrhea lasting more than 4 weeks warrants evaluation to exclude serious underlying pathology. In contrast to acute diarrhea, most of the many causes of chronic diarrhea are noninfectious. The classification of chronic diarrhea by pathophysiologic mechanism facilitates a rational approach to management (Table 42-3).

Secretory Causes Secretory diarrheas are due to derangements in fluid and electrolyte transport across the enterocolic mucosa. They are characterized clinically by watery, large-volume fecal outputs that are typically painless and persist with fasting. Because there is no malabsorbed solute, stool osmolality is accounted for by normal endogenous electrolytes with no fecal osmotic gap.

Medications Side effects from regular ingestion of drugs and toxins are the most common secretory causes of chronic diarrhea. Hundreds of prescription and over-the-counter medications (see "Other Causes of Acute Diarrhea," above) may produce unwanted diarrhea. Surreptitious or habitual use of stimulant laxatives [e.g., senna, cascara, bisacodyl, ricinoleic acid (castor oil)] must also be considered. Chronic ethanol consumption may cause a secretory-type diarrhea due to enterocyte injury with impaired sodium and water absorption as well as to rapid transit and other alterations. Inadvertent ingestion of certain environmental toxins (e.g., arsenic) may lead to chronic rather than acute forms of diarrhea. Certain bacterial infections may occasionally persist and be associated with a secretory-type diarrhea.

Bowel resection, mucosal disease, or enterocolic fistula These conditions may result in a secretory-type diarrhea because of inadequate surface for resorption of secreted fluids and electrolytes. Unlike other secretory diarrheas, this subset of conditions tends to worsen with eating. With disease (e.g., Crohn's ileitis) or resection of <100 cm of terminal ileum, dihydroxy bile acids may escape absorption and stimulate colonic secretion (cholorrheic diarrhea). This mechanism may contribute to so-called *idiopathic secretory diarrhea*, in which bile acids are functionally malabsorbed from a normal-appearing terminal ileum. Partial bowel obstruction, ostomy stricture, or fecal impaction may paradoxically lead to increased fecal output due to hypersecretion.

Hormones Although uncommon, the classic examples of secretory diarrhea are those mediated by hormones. *Metastatic gastrointestinal carcinoid tumors* or, rarely, *primary bronchial carcinoids* may produce watery diarrhea alone or as part of the carcinoid syndrome that comprises episodic flushing, wheezing, dyspnea, and right-sided valvular heart disease. Diarrhea is due to the release into the circulation of potent intestinal secretagogues including serotonin, histamine, prostaglandins, and various kinins. Pellagra-like skin lesions may rarely occur as the result of serotonin overproduction with niacin depletion. *Gastrinoma*, one of the most common neuroendocrine tumors, most typically presents with refractory peptic ulcers, but diarrhea occurs in up to one-third of cases and may be the only clinical manifestation in 10%. While various secretagogues released with gastrin may play a role, the diarrhea most often results from fat maldigestion owing to pancreatic enzyme inactivation by low intraduodenal pH. The watery diarrhea hypokalemia achlorhydria (WDHA) syndrome, also called *pancreatic cholera*, is due to a non-b cell pancreatic adenoma, referred to as a VIPoma, that secretes vasoactive intestinal peptide (VIP) and a host of other peptide hormones including pancreatic polypeptide, secretin, gastrin, gastrin-inhibitory polypeptide, neurotensin, calcitonin, and prostaglandins. The secretory diarrhea is often massive with stool volumes>3 L/d; daily volumes as high as 20 L have been reported. Life-threatening dehydration, neuromuscular dysfunction from associated hypokalemia, hypomagnesemia, or hypercalcemia, flushing, and hyperglycemia may accompany vipoma. *Medullary carcinoma of the thyroid* may present with watery diarrhea caused by calcitonin, other secretory peptides, or prostaglandins. This tumor occurs sporadically or, in 25 to 50% of cases, as a feature of multiple endocrine neoplasia type IIa with pheochromocytomas and hyperparathyroidism. Prominent diarrhea is often associated with metastatic disease and poor prognosis. *Systemic mastocytosis*, which may be associated with the skin lesion urticaria pigmentosa, may cause diarrhea that is either secretory and mediated by histamine, or inflammatory and due to intestinal filtration by mast cells. Large *colorectal villous adenomas* may rarely be associated with a secretory diarrhea that may cause hypokalemia, can be inhibited by NSAIDs, and is apparently mediated by prostaglandins.

Congenital defects in ion absorption Rarely, these defects cause watery diarrhea from birth and include defective Cl^-/HCO_3^-exchange (*congenital chloridorrhea*) with alkalosis and defective Na^+/H^+exchange with acidosis. Some hormone deficiencies may be associated with watery diarrhea, such as occurs with adrenocortical insufficiency (Addison's disease) that may be accompanied by hyperpigmentation.

Osmotic Causes Osmotic diarrhea occurs when ingested, poorly absorbable, osmotically active solutes draw enough fluid lumenward to exceed the resorptive capacity of the colon. Fecal water output increases in proportion to such a solute load. Osmotic diarrhea characteristically ceases with fasting or with discontinued oral intake of the offending agent.

Osmotic laxatives Ingestion of magnesium-containing antacids, health supplements, or laxatives may induce osmotic diarrhea typified by a stool osmotic gap: 2([Na] + [K]) <<290 mosm/kg. Anionic laxatives containing sulfates or phosphates produce osmotic diarrhea without an osmotic gap, as sodium accompanies the anionic solutes; direct measurement of stool sulfates and phosphates may be necessary to confirm the cause of diarrhea.

Carbohydrate malabsorption Carbohydrate malabsorption due to acquired or congenital defects in brush-border disaccharidases and other enzymes leads to osmotic diarrhea with a low pH. One of the most common causes of chronic diarrhea in adults is *lactase deficiency*, which affects three-fourths of non-Caucasians worldwide and 5 to 30% of persons in the United States; most learn to avoid milk products without an intervention. Some sugars, such as sorbitol, are universally malabsorbed, and diarrhea ensues with ingestion of ample medications, gum, or candies sweetened with these nonabsorbable sugars. Lactulose, used to acidify stools in patients with hepatic failure, also causes diarrhea on this basis.

Steatorrheal Causes Fat malabsorption may lead to greasy, foul-smelling, difficult-to-flush diarrhea often associated with weight loss and nutritional deficiencies due to concomitant malabsorption of amino acids and vitamins. Increased fecal output is caused by the osmotic effects of fatty acids, especially after bacterial hydroxylation, and, to a lesser extent, by the burden of neutral fat. Quantitatively, steatorrhea is defined as stool fat exceeding the normal 7 g/d; daily fecal fat averages 15 to 25 g with small intestinal diseases and often exceeds 40 g with pancreatic exocrine insufficiency. Intraluminal maldigestion, mucosal malabsorption, or lymphatic obstruction may produce steatorrhea.

Intraluminal maldigestion This condition most commonly results from pancreatic exocrine insufficiency, which occurs when >90% of pancreatic secretory function is lost. *Chronic pancreatitis*, usually a sequela of ethanol abuse, most frequently causes pancreatic insufficiency. Other causes include *cystic fibrosis*, *pancreatic duct obstruction*, and rarely, *somatostatinoma*. Bacterial overgrowth in the small intestine may deconjugate bile acids and alter micelle formation that impair fat digestion; it occurs with stasis from a blind-loop, small bowel diverticulum, or dysmotility and is especially likely in the elderly. Finally, cirrhosis or biliary obstruction may lead to mild steatorrhea due to deficient intraluminal bile acid concentration.

Mucosal Malabsorption Mucosal malabsorption occurs from a variety of enteropathies, but most prototypically and perhaps most commonly from *celiac sprue*. This gluten-sensitive enteropathy characterized by villous atrophy and crypt hyperplasia in the proximal small bowel often presents with fatty diarrhea associated with multiple nutritional deficiencies of varying severity and affects all ages. *Tropical sprue* may produce a similar histologic and clinical syndrome, but it occurs in residents of or travelers to tropical climates; its often abrupt onset and response to antibiotics suggest an infectious etiology. *Whipple's disease*, due to the actinomycete *Treponema whippleii* and histiocytic infiltration of the small bowel mucosa, is a less common cause of steatorrhea that most typically occurs in young or middle-aged men; it is frequently associated with arthralgias, fever, lymphadenopathy, and extreme fatigue and may affect the central nervous system and endocardium. A similar clinical and histologic picture results from *Mycobacterium avium intracellulare* infection in patients with AIDS. *Abetalipoproteinemia* is a rare defect of chylomicron formation and fat malabsorption in children associated with acanthocytic erythrocytes, ataxia, and retinitis pigmentosa. Several other conditions may cause mucosal malabsorption including infections, especially with protozoa like *Giardia*, numerous medications (e.g., colchicine, cholestyramine, neomycin), and chronic ischemia.

Postmucosal lymphatic obstruction The pathophysiology of this condition, which is due to the rare *congenital intestinal lymphangiectasia* or to *acquired lymphatic obstruction* secondary to trauma, tumor, or infection, leads to the unique constellation of fat malabsorption with enteric losses of protein (often causing edema) and lymphocytes (with resultant lymphocytopenia) that enter the portal circulation directly. Carbohydrate and amino acid absorption are preserved.

Inflammatory Causes Inflammatory diarrheas are generally accompanied by pain, fever, bleeding, or other manifestations of inflammation. The mechanism of diarrhea may not only be exudation but, depending on lesion site, may include fat malabsorption, disrupted fluid/electrolyte absorption, and hypersecretion or hypermotility from release of cytokines and other inflammatory mediators. The unifying feature on stool analysis is the presence of leukocytes or leukocyte-derived proteins such as calprotectin. With severe inflammation, exudative protein loss can lead to anasarca (generalized edema). Any middle-aged or older person with chronic inflammatory-type diarrhea, especially with blood, should be carefully evaluated to exclude a colorectal or large enteric tumor.

Idiopathic inflammatory bowel disease The illnesses in this category, which include *Crohn's disease* and *chronic ulcerative colitis*, are among the most common organic causes of chronic diarrhea in adults and range in severity from mild to fulminant and life threatening. They may be associated with uveitis, polyarthralgias, cholestatic liver disease (primary sclerosing cholangitis), and various skin lesions (erythema nodosum, pyoderma gangrenosum). *Microscopic colitis*, including *collagenous colitis*, is an increasingly recognized cause of chronic watery diarrhea; biopsy of a normal appearing colorectum is required for histologic diagnosis.

Primary or secondary forms of immunodeficiency Immunodeficiency may lead to prolonged infectious diarrhea. With common, variable *hypogammaglobulinemia*, diarrhea is particularly prevalent and often the result of giardiasis.

Eosinophilic gastroenteritis Eosinophil infiltration of the mucosa, muscularis, or serosa at any level of the gastrointestinal tract may cause diarrhea, pain, vomiting, or ascites. Affected patients often have an atopic history, Charcot-Leyden crystals due to extruded eosinophil contents may be seen on microscopic inspection of stool, and peripheral eosinophilia is present in 50 to 75% of patients. While hypersensitivity to certain foods occurs in adults, true food allergy causing chronic diarrhea is rare.

Other Causes Chronic inflammatory diarrhea may be caused by *radiation enterocolitis*, *chronic graft-versus-host disease*, *Behcet's syndrome*, and *Cronkite-Canada syndrome*, among others.

Dysmotile Causes Rapid transit may accompany many diarrheas as a secondary or contributing phenomenon, but primary dysmotility is an unusual etiology of true diarrhea. Stool features often suggest a secretory diarrhea, but mild steatorrhea up to 14 g of fat per day can be produced by maldigestion from rapid transit alone. *Hyperthyroidism*, *carcinoid syndrome*, and certain drugs (e.g., prostaglandins, prokinetic agents) may produce hypermotility with resultant diarrhea. Primary visceral neuromyopathies or idiopathic acquired intestinal pseudo-obstruction may lead to stasis with secondary bacterial overgrowth causing diarrhea. *Diabetic diarrhea*, often accompanied by peripheral and generalized autonomic neuropathies, may occur in part because of intestinal dysmotility.

The exceedingly common *irritable bowel syndrome* (10% point prevalence, 1 to 2% per year incidence) is characterized by disturbed intestinal and colonic motor and sensory responses to various stimuli. Symptoms of stool frequency typically cease at night, alternate with periods of constipation, are accompanied by abdominal pain relieved with defecation, and rarely result in weight loss or true diarrhea.

Factitial Causes Factitial diarrhea accounts for up to 15% of unexplained diarrheas referred to tertiary care centers. Either as a form of *Munchausen syndrome* (deception or self-injury for secondary gain) or *bulimia*, some patients covertly self-administer laxatives alone or in combination with other medications (e.g., diuretics) or surreptitously add water or urine to stool sent for analysis. Such patients are typically women, often with histories of psychiatric illness and disproportionately from careers in health care. Hypotension and hypokalemia are common co-presenting features. Such patients often deny this possibility when confronted, but they do benefit from psychiatric counseling when they acknowledge their behavior.

Approach to the Patient

The laboratory tools available to evaluate the very common problem of chronic diarrhea are extensive, and many are costly and invasive. As such, the diagnostic evaluation must be rationally directed by a careful history and physical examination, and simple triage tests are often warranted before complex investigations are launched (Fig. 42-3). The history, physical examination, and routine blood studies should attempt to characterize the mechanism of diarrhea, identify diagnostically helpful associations, and assess the patient's fluid/electrolyte and nutritional status. Patients should be questioned about the onset, duration, pattern, aggravants (especially diet), relieving factors, and stool characteristics of their diarrhea. The presence or absence of fecal incontinence, fever, weight loss, pain, certain exposures (travel, medications, contacts with diarrhea), and common extraintestinal manifestations (skin changes, arthralgias, oral aphtha) should be noted. Physical findings may offer clues such as a thyroid mass, wheezing, heart murmurs, edema, hepatomegaly, abdominal masses, lymphadenopathy, mucocutaneous abnormalities, perianal fistulae, or anal sphincter laxity. Peripheral blood counts may reveal leukocytosis that suggests inflammation; anemia that reflects blood loss or nutritional deficiencies; or eosinophilia that may occur with parasitoses, neoplasia, collagen-vascular disease, allergy, or eosinophilic gastroenteritis. Blood chemistries may demonstrate electrolyte, hepatic, or other metabolic disturbances.

A therapeutic trial is often appropriate, definitive, and highly cost-effective when a specific diagnosis is suggested on the initial physician encounter. For example, chronic watery diarrhea, which ceases with fasting in an otherwise healthy young adult, may justify a trial of a lactose-restricted diet; bloating and diarrhea persisting since a mountain backpacking trip may warrant a trial of metronidazole for likely giardiasis; and postprandial diarrhea persisting since an ileal resection might be treated with cholestyramine before further evaluation. Persistent symptoms require additional investigation.

Certain diagnoses may be suggested on the initial encounter, e.g., idiopathic inflammatory bowel disease; however, additional focused evaluations may be necessary to confirm the diagnosis and characterize the severity or extent of disease so that treatment can be best guided. Patients suspected of having irritable bowel syndrome should be initially evaluated with proctosigmoidoscopy and mucosal biopsies; those with normal findings might be reassured and, as indicated, treated empirically with antispasmodics, antidiarrheals, bulk agents, anxiolytes, or antidepressants. Any patient who presents with chronic diarrhea and hematochezia should be evaluated with stool microbiologic studies and colonoscopy.

In an estimated two-thirds of cases, the cause for chronic diarrhea remains unclear after the initial encounter, and further testing is required. Quantitative stool collection and analyses can yield important objective data that may establish a diagnosis or characterize the type of diarrhea as a triage for focused additional studies (Fig. 42-3). If stool weight exceeds 200 g/d, additional stool analyses should be performed that might include electrolyte concentration, pH, occult blood testing, leukocyte inspection (or leukocyte protein assay), fat quantitation, and laxative screens.

For secretory diarrheas (watery, normal osmotic gap), possible medication-related side effects or surreptitious laxative use should be reconsidered. Microbiologic studies should be done including fecal bacterial cultures (including media for *Aeromonas* and *Pleisiomonas*), inspection for ova and parasites, and *Giardia* antigen assay (the most sensitive test for giardiasis). Small bowel bacterial overgrowth can be excluded by intestinal aspirates with quantitative cultures or with glucose or xylose breath tests involving measurement of breath hydrogen or other metabolite (e.g.,$_{14}CO_2$). However, interpretation of these breath tests may be confounded by disturbances of intestinal transit. When suggested by history or other findings, screens for peptide hormones should be pursued (e.g., serum gastrin,VIP, calcitonin, and thyroid hormone/thyroid stimulating hormone, or urinary 5-hydroxyindolacetic acid and histamine). Upper endoscopy and colonoscopy with biopsies and small bowel barium x-rays are helpful to rule out structural or occult inflammatory disease.

Further evaluation of osmotic diarrhea should include tests for lactose intolerance and magnesium ingestion, the two most common causes. Low fecal pH suggests carbohydrate malabsorption; lactose malabsorption can be confirmed by lactose breath testing or by a therapeutic trial with lactose exclusion and observation of the effect of lactose challenge (e.g., a quart of milk). Lactase determination on small bowel biopsy is generally not available. If fecal Mg^{2+} or laxative levels are elevated, then inadvertent or surreptitious ingestion should be considered and psychiatric help should be sought.

For those with proven fatty diarrhea, endoscopy with small bowel biopsy (including aspiration for *Giardia* and quantitative cultures) should be performed; if this procedure is unrevealing, a small bowel radiograph is often an appropriate next step. If small bowel studies are negative or if pancreatic disease is suspected, pancreatic exocrine insufficiency should be excluded with direct tests, such as the secretin-cholecystokinin stimulation test, or by indirect tests, such as assay of fecal chymotrypsin activity or a bentiromide test.

Chronic inflammatory-type diarrheas should be suspected by the presence of blood or leukocytes in the stool. Such findings warrant stool cultures, inspection for ova and parasites, *C. difficile* toxin assay, colonoscopy with biopsies, and if indicated, small bowel oral contrast studies.

TREATMENT

Treatment of chronic diarrhea depends on the specific etiology and may be curative, suppressive, or empiric. If the cause can be eradicated, treatment is curative as with resection of a colorectal cancer, antibiotic administration for Whipple's disease, or discontinuation of an offending drug. For many chronic conditions, diarrhea can be controlled by suppression of the underlying mechanism. Examples include elimination of dietary lactose for lactase deficiency or gluten for celiac sprue, use of glucocorticoids or other anti-inflammatory agents for idiopathic inflammatory bowel diseases, adsorptive agents such as cholestyramine for ileal bile acid malabsorption, proton pump inhibitors such as omeprazole for the gastric hypersecretion of gastrinomas, somatostatin analogues such as octreotide for malignant carcinoid, prostaglandin inhibitors such as indomethacin for medullary carcinoma of the thyroid, and pancreatic enzyme replacement for pancreatic insufficiency. When the specific cause or mechanism of chronic diarrhea evades diagnosis, empiric therapy may be beneficial. Mild opiates such as diphenoxylate or loperamide are often helpful in mild or moderate watery diarrhea. For those with more severe diarrhea, codeine or tincture of opium may be beneficial. Such antimotility agents should be avoided with inflammatory bowel disease, as toxic megacolon may be precipitated. Clonidine, ana_2-adrenergic agonist, may allow control of diabetic diarrhea. For all patients with chronic diarrhea, fluid and electrolyte repletion is an important component of management (see "Acute Diarrhea," above). Replacement of fat-soluble vitamins may also be necessary in patients with chronic steatorrhea.

CONSTIPATION

DEFINITION

Constipation is a common complaint in clinical practice and usually refers to persistent, difficult, infrequent, or seemingly incomplete defecation. Because of the wide range of normal bowel habits, constipation is difficult to define precisely. Most persons have at least three bowel movements per week; however, stool frequency alone is not a sufficient criterion for the diagnosis of constipation because many constipated patients describe a normal frequency of defecation but subjective complaints of excessive straining, hard stools, lower abdominal fullness, and a sense of incomplete evacuation. The individual patient's symptoms must be analyzed in detail to ascertain what is meant by "constipation" or "difficulty" with defecation.

Stool form and consistency are well correlated with the time elapsed from the preceding defecation. Hard, pellety stools occur with slow transit, while loose watery stools are associated with rapid transit. Small, pellety stools are more difficult to expel than large ones.

The perception of hard stools or excessive straining is more difficult to assess objectively, and the need for enemas or digital disimpaction is a clinically useful way to corroborate the patient's perceptions of difficult defecation.

Psychosocial factors may also be important. A person whose parents attached great importance to daily defecation will become greatly concerned when he or she misses a daily bowel movement; some children withhold stool to gain attention; and some adults are simply too busy or too embarrassed to interrupt their work when the call to have a bowel movement is sensed.

CAUSES

Pathophysiologically, chronic constipation generally results from inadequate fiber intake or from disordered colonic transit or anorectal function as a result of a neurogastroenterologic disturbance, certain drugs, or in association with a large number of systemic diseases that affect the gastroinestinal tract (Table 42-4). Constipation of recent onset may be a symptom of significant organic disease such as tumor or stricture. In *idiopathic constipation*, a subset of patients exhibit delayed emptying of the ascending and transverse colon with prolongation of transit (often in the proximal colon) and a reduced frequency of propulsive colonic contractions (HAPCs). *Outlet obstruction to defecation* (also called *evacuation disorders*) may cause delayed colonic transit, which is usually corrected by biofeedback retraining of the disordered defecation. Constipation of any cause may be exacerbated by chronic illnesses that lead to physical or mental impairment and result in inactivity or physical immobility.

Approach to the Patient

A careful history should explore the patient's symptoms and confirm whether he or she is indeed constipated based on frequency (e.g., <3 bowel movements per week), consistency (lumpy/hard), excessive straining, prolonged defecation time, or need to support the perineum or digitate the anorectum. In the vast majority of cases (probably >90%), there is no underlying cause (e.g., cancer, depression, or hypothyroidism), and constipation responds to ample hydration, exercise, and supplementation of dietary fiber (15 to 25 g/d). A good diet and medication history and attention to psychosocial issues are key. Physical examination and, particularly, a rectal examination should exclude most of the important diseases that present with constipation and possibly indicate features suggesting an evacuation disorder (e.g., high anal sphincter tone).

There is broad consensus on the selection of patients for further investigation. The presence of weight loss, rectal bleeding, or anemia with constipation mandates either sigmoidoscopy plus barium enema or colonoscopy alone, particularly in patients over 40 years of age, to exclude structural diseases such as cancer or strictures. Colonoscopy alone is most cost effective in this setting since it provides an opportunity to biopsy mucosal lesions, perform polypectomy, or dilate strictures. Barium enema has advantages over colonoscopy in the patient with isolated constipation, since it is less costly and identifies colonic dilatation and all significant mucosal lesions or strictures that are likely to present with constipation. Melanosis coli, or pigmentation of the colon mucosa, indicates the use of anthraquinone laxatives such as cascara or senna; however, this is usually apparent from a careful history. An unexpected disorder such as megacolon or cathartic colon may also be detected by colonic radiographs. Measurement of serum calcium and thyroid stimulating hormone levels will identify rare patients with metabolic disorders.

Patients with more troublesome constipation may not respond to fiber alone and may be helped by a bowel training regimen: taking an osmotic laxative and evacuating with enema or glycerine suppository as needed. After breakfast, a distraction-free 15 to 20 min on the toilet without straining is encouraged. Excessive straining may lead to development of hemorrhoids, and, if there is weakness of the pelvic floor or injury to the pudendal nerve, may result in obstructed defecation from descending perineum syndrome several years later. Those few who do not benefit from the simple measures delineated above or require long-term treatment with stimulant laxatives with the attendant risk of developing laxative abuse syndrome are assumed to have severe or intractable constipation and should have further investigation (Fig. 42-4).

INVESTIGATION OF SEVERE CONSTIPATION

A small minority (probably <5%) of all patients with constipation have cases that are considered severe or "intractable"; these are the patients most likely to be seen by gastroenterologists or in referral centers. Further observation of the patient may occasionally reveal a previously unrecognized cause, such as an evacuation disorder, laxative abuse, malingering, or psychiatric disorder. In these patients, recent studies suggest that evaluations of the physiologic function of the colon and pelvic floor and of psychological status aid in the rational choice of treatment. Even among these highly selected patients with severe constipation, a cause can be identified in only about 30% (see below).

Measurement of Colonic Transit Radiopaque marker transit tests are easy, repeatable, generally safe, inexpensive, reliable, and highly applicable in evaluating constipated patients in clinical practice. There are several validated methods that are very simple. For example, radiopaque markers are ingested, and an abdominal flat film taken 5 d later should indicate passage of 80% of the markers out of the colon. This test does not provide useful information about the transit profile of the stomach and small bowel, and avoidance of laxatives or enemas during the testing period is essential.

Radioscintigraphy with a delayed-release capsule containing radiolabeled particles has been used to noninvasively characterize normal, accelerated, or delayed colonic function over 24 to 48 h with low radiation exposure. This approach simultaneously assesses gastric, small bowel, and colonic transit. The disadvantages are the greater cost and the need for specific materials prepared in a nuclear medicine laboratory.

Anorectal and Pelvic Floor Tests Pelvic floor dysfunction is suggested by the inability to evacuate the rectum, a feeling of persistent rectal fullness, rectal pain, the need to extract stool from the rectum digitally, application of pressure on the posterior wall of the vagina, support of the perineum during straining, and excessive straining. These significant symptoms should be contrasted with the sense of incomplete rectal evacuation, which is common in irritable bowel syndrome.

Patients with clinically suspected obstruction of defecation should also be evaluated by a psychologist to identify eating disorders or a "need to control," to provide stress management or relaxation training, and to identify depression.

A simple clinical test in the office to document a nonrelaxing puborectalis muscle is to have the patient strain to expel the index finger during a digital rectal exam. Motion of the puborectalis posteriorly during straining indicates proper coordination of the pelvic floor muscles.

Measurement of perineal descent is relatively easy to gauge clinically by placing the patient in the left decubitus position and watching the perineum to assess either paucity or lack of descent (<1.5 cm, a sign of pelvic floor dysfunction) or perineal ballooning during straining relative to bony landmarks (>4 cm, suggesting excessive perineal descent).

A useful overall test of evacuation is the balloon expulsion test. A urinary catheter is placed in the rectum, the balloon is inflated to 50 ml with water, and a determination is made about whether the patient can expel it while seated on a toilet or in the left lateral decubitus position. In the lateral position, the weight needed to facilitate expulsion of the balloon (normal, 0 to 200 g) is determined.

Anorectal manometry is not often contributory in the evaluation of patients presenting with severe constipation, except when an excessively high resting or squeeze anal sphincter tone suggests anismus (anal sphincter spasm). This test also identifies rare syndromes, such as adult Hirschsprung's disease, by the absence of the rectoanal inhibitory reflex or the presence of occult incontinence.

Defecography (a dynamic barium enema including lateral views obtained during barium expulsion) reveals "soft abnormalities" in many patients; the most relevant findings are the measured changes in rectoanal angle, anatomic defects of the rectum, and enteroceles or rectoceles. In a very small proportion of patients, significant anatomic defects associated with intractable constipation respond best to surgical treatment. These defects include severe intussusception with complete outlet obstruction due to funnel-shaped plugging at the anal canal or an extremely large rectocele that is preferentially filled during attempts at defecation instead of expulsion of the barium through the anus. In summary, defecography requires an interested and experienced radiologist, and abnormalities are not pathognomonic for pelvic floor dysfunction. More commonly, outlet obstruction results from a nonrelaxing puborectalis muscle, which impedes rectal emptying, rather than from defects identified by defecography.

Dynamic imaging studies such as proctography during defecation or scintigraphic expulsion of artificial stool help measure perineal descent and the rectoanal angle during rest, squeezing and straining, and scintigraphic expulsion quantitates the amount of "artificial stool" emptied. Failure of the rectoanal angle to increase significantly (~15°) during straining confirms pelvic floor dysfunction.

Neurologic testing (EMG) is more helpful in the evaluation of patients with incontinence than of those with symptoms suggesting obstructed defecation. The absence of neurologic signs in the lower extremities suggests that any documented denervation of the puborectalis results from pelvic (e.g., obstetric) injury or from stretching of the pudendal nerve by chronic, long-standing straining.

Ultrasonography identifies sphincter or rectal wall defects and may help select patients for surgical correction. Spinal-evoked responses during electrical rectal stimulation or stimulation of external anal sphincter contraction by applying magnetic stimulation over the lumbosacral cord identify patients with limited sacral neuropathies with sufficient residual nerve conduction to attempt biofeedback training.

In summary, a balloon expulsion test is an important screening test for anorectal dysfunction. If positive, an anatomic evaluation of the rectum or anal sphincters and an assessment of pelvic floor relaxation are the tools for evaluating patients in whom obstructed defecation is suspected.

TREATMENT

After the cause of constipation is characterized, a treatment decision can be made. Slow transit constipation requires aggressive medical or surgical treatment; anismus or pelvic floor dysfunction usually responds to biofeedback management (Fig. 42-4). However, only about 30% of patients with severe constipation are found to have such a physiologic disorder.

Patients with slow transit constipation are treated with bulk, osmotic, and stimulant laxatives, including fiber, psyllium, milk of magnesia, lactulose, polyethylene glycol (colonic lavage solution), and bisacodyl. If a 2- to 3-month trial of medical therapy fails and patients continue to have documented slow transit constipation unassociated with obstructed defecation, colectomy with ileorectostomy is indicated. The decision to resort to surgery is facilitated in the presence of megacolon and megarectum. The complications after surgery include small bowel obstruction (11%) and fecal soiling, particularly at night during the first postoperative year.

Patients who have a combined disorder should pursue pelvic floor retraining (biofeedback and muscle relaxation), psychological counseling, and dietetic advice first, followed by colectomy and ileorectosomy if colonic transit studies do not normalize with biofeedback alone. In patients with pelvic floor dysfunction alone, biofeedback training has a 70 to 80% success rate, measured by the acquisition of comfortable stool habits. Attempts to manage pelvic floor dysfunction with operations (internal anal sphincter or puborectalis muscle division) have achieved only mediocre success and have been largely abandoned.

(Bibliography omitted in Palm version)

Back to Table of Contents

43. WEIGHT LOSS - *Carol M. Reife*

Significant unintentional weight loss in a previously healthy individual is often a harbinger of underlying systemic disease. During the routine medical history, therefore, inquiry should always be made about changes in weight; loss of 5% of body weight over 6 to 12 months should prompt further evaluation.

PHYSIOLOGY OF WEIGHT REGULATION

The normal individual maintains weight at a remarkably stable "set point," given the wide variation in daily caloric intake and level of activity. Because of the physiologic importance of maintaining energy stores, voluntary weight loss is difficult to achieve and sustain.

Appetite and metabolism are regulated by an intricate network of neural and hormonal factors. The hypothalamic feeding and satiety centers play a central role in these processes (Chap. 77). Neuropeptides, like corticotropin-releasing hormone (CRH), a-melanocyte stimulating hormone (a-MSH), and cocaine and amphetamine-related transcript (CART) induce anorexia by acting centrally on satiety centers. Epinephrine and norepinephrine cause a decrease in food intake and an increase in metabolic rate (Chap. 72). Amphetamines and related drugs used to suppress appetite act by releasing norepinephrine in the central nervous system. The gastrointestinal peptides glucagon, somatostatin, and particularly cholecystokinin induce a decrease in food intake by acting through a vagal mechanism to signal satiety. Hypoglycemia decreases levels of insulin which reduces glucose utilization and inhibits activity of the satiety center.

Leptin plays a central role in the long-term maintenance of weight homeostasis (Chap. 77). Leptin is produced by adipose tissue and acts on the hypothalamus to decrease food intake and increase energy expenditure. It suppresses expression of hypothalamic neuropeptide Y, a potent appetite stimulatory peptide. In parallel, leptin increases the expression ofa-MSH, which decreases appetite by acting on the MC4R melanocortin receptor. Thus, leptin activates a series of downstream neural pathways that alter food-seeking behavior and metabolism. However, leptin deficiency, which occurs in conjunction with the loss adipose tissue, stimulates appetite and induces other adaptive responses including inhibition of hypothalamic thyrotropin releasing hormone (TRH) and gonadotropin releasing hormone (GnRH).

A variety of cytokines, including tumor necrosis factor a(TNF-a), interleukin (IL) (IL-6), IL-1, interferong (IFN-g), ciliary neurotrophic factor (CNTF), and leukemia inhibitory factor (LIF), can contribute to cachexia (Chap. 17). In addition to causing anorexia, these factors may induce fever, depress myocardial function, modulate immune and inflammatory responses, and induce a variety of specific metabolic alterations. TNF-a, for example, preferentially mobilizes fat but spares skeletal muscle. Levels of one or more of these cytokines may be increased in patients with cancer, sepsis, chronic inflammatory conditions, AIDS, and congestive heart failure.

Weight loss occurs when energy expenditure exceeds calories available for energy utilization. In most individuals, approximately half of food energy is utilized for basal processes such as maintenance of body temperature. In a 70-kg person, basal activity consumes about 1800 kcal/d. About 40% of caloric intake is used for physical activity, although athletes may use more than 50% during vigorous exercise. About 10% of caloric intake is used for dietary thermogenesis, the energy expended for digestion, absorption, and metabolism of food.

Mechanisms of weight loss include decreased food intake, malabsorption, loss of calories, and increased energy requirements (Fig. 43-1). Changes in weight may reflect alterations in either tissue mass or body fluid content. A deficit of 3500 kcal generally correlates with the loss of 1 lb (0.45 kg) of body fat, but one must also consider water weight (2.2 lb/L) gained or lost. Weight loss that persists over weeks to months is almost invariably due to loss of tissue mass.

Food intake may be influenced by a wide variety of visual, olfactory, and gustatory stimuli as well by genetic, psychological, and social factors. Absorption may be impaired because of pancreatic insufficiency, cholestasis, celiac sprue, intestinal tumors, radiation injury, inflammatory bowel disease, infection, or medication effect. Manifestations of these disease processes may be suggested by changes in stool frequency and consistency. Calories also may be lost due to vomiting or diarrhea, glucosuria in diabetes mellitus, or fistulous drainage. Resting energy expenditure decreases with age and can be affected by thyroid status. Beginning at about age 60, body weight declines by an average of 0.5% per year. Body composition is also affected by aging; adipose tissue increases and lean muscle mass decreases with age.

SIGNIFICANCE OF WEIGHT LOSS

Unintentional weight loss, especially in the elderly, is not uncommon and is associated with increased morbidity and mortality rates, even after comorbid conditions have been taken into account. Prospective studies indicate that significant involuntary weight loss is associated with a mortality rate of 25% over the next 18 months. Retrospective studies of significant weight loss in the elderly document mortality rates of 9 to 38% over a 2- to 3-year period.

Cancer patients with weight loss have decreased performance status, response to chemotherapy, and median survival (Chap. 79). Marked degrees of weight loss also predispose to infection. Patients undergoing elective surgery, who have lost more than 10 lb (4.5 kg) in 6 months, have higher surgical mortality rates. Vitamin and nutrient deficiencies also can accompany significant weight loss (Chap. 74).

CAUSES OF WEIGHT LOSS

The list of possible causes of weight loss is extensive (Table 43-1). In the elderly, the most common causes of weight loss are depression, cancer, and benign gastrointestinal disease. Lung and gastrointestinal cancer are the most common malignancies in patients presenting with weight loss. In younger individuals, diabetes mellitus, hyperthyroidism, psychiatric disturbances including eating disorders, and infection, especially with HIV, should be considered.

The cause of involuntary weight loss is rarely occult. Careful history and physical examination, in association with directed diagnostic testing, will identify the cause of weight loss in 75% of patients. The etiology of weight loss will not be found in the remaining patients, despite extensive testing. Patients with negative evaluations tend to have lower mortality rates than those found to have organic disease.

Patients with medical causes of weight loss usually have signs or symptoms that suggest involvement of a particular organ system. Gastrointestinal tumors, including those of the pancreas and liver, may affect food intake early in the course of illness, causing weight loss before other symptoms are apparent. Lung cancer may present with post-obstructive pneumonia, dyspnea, or cough and hemoptysis; however, it may be silent and should be considered even in those without a history of cigarette smoking. Depression and isolation can cause profound weight loss, especially in the elderly. Chronic pulmonary disease and congestive heart failure can produce anorexia and may also increase resting energy expenditure. Weight loss may be the presenting sign of infectious diseases such as HIV infection, tuberculosis, endocarditis, and fungal and parasitic infections. Hyperthyroidism or pheochromocytoma increase metabolism; elderly patients with apathetic hyperthyroidism may present with weight loss alone. New onset diabetes mellitus is often accompanied by weight loss, reflecting glucosuria and loss of the anabolic actions of insulin. Adrenal insufficiency may be suggested by increased pigmentation, hyponatremia, and hyperkalemia.

Approach to the Patient

Before extensive evaluation is undertaken, it is important to confirm that weight loss has occurred. Almost half of patients who claim significant weight loss have no actual change in weight when it is measured objectively. If weight loss is present, efforts should be made to determine the time interval over which it has occurred. In the absence of documentation, changes in belt notch size or the fit of clothing may help confirm loss of weight. Not infrequently, patients who have actually sustained significant weight loss are unaware that it has occurred. Routine documentation of weight during office visits is therefore important.

The review of systems should focus on signs or symptoms that are associated with disorders that commonly cause weight loss. These include fever, pain, shortness of breath or cough, palpitations, changes in pattern of urination, and evidence of neurologic disease. Gastrointestinal disturbances, including difficulty eating, dysphagia, anorexia, nausea, and change in bowel habits, should be sought. Use of cigarettes, alcohol, and all medications should be reviewed, and patients should be questioned about previous illness or surgery as well as diseases in family members. Risk factors for HIV infection should be assessed. Signs of depression, evidence of dementia, and social factors, including financial issues that might affect food intake, should be considered.

Physical examination should begin with weight determination and documentation of vital signs. The skin should be examined for pallor, jaundice, turgor, scars from prior surgery, and stigmata of systemic disease. The search for oral thrush or dental disease, thyroid gland enlargement, adenopathy, and respiratory or cardiac abnormalities and a detailed examination of the abdomen often lead to clues for further evaluation. Rectal examination, including prostate exam and testing of stool for occult blood, should be performed in men; and all women should have a pelvic examination, even if they have had a hysterectomy. Neurologic examination should include mental status assessment and screening for depression.

Laboratory testing should confirm or exclude possible diagnoses elicited from the history and physical examination (Table 43-2). An initial phase of testing should include a complete blood count with differential, serum chemistry tests including glucose, electrolytes, renal and liver tests, calcium, thyroid stimulating hormone (TSH), urinalysis, and chest x-ray. Patients at risk for HIV infection should have HIV antibody testing. In all cases, recommended cancer screening tests appropriate for the gender and age group, such as mammograms and Pap smears, should be updated (Chap. 80). If gastrointestinal signs or symptoms are present, upper and/or lower endoscopy and abdominal imaging with either computed tomography (CT) or magnetic resonance imaging (MRI) have a relatively high yield, consistent with the high prevalence of gastrointestinal disorders in patients with weight loss. If an etiology of weight loss is not found, careful clinical follow-up, rather than persistent undirected testing, is reasonable.

(Bibliography omitted in Palm version)

Back to Table of Contents

44. GASTROINTESTINAL BLEEDING - *Loren Laine*

Bleeding from the gastrointestinal (GI) tract may present in 5 ways. *Hematemesis* is vomitus of red blood or "coffee-grounds" material. *Melena* is black, tarry, foul-smelling stool. *Hematochezia* is the passage of bright red or maroon blood from the rectum. *Occult GI bleeding (GIB)* may be identified in the absence of overt bleeding by special examination of the stool (e.g., guaiac testing). Finally, patients may present only with *symptoms of blood loss or anemia* such as lightheadedness, syncope, angina, or dyspnea.

SOURCES OF GASTROINTESTINAL BLEEDING

UPPER GASTROINTESTINAL SOURCES OF BLEEDING (Table 44-1)

The annual incidence of hospital admissions for upper GIB (UGIB) in the United States and Europe is approximately 0.1%, with a mortality rate of ~10%. Patients rarely die from exsanguination; rather, they die due to decompensation from other underlying illnesses. The mortality rate for patients under 60 years of age in the absence of malignancy or organ failure is<1%.

Peptic ulcers are the most common cause ofUGIB, accounting for about 50% of cases. Mallory-Weiss tears account for 5 to 15% of cases. The proportion of patients bleeding from varices varies widely from ~5 to 30%, depending on the population. Hemorrhagic or erosive gastropathy [e.g., due to nonsteroidal anti-inflammatory drugs (NSAIDs) or alcohol] and erosive esophagitis often cause mild UGIB, but major bleeding is rare.

Peptic Ulcers Clinical features that predict poorer outcome include hemodynamic instability, the number of units of blood transfused, red blood in the emesis and the stool, increasing age, and the presence of concurrent illness. Characteristics of the ulcer at endoscopy also provide important prognostic information. One-third of patients with active bleeding or a non-bleeding visible vessel have further bleeding that requires urgent surgery if they are treated conservatively. These patients clearly benefit from endoscopic therapy with bipolar electrocoagulation, heater probe, or injection therapy (e.g., absolute alcohol, 1:10,000 epinephrine), with reductions in bleeding, hospital stay, mortality rate, and costs. In contrast, patients with clean-based ulcers have rates of recurrent bleeding approaching zero. If there is no other reason for hospitalization, such patients may be discharged on the first hospital day, following stabilization. Patients without clear-based ulcers should usually remain in the hospital for 3 days, since most episodes of recurrent bleeding occur within 3 days.

Various pharmacologic agents have been assessed in the past for the treatment of ulcer bleeding without clearcut benefit. However, in recent controlled trials in Europe and Asia, high-dose intravenous omeprazole used to raise intragastric pH to 6 to 7 and enhance clot stability decreased further bleeding (but not mortality), even after the use of appropriate endoscopic therapy.

Approximately one-third of patients with a bleeding ulcer will rebleed within the next 1 to 2 years. Prevention of recurrent bleeding focuses on the three main factors in ulcer pathogenesis, *Helicobacter pylori*,NSAIDs, and acid. Eradication of *H. pylori* in patients with bleeding ulcers dramatically decreases rates of rebleeding to < 5%. If a bleeding ulcer develops in a patient taking NSAIDs, the NSAIDs should be discontinued if possible. If NSAIDs must be continued, initial treatment should be with a proton pump inhibitor, and subsequent prophylactic therapy with a proton pump inhibitor or misoprostol should be continued as long as the patient is taking NSAIDs. Changing from a standard NSAID to a COX-2-specific inhibitor should markedly lower the risk of recurrentUGIB. Patients with bleeding ulcers unrelated to *H. pylori* or NSAIDs should remain on full-dose antisecretory therapy indefinitely.**Peptic ulcers are discussed in Chap. 285.*

Mallory-Weiss Tears The classic history is vomiting, retching, or coughing preceding hematemesis, especially in an alcoholic patient. Bleeding from these tears, which are usually on the gastric side of the gastroesophageal junction, stops spontaneously in 80 to 90% of patients and recurs in only 0 to 5%. Endoscopic therapy is effective for actively bleeding Mallory-Weiss tears. Angiographic therapy with intra-arterial infusion of vasopressin or embolization also may be useful. Rarely, operative therapy with oversewing of the tear may be required.**Mallory-Weiss tears are discussed in Chap. 284.*

Esophageal Varices Patients withUGIB and clinical evidence suggesting the possibility of liver disease should undergo early endoscopy to determine if varices are the sources of bleeding, because patients with variceal hemorrhage have poorer outcomes than patients with other sources of UGIB. Endoscopic therapy at this time decreases further bleeding, and repeated sessions with endoscopic therapy to eradicate esophageal varices significantly reduces rebleeding and mortality. Endoscopic ligation therapy is the endoscopic therapy of choice for esophageal varices because it has less rebleeding, a lower mortality rate, fewer local complications, and requires fewer treatment sessions to achieve variceal eradication as compared to sclerotherapy.

Acute treatment with octreotide (50 ug bolus and 50 ug/h intravenous infusion for 2 to 5 days) or somatostatin may help in the control of acute bleeding, and these agents have replaced vasopressin as the medical therapy of choice for acute variceal bleeding. Over the long term, treatment with nonselective beta blockers (e.g., propranolol) has also been shown to decrease recurrent bleeding from esophageal varices. These agents commonly are given along with chronic endoscopic therapy.

In patients who have persistent or recurrent bleeding despite endoscopic and medical therapy, more invasive therapy is warranted. Transjugular intrahepatic portosystemic shunt (TIPS) decreases rebleeding more effectively than endoscopic therapy, although hepatic encephalopathy is more common and the mortality rates are comparable. Most patients with TIPS have shunt stenosis within 1 to 2 years and require re-instrumentation. Therefore, TIPS is most appropriate for patients with more severe liver disease and those in whom transplant is anticipated. Patients with milder, well-compensated cirrhosis probably should undergo decompressive surgery (e.g., distal splenorenal shunt).

Portal hypertension is also responsible for bleeding from gastric varices, ectopic varices in the small and large intestine, and portal hypertensive gastropathy and enterocolopathy.

Hemorrhagic and Erosive Gastropathy ("Gastritis") Hemorrhagic and erosive gastropathy or gastritis refers to endoscopically visualized subepithelial hemorrhages and erosions. These are mucosal lesions and thus do not cause major bleeding. They develop in various clinical settings, the most important of which are ingestion ofNSAIDs, alcohol, and stress. Half of patients who chronically ingest NSAIDs have erosions (15 to 30% have ulcers), while up to 20% of actively drinking alcoholic patients with symptoms ofUGIBhave evidence of subepithelial hemorrhages or erosions.

Stress-related gastric mucosal injury occurs only in extremely sick patients: those who have experienced serious trauma, major surgery, burns covering more than one-third of the body surface area, major intracranial disease, and severe medical illness (ventilator dependency, coagulopathy). Significant bleeding probably does not develop unless ulceration occurs. The mortality rate in these patients is quite high because of their serious underlying illnesses.

The incidence of bleeding from stress-related gastric mucosal injury or ulceration has decreased dramatically in recent years, most likely due to better care of critically ill patients. Pharmacologic prophylaxis for bleeding may be considered in the high-risk patients mentioned above. The best clinical data suggest that intravenous H_2-receptor antagonist therapy is the treatment of choice, although sucralfate also is effective. Prophylactic therapy decreases bleeding, but it does not lower the mortality rate.

Other Causes Other, less frequent causes ofUGIBinclude erosive duodenitis, neoplasms, aortoenteric fistulas, vascular lesions [including hereditary hemorrhagic telengectasias (Osler-Weber-Rendu) and gastric antral vascular ectasia ("watermelon stomach")], Dieulafoy's lesion (in which an aberrant vessel in the mucosa bleeds from a pinpoint mucosal defect), prolapse gastropathy (prolapse of proximal stomach into esophagus with retching, especially in alcoholics), and hemobilia and hemosuccus pancreaticus (bleeding from the bile duct or pancreatic duct).

SMALL INTESTINAL SOURCES OF BLEEDING

Small intestinal sources of bleeding (bleeding from sites beyond the reach of the standard upper endoscope) are difficult to diagnose and are responsible for the majority of cases of obscureGIB. Fortunately, small intestinal bleeding is uncommon. The most common causes are vascular ectasias and tumors (e.g., adenocarcinoma, leiomyoma, lymphoma, benign polyps, carcinoid, metastases, and lipoma). Other less common causes include Crohn's disease, infection, ischemia, vasculitis, small bowel varices, diverticula, Meckel's diverticula, duplication cysts, and intussusception.NSAIDsinduce small intestinal erosions and ulcers and may be a relatively common cause of chronic, obscure GIB.

Meckel's diverticulum is the most common cause of significant lower GIB (LGIB) in children, decreasing in frequency as a cause of bleeding with age. In adults younger than 40 to 50 years, small bowel tumors often account for obscureGIB, while in patients older than 50 to 60 years, vascular ectasias are usually responsible.

Vascular ectasias should be treated with endoscopic therapy if possible. Surgical therapy can be used for vascular ectasias isolated to a segment of the small intestine when endoscopic therapy is unsuccessful; estrogen/progesterone compounds may also be tried. Isolated lesions, such as tumors, diverticula, or duplications, generally are treated with surgical resection.

COLONIC SOURCES OF BLEEDING

The incidence of hospitalizations forLGIB is about one-fifth that forUGIB. Hemorrhoids are probably the most common cause of LGIB; anal fissures also cause minor bleeding and pain. If these local anal processes, which rarely require hospitalization, are excluded, the most common causes of LGIB in adults are diverticula, vascular ectasias (especially in the proximal colon of patients> 70 years), neoplasms (adenomatous polyps and adenocarcinoma), and colitis -- most commonly infectious or idiopathic inflammatory bowel disease, but occasionally ischemic or radiation-induced. Uncommon causes include post-polypectomy bleeding, solitary rectal ulcer syndrome,NSAID-induced ulcers or colitis, other neoplasms, trauma, ectopic varices (most commonly rectal), lymphoid nodular hyperplasia, vasculitis, and aorto-colic fistulas. In children and adolescents, the most common colonic causes of significantGIB are inflammatory bowel disease and juvenile polyps.

Diverticular bleeding is abrupt in onset, usually painless, sometimes massive, and often from the right colon; minor and occult bleeding is not characteristic. Clinical reports suggest that bleeding colonic diverticula stop bleeding spontaneously in approximately 80% of patients, and rebleed in 20 to 25% of patients. Intraarterial vasopressin may halt the bleeding, at least temporarily. If bleeding persists or recurs, segmental surgical resection is indicated.

Bleeding from right colonic vascular ectasias in the elderly may be overt or occult; it tends to be chronic and only occasionally is hemodynamically significant. Endoscopic hemostatic therapy may be useful in the treatment of vascular ectasias, as well as discrete bleeding ulcers and post-polypectomy bleeding, while endoscopic polypectomy, if possible, is used for bleeding colonic polyps. Surgical therapy is generally required for major, persistent, or recurrent bleeding from the wide variety of colonic sources of GIB that cannot be treated medically or endoscopically.

Approach to the Patient

Measurement of the heart rate and blood pressure is the best way to assess a patient withGIB. Clinically significant bleeding leads to postural changes in heart rate or blood pressure, tachycardia, and, finally, recumbent hypotension. Patients also may have a vasovagal reaction with bradycardia during bleeding episodes.

In contrast, the hemoglobin does not fall immediately with acuteGIB, due to proportionate reductions in plasma and red cell volumes (i.e., "people bleed whole blood"). Thus, hemoglobin may be normal or only minimally decreased at the initial presentation of a severe bleeding episode. As extravascular fluid enters the vascular space to restore volume, the hemoglobin falls, but this process may take up to 72 h. Patients with slow, chronic GIB may have very low hemoglobin values despite normal blood pressure and heart rate. With the development of iron deficiency anemia, the mean corpuscular volume will be low and red blood cell distribution width will be increased.

Differentiation of Upper from Lower GIB Hematemesis indicates an upperGIsource of bleeding (above the ligament of Treitz). Melena indicates that blood has been present in the GI tract for at least 14 h. Thus, the more proximal the bleeding site, the more likely melena will occur. Hematochezia usually represents a lower GI source of bleeding, although an upper GI lesion may bleed so rapidly that blood does not remain in the bowel long enough for melena to develop. When hematochezia is the presenting symptom ofUGIB, it is associated with hemodynamic instability and dropping hemoglobin. Bleeding lesions of the small bowel may present as melena or hematochezia.

A non-bloody nasogastric aspirate may be seen in up to 16% of patients withUGIB -- usually from a duodenal source. Even a bile-stained appearance does not exclude a bleeding post-pyloric lesion since reports of bile in the aspirate are incorrect in about 50% of cases. Testing of aspirates that are not grossly bloody for occult blood is of no clinical value. Other clues to UGIB include hyperactive bowel sounds and an elevated BUN (due to volume depletion and absorbed blood proteins).

Diagnostic Evaluation of the Patient withGIB

UPPER GIB (Fig. 44-1) The history and physical exam seldom are diagnostic of the source of GIB. Upper endoscopy is the test of choice in patients withUGIB, and should be performed urgently in patients with hemodynamic instability (hypotension, tachycardia, or postural changes in heart rate or blood pressure). Early routine endoscopy is also beneficial in cases of milder bleeding for management decisions. Patients with major bleeding and high risk endoscopic findings (varices, ulcers with active bleeding or a visible vessel) benefit from endoscopic hemostatic therapy, while patients with low-risk lesions (e.g., clean based ulcers, non-bleeding Mallory-Weiss tears, erosive or hemorrhagic gastropathy) who have stable vital signs and hemoglobin, and no other medical problems, can be discharged home.

LOWERGIB(Fig. 44-2) Patients with presumedLGIB may undergo early sigmoidoscopy for the detection of obvious, low-lying lesions. However, the procedure is difficult with brisk bleeding, and it often is impossible to identify the area of bleeding. Sigmoidoscopy is useful primarily in patients < 40 years with relatively minor bleeding. Patients with hematochezia and hemodynamic instability should have upper endoscopy to rule out an upperGIsource before evaluation of the lower GI tract.

Colonoscopy after an oral lavage solution is the procedure of choice in patients withLGIBunless bleeding is too massive or unless sigmoidoscopy has disclosed an obvious actively bleeding lesion.99MTc-labeled red cell scan allows repeated imaging for up to 24 h and may identify the general location of bleeding. However, radionuclide scans should be interpreted with caution because results are highly variable. In active LGIB, angiography can detect the site of bleeding (extravasation of contrast into the gut) and permits treatment with intraarterial infusion of vasopressin or embolization. Even after bleeding has stopped, angiography may identify lesions with abnormal vasculature such as vascular ectasias or tumors.

GIBOF OBSCURE ORIGIN Obscure GIB is defined as recurrent acute or chronic bleeding for which no source has been identified by routine endoscopic and contrast studies. Push enteroscopy, with a specially designed enteroscope or a pediatric colonoscope to inspect the entire duodenum and part of the jejunum, is generally the next step. Push enteroscopy may identify probable bleeding sites in 20 to 40% of patients with obscure GIB. If enteroscopy is negative or unavailable, a specialized radiographic examination of the small bowel (e.g., enteroclysis) should be performed.

Patients with recurrent bleeding who require transfusions or repeated hospitalizations warrant further investigations.^{99M}Tc-labeled red blood cell scintigraphy should be employed. Angiography is useful even if bleeding has subsided, since it may disclose vascular anomalies or tumor vessels.^{99M}Tc-pertechnetate scintigraphy for diagnosis of Meckel's diverticulum should be done, especially in the evaluation of young patients withLGIB. When all tests are unrevealing, intraoperative endoscopy is indicated in patients with severe recurrent or persistent bleeding requiring repeated transfusions.

*OCCULTGIB*Occult GIB is manifested by either a positive test for fecal occult blood or iron deficiency anemia. Unless a patient has upperGIsymptoms, evaluation of occult bleeding generally should begin with colonoscopy, particularly in patients older than 40 years. If evaluation of the colon is negative, some perform upper endoscopy only if iron deficiency anemia or upper GI symptoms are present, while others recommend upper endoscopy in all patients since up to 25 to 40% of these patients have some abnormality noted on upper endoscopy. If standard endoscopic tests are unrevealing, enteroscopy and/or enteroclysis may be considered in patients with iron-deficiency anemia.

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

45. JAUNDICE - *Daniel S. Pratt*, *Marshall M. Kaplan*

Jaundice, or icterus, is a yellowish discoloration of tissue resulting from the deposition of bilirubin. Tissue deposition of bilirubin occurs only in the presence of serum hyperbilirubinemia and is a sign of either liver disease or, less often, a hemolytic disorder. The degree of serum bilirubin elevation can be estimated by physical examination. Slight increases in serum bilirubin are best detected by examining the sclerae which have a particular affinity for bilirubin due to their high elastin content. The presence of scleral icterus indicates a serum bilirubin of at least 3.0 mg/dL. The ability to detect scleral icterus is made more difficult if the examining room has fluorescent lighting. If the examiner suspects scleral icterus, a second place to examine is underneath the tongue. As serum bilirubin levels rise, the skin will eventually become yellow in light-skinned patients and even green if the process is longstanding; the green color is produced by oxidation of bilirubin to biliverdin.

The differential diagnosis for yellowing of the skin is limited. In addition to jaundice, it includes carotenoderma, the use of the drug quinacrine, and excessive exposure to phenols. Carotenoderma is the yellow color imparted to the skin by the presence of carotene; it occurs in healthy individuals who ingest excessive amounts of vegetables and fruits that contain carotene, such as carrots, leafy vegetables, squash, peaches, and oranges. Unlike jaundice, where the yellow coloration of the skin is uniformly distributed over the body, in carotenoderma the pigment is concentrated on the palms, soles, forehead, and nasolabial folds. Carotenoderma can be distinguished from jaundice by the sparing of the sclerae. Quinacrine causes a yellow discoloration of the skin in 4 to 37% of patients treated with it. Unlike carotene, quinacrine can cause discoloration of the sclerae.

Another sensitive indicator of increased serum bilirubin is darkening of the urine, which is due to the renal excretion of conjugated bilirubin. Patients often describe their urine as tea or cola colored. Bilirubinuria indicates an elevation of the direct serum bilirubin fraction and therefore the presence of liver disease.

Increased serum bilirubin levels occur when an imbalance exists between bilirubin production and clearance. A logical evaluation of the patient who is jaundiced requires an understanding of bilirubin production and metabolism.

PRODUCTION AND METABOLISM OF BILIRUBIN (See also Chap. 294)

Bilirubin, a tetrapyrrole pigment, is a breakdown product of heme (ferroprotoporphyrin IX). About 70 to 80% of the 250 to 300 mg of bilirubin produced each day is derived from the breakdown of hemoglobin in senescent red blood cells. The remainder comes from prematurely destroyed erythroid cells in bone marrow and from the turnover of hemoproteins such as myoglobin and crytochromes found in tissues throughout the body.

The formation of bilirubin occurs in reticuloendothelial cells, primarily in the spleen and liver. The first reaction, catalyzed by the enzyme heme oxygenase, oxidatively cleaves thea bridge of the porphyrin group and opens the heme ring. The end products of this reaction are biliverdin, carbon monoxide, and iron. The second reaction, catalyzed by the cytosolic enzyme biliverdin reductase, reduces the central methylene bridge of biliverdin and converts it to bilirubin. Bilirubin formed in the reticuloendothelial cells is virtually insoluble in water. To be transported in blood, it must be solubilized. This is accomplished by its reversible, noncovalent binding to albumin. Unconjugated bilirubin bound to albumin is transported to the liver, where it, but not the albumin, is taken up by hepatocytes via a process that at least partly involves carrier-mediated membrane transport.

In the cytosol of the hepatocyte, unconjugated bilirubin is coupled predominantly to the protein ligandin (formerly called the Y protein). Ligandin was initially thought to be a transport protein facilitating the movement of bilirubin from the sinusodial membrane to the endoplasmic reticulum. It is now thought to slow the cytosolic diffusion of bilirubin and to reduce its efflux back into serum. In the endoplasmic reticulum, bilirubin is solubilized by conjugation to glucuronic acid, forming bilirubin monoglucuronide and diglucuronide. The conjugation of glucuronic acid to bilirubin is catalyzed by bilirubin uridine-diphosphate (UDP) glucuronosyltransferase.

The now hydrophilic bilirubin conjugates diffuse from the endoplasmic reticulum to the canalicular membrane, where bilirubin monoglucuronide and diglucuronide are actively transported into canalicular bile by an energy-dependent mechanism involving the multiple organic ion transport protein/multiple drug resistance protein. The conjugated bilirubin excreted into bile drains into the duodenum and passes unchanged through the proximal small bowel. Conjugated bilirubin is not taken up by the intestinal mucosa. When the conjugated bilirubin reaches the distal ileum and colon, it is hydrolyzed to unconjugated bilirubin by bacterialb-glucuronidases. The unconjugated bilirubin is reduced by normal gut bacteria to form a group of colorless tetrapyrroles called urobilinogens. About 80 to 90% of these products are excreted in feces, either unchanged or oxidized to orange derivatives called urobilins. The remaining 10 to 20% of the urobilinogens are passively absorbed, enter the portal venous blood, and are reexcreted by the liver. A small fraction (usually less than 3 mg/dL) escapes hepatic uptake, filters across the renal glomerulus, and is excreted in urine.

MEASUREMENT OF SERUM BILIRUBIN

The terms direct- and indirect-reacting bilirubin are based on the original van den Bergh reaction. This assay, or a variation of it, is still used in most clinical chemistry laboratories to determine the serum bilirubin level. In this assay, bilirubin is exposed to diazotized sulfanilic acid, splitting into two relatively stable dipyrrylmethene azopigments that absorb maximally at 540 nm, allowing for photometric analysis. The direct fraction is that which reacts with diazotized sulfanilic acid in the absence of an accelerator substance such as alcohol. The direct fraction provides an approximate determination of the conjugated bilirubin in serum. The total serum bilirubin is the amount that reacts after the addition of alcohol. The indirect fraction is the difference between the total and the direct bilirubin and provides an estimate of the unconjugated bilirubin in serum.

With the van den Bergh method, the normal serum bilirubin concentration usually is<1 mg/dL (17 umol/L). Up to 30%, or 0.3 mg/dL (5.1 umol/L), of the total may be direct-reacting (conjugated) bilirubin. Total serum bilirubin concentrations are between 0.2 and 0.9 mg/dL in 95% of a normal population.

Several new techniques, although less convenient to perform, have added considerably to our understanding of bilirubin metabolism. First, they demonstrate that in normal people or those with Gilbert's syndrome, almost 100% of the serum bilirubin is unconjugated; less than 3% is monoconjugated bilirubin. Second, in jaundiced patients with hepatobiliary disease, the total serum bilirubin concentration measured by these new, more accurate methods is lower than the values found with diazo methods. This suggests that there are diazo-positive compounds distinct from bilirubin in the serum of patients with hepatobiliary disease. Third, these studies indicate that in jaundiced patients with hepatobiliary disease, monoglucuronides of bilirubin predominate over the diglucuronides. Fourth, part of the direct-reacting bilirubin fraction includes conjugated bilirubin that is covalently linked to albumin. This albumin-linked bilirubin fraction (*delta fraction* or *biliprotein*) represents an important fraction of total serum bilirubin in patients with cholestasis and hepatobiliary disorders. Albumin-bound conjugated bilirubin is formed in serum when hepatic excretion of bilirubin glucuronides is impaired and the glucuronides are present in serum in increasing amounts. By virtue of its tight binding to albumin, the clearance rate of albumin-bound bilirubin from serum approximates the half-life of albumin, 12 to 14 days, rather than the short half-life of bilirubin, about 4 h.

The prolonged half-life of albumin-bound conjugated bilirubin explains two previously unexplained enigmas in jaundiced patients with liver disease: (1) that some patients with conjugated hyperbilirubinemia do not exhibit bilirubinuria during the recovery phase of their disease because the bilirubin is bound to albumin and therefore not filtered by the renal glomeruli and (2) that the elevated serum bilirubin level declines more slowly than expected in some patients who otherwise appear to be recovering satisfactorily. Late in the recovery phase of hepatobiliary disorders, all the conjugated bilirubin may be in the albumin-linked form. Its value in serum falls slowly because of the long half-life of albumin.

MEASUREMENT OF URINE BILIRUBIN

Unconjugated bilirubin is always bound to albumin in the serum, is not filtered by the kidney, and is not found in the urine. Conjugated bilirubin is filtered at the glomerulus and the majority is reabsorbed by the proximal tubules; a small fraction is excreted in the urine. Any bilirubin found in the urine is conjugated bilirubin. The presence of bilirubinuria implies the presence of liver disease. A urine dipstick test (Ictotest) gives the same information as fractionation of the serum bilirubin. This test is very accurate. A false-negative test is possible in patients with prolonged cholestasis due to the predominance of conjugated bilirubin covalently bound to albumin.

THE EVALUATION OF JAUNDICE

The bilirubin present in serum represents a balance between input from production of bilirubin and hepatic/biliary removal of the pigment. Hyperbilirubinemia may result from (1) overproduction of bilirubin; (2) impaired uptake, conjugation, or excretion of bilirubin; or (3) regurgitation of unconjugated or conjugated bilirubin from damaged hepatocytes or bile ducts. An increase in unconjugated bilirubin in serum results from either overproduction, impairment of uptake, or conjugation of bilirubin. An increase in conjugated bilirubin is due to decreased excretion into the bile ductules or backward leakage of the pigment. The initial steps in evaluating the patient with jaundice are to determine (1) whether the hyperbilirubinemia is predominantly conjugated or unconjugated in nature, and (2) whether other biochemical liver tests are abnormal. The thoughtful interpretation of limited data will allow for a rational evaluation of the patient (Fig. 45-1). This discussion will focus solely on the evaluation of the adult patient with jaundice.

ISOLATED ELEVATION OF SERUM BILIRUBIN

Unconjugated Hyperbilirubinemia The differential diagnosis of an isolated unconjugated hyperbilirubinemia is limited (Table 45-1). The critical determination is whether the patient is suffering from a hemolytic process resulting in an overproduction of bilirubin (hemolytic disorders and ineffective erythropoiesis) or from impaired hepatic uptake/conjugation of bilirubin (drug effect or genetic disorders).

Hemolytic disorders that cause excessive heme production may be either inherited or acquired. Inherited disorders include spherocytosis, sickle cell anemia, and deficiency of red cell enzymes such as pyruvate kinase and glucose-6-phosphate dehydrogenase. In these conditions, the serum bilirubin rarely exceeds 5 mg/dL. Higher levels may occur when there is coexistent renal or hepatocellular dysfunction, or in acute hemolysis such as a sickle cell crisis. In evaluating jaundice in patients with chronic hemolysis, it is important to remember the high incidence of pigmented (calcium bilirubinate) gallstones found in these patients, which increases the likelihood of choledocholithiasis as an alternative explanation for hyperbilirubinemia.

Acquired hemolytic disorders include microangiopathic hemolytic anemia (e.g., hemolytic-uremic syndrome), paroxysmal nocturnal hemoglobinuria, and immune hemolysis. Ineffective erythropoiesis occurs in cobalamin, folate, and iron deficiencies.

In the absence of hemolysis, the physician should consider a problem with the hepatic uptake or conjugation of bilirubin. Certain drugs, including rifampicin and probenecid, may cause unconjugated hyperbilirubinemia by diminishing hepatic uptake of bilirubin. Impaired bilirubin conjugation occurs in three genetic conditions: *Crigler-Najjar syndrome, types I and II*, and *Gilbert's syndrome*. *Crigler-Najjar type I* is an exceptionally rare condition found in neonates and characterized by severe jaundice (bilirubin> 20 mg/dL) and neurologic impairment due to kernicterus, frequently leading to death in infancy or childhood. These patients have a complete absence of bilirubinUDPglucuronosyltransferase activity, usually due to mutations in the critical 3¢ domain of the UDP glucuronosyltransferase gene, and are totally unable to conjugate, hence cannot excrete bilirubin. The only effective treatment is orthotopic liver transplantation. Use of gene therapy and allogeneic hepatocyte infusion are experimental approaches of future promise for this devastating disease.

Crigler-Najjar type II is somewhat more common than type I. Patients live into adulthood with serum bilirubin levels that range from 6 to 25 mg/dL. In these patients, mutations in the bilirubinUDPglucuronosyltransferase gene cause reduced but not completely absent activity of the enzyme. Bilirubin UDP glucuronosyltransferase activity can be induced by the administration of phenobarbital, which can reduce serum bilirubin levels in these patients. Despite marked jaundice, these patients usually survive into adulthood, although they may be susceptible to kernicterus under the stress of intercurrent illness or surgery.

Gilbert's syndrome is also marked by the impaired conjugation of bilirubin due to reduced bilirubinUDPglucuronosyltransferase activity. Molecular analyses show that Gilbert's syndrome is due to reduced expression of UDP glucuronosyltransferase activity caused by lengthening of the TATAA box from $A(TA)_6TAA$ to $A(TA)_7TAA$ in the promoter element of the gene. This results in mild unconjugated hyperbilirubinemia with serum levels almost always less than 6 mg/dL. The serum levels may fluctuate and jaundice is often identified only during periods of fasting. Unlike both Crigler-Najjar syndromes, Gilbert's syndrome is very common. The reported incidence is 3 to 7% of the population with males predominating over females by a ratio of 2-7:1.

Conjugated Hyperbilirubinemia Elevated conjugated hyperbilirubinemia is found in two rare inherited conditions: *Dubin-Johnson syndrome* and *Rotor's syndrome* (Table 45-1). Patients with both conditions present with asymptomatic jaundice, typically in the second generation of life. The defect in Dubin-Johnson syndrome is a point mutation in the gene for the canalicular multispecific organic anion transporter. These patients have altered excretion of bilirubin into the bile ducts. Rotor's syndrome seems to be a problem with the hepatic storage of bilirubin. Differentiating between these syndromes is possible, but clinically unnecessary, due to their benign nature.

ELEVATION OF SERUM BILIRUBIN WITH OTHER LIVER TEST ABNORMALITIES

The remainder of this chapter will focus on the evaluation of the patient with a conjugated hyperbilirubinemia in the setting of other liver test abnormalities. This group of patients can be divided into those with a primary hepatocellular process and those with intra- or extrahepatic cholestasis. Being able to make this differentiation will guide the physician's evaluation (Fig. 45-1). This differentiation is made on the basis of the history and physical examination as well as the pattern of liver test abnormalities.

History A complete medical history is perhaps the single most important part of the evaluation of the patient with unexplained jaundice. Important considerations include the use of or exposure to any chemical or medication, either physician-prescribed or over-the-counter, such as herbal and vitamin preparations and other drugs such as anabolic steroids. The patient should be carefully questioned about possible parenteral exposures, including transfusions, intravenous and intranasal drug use, tattoos, and sexual activity. Other important questions include recent travel history, exposure to people with jaundice, exposure to possibly contaminated foods, occupational exposure to hepatotoxins, alcohol consumption, the duration of jaundice, and the presence of any accompanying symptoms such as arthralgias, myalgias, rash, anorexia, weight loss, abdominal pain, fever, pruritis, and changes in the urine and stool. While none of these latter symptoms are specific for any one condition, they can suggest a particular diagnosis. A history of arthralgias and myalgias predating jaundice suggests hepatitis, either viral or drug-related. Jaundice associated with the sudden onset of severe right upper quadrant pain and shaking chills suggests choledocholithiasis and ascending cholangitis.

Physical Examination The general assessment should include assessment of the patient's nutritional status. Temporal and proximal muscle wasting suggests longstanding diseases such as pancreatic cancer or cirrhosis. Stigmata of chronic liver disease, including spider nevi, palmar erythema, gynecomastia, caput medusae, Dupuytren's contractures, parotid gland enlargement, and testicular atrophy are commonly seen in advanced alcoholic (Laennec's) cirrhosis and occasionally in other types of cirrhosis. An enlarged left supraclavicular node (Virchow's node) or periumbilical nodule (Sister Mary Joseph's nodule) suggest an abdominal malignancy. Jugular venous distention, a sign of right-sided heart failure, suggests hepatic congestion. Right pleural effusion, in the absence of clinically apparent ascites, may be seen in advanced cirrhosis.

The abdominal examination should focus on the size and consistency of the liver, whether the spleen is palpable and hence enlarged, and whether there is ascites present. Patients with cirrhosis may have an enlarged left lobe of the liver which is felt below the xiphoid and an enlarged spleen. A grossly enlarged nodular liver or an obvious abdominal mass suggests malignancy. An enlarged tender liver could be viral or alcoholic hepatitis or, less often, an acutely congested liver secondary to right-sided heart failure. Severe right upper quadrant tenderness with respiratory arrest on inspiration (Murphy's sign) suggests cholecystitis or, occasionally, ascending cholangitis. Ascites in the presence of jaundice suggests either cirrhosis or malignancy with peritoneal spread.

Laboratory Tests When the physician encounters a patient with unexplained jaundice, there are a battery of tests that are helpful in the initial evaluation. These include total and direct serum bilirubin with fractionation, aminotransferases, alkaline phosphatase, albumin, and prothrombin time tests. Enzyme tests [alanine aminotransferase (ALT), aspartate aminotransferase (AST), and alkaline phosphatase] are helpful in differentiating between a hepatocellular process and a cholestatic process (see Table 293-1 andFig. 45-1), a critical step in determining what additional workup is indicated. Patients with a hepatocellular process generally have a disproportionate rise in the aminotransferases compared to the alkaline phosphatase. Patients with a cholestatic process have a disproportionate rise in the alkaline phosphatase compared to the aminotransferases. The bilirubin can be prominently elevated in both hepatocellular and cholestatic conditions and therefore is not necessarily helpful in differentiating between the two.

In addition to the enzyme tests, all jaundiced patients should have additional blood tests, specifically an albumin and a prothrombin time, to assess liver function. A low albumin suggests a chronic process such as cirrhosis or cancer. A normal albumin is suggestive of a more acute process such as viral hepatitis or choledocholithiasis. An elevated prothrombin time indicates either vitamin K deficiency due to prolonged jaundice and malabsorption of vitamin K or significant hepatocellular dysfunction. The failure of the prothrombin time to correct with parenteral administration of vitamin K indicates severe hepatocellular injury.

The results of the bilirubin, enzyme, albumin, and prothrombin time tests will usually indicate whether a jaundiced patient has a hepatocellular or a cholestatic disease. The causes and evaluation of each of these is quite different.

Hepatocellular Conditions Hepatocellular diseases that can cause jaundice include viral hepatitis, drug or environmental toxicity, alcohol, and end-stage cirrhosis from any cause (Table 45-2). Wilson's disease should be considered in young adults. Autoimmune hepatitis is typically seen in young to middle-aged women, but may affect men and women of any age. Alcoholic hepatitis can be differentiated from viral and toxin-related hepatitis by the pattern of the aminotransferases. Patients with alcoholic hepatitis typically have anAST:ALTratio of at least 2:1. The AST rarely exceeds 300 U/L. Patients with acute viral hepatitis and toxin-related injury severe enough to produce jaundice typically have aminotransferases greater than 500 U/L, with the ALT greater than or equal to the AST. The degree of aminotransferase elevation can occasionally help in differentiating between hepatocellular and cholestatic processes. While ALT and AST values less than 8 times normal may be seen in either hepatocellular or cholestatic liver disease, values 25 times normal or higher are seen primarily in acute hepatocellular diseases. Patients with jaundice from cirrhosis can have normal or only slight elevations of the aminotransferases.

When the physician determines that the patient has a hepatocellular disease, appropriate testing for acute viral hepatitis includes a hepatitis A IgM antibody, a hepatitis B surface antigen and core IgM antibody, and a hepatitis C viral RNA test. It can take many weeks for the hepatitis C antibody to become detectable, making it an unreliable test if acute hepatitis C is suspected. Depending on circumstances, studies for hepatitis D, E, Epstein-Barr virus (EBV), and cytomegalovirus (CMV) may be indicated. Ceruloplasmin is the initial screening test for Wilson's disease. Testing for autoimmune hepatitis usually includes an antinuclear antibody and measurement of specific immunoglobulins.

Drug-induced hepatocellular injury can be classified either as predictable or unpredictable. Predictable drug reactions are dose-dependent and affect all patients who ingest a toxic dose of the drug in question. The classic example is acetaminophen hepatotoxicity. Unpredictable or idiosyncratic drug reactions are not dose-dependent and occur in a minority of patients. A great number of drugs can cause idiosyncratic hepatic injury. Environmental toxins are also an important cause of hepatocellular injury. Examples include industrial chemicals such as vinyl chloride, herbal preparations containing pyrrolizidine alkaloids (Jamaica bush tea), and the mushrooms Amanita phalloides or verna containing highly hepatotoxic amatoxins.

Cholestatic Conditions When the pattern of the liver tests suggests a cholestatic disorder, the next step is to determine whether it is intra- or extrahepatic cholestasis (Fig. 45-1). Distinguishing intrahepatic from extrahepatic cholestasis may be difficult. History, physical examination, and laboratory tests are often not helpful. The next appropriate test is an ultrasound. The ultrasound is inexpensive, does not expose the patient to ionizing radiation, and can detect dilation of the intra- and extrahepatic biliary tree with a high degree of sensitivity and specificity. The absence of biliary dilatation suggests intrahepatic cholestasis, while the presence of biliary dilatation indicates extrahepatic cholestasis. False-negative results occur in patients with partial obstruction of the common bile duct or in patients with cirrhosis or primary sclerosing cholangitis (PSC) where scarring prevents the intrahepatic ducts from dilating.

Although ultrasonography may indicate extrahepatic cholestasis, it rarely identifies the site or cause of obstruction. The distal common bile duct is a particularly difficult area to visualize by ultrasound because of overlying bowel gas. Appropriate next tests include computed tomography (CT) and endoscopic retrograde cholangiopancreatography (ERCP). CT scanning is better than ultrasonography for assessing the head of the pancreas and for identifying choledocholithiasis in the distal common bile duct, particularly when the ducts are not dilated. ERCP is the gold standard for identifying choledocholithiasis. It is performed by introducing a side-viewing endoscope perorally into the duodenum. The ampulla of Vater is visualized and a catheter is advanced through the ampulla. Injection of dye allows for the visualization of the common bile duct and the pancreatic duct. The success rate for cannulation of the common bile duct ranges from 80 to 95%, depending on the operator's experience. Beyond its diagnostic capabilities, ERCP allows for therapeutic interventions, including the removal of common bile duct stones and the placement of stents. In patients in whom ERCP is unsuccessful, transhepatic cholangiography can provide the same information. Magnetic resonance cholangiopancreatography (MRCP) is a rapidly developing, noninvasive technique for imaging the bile and pancreatic ducts; this may replace ERCP as the initial diagnostic test in cases where the need for intervention is felt to be small.

In patients with apparent *intrahepatic cholestasis*, the diagnosis is often made by serologic testing in combination with percutaneous liver biopsy. The list of possible causes of intrahepatic cholestasis is long and varied (Table 45-3). A number of conditions that typically cause a hepatocellular pattern of injury can also present as a cholestatic variant. Both hepatitis B and C can cause a cholestatic hepatitis (fibrosing cholestatic hepatitis) that has histologic features that mimic large duct obstruction. This disease variant has been reported in patients who have undergone solid organ transplantation. Hepatitis A, alcoholic hepatitis,EBV, andCMV may also present as cholestatic liver disease.

Drugs may cause intrahepatic cholestasis, a variant of drug-induced hepatitis. Drug-induced cholestasis is usually reversible after eliminating the offending drug, although it may take many months for cholestasis to resolve. Drugs most commonly associated with cholestasis are the anabolic and contraceptive steroids. Cholestatic hepatitis has been reported with chlorpromazine, imipramine, tolbutamide, sulindac, cimetidine, and erythromycin estolate. It also occurs in patients taking trimethoprim, sulfamethoxazole, and penicillin-based antibiotics such as ampicillin, dicloxacillin, and clavulinic acid. Rarely, cholestasis may be chronic and associated with progressive fibrosis despite early discontinuation of the drug. Chronic cholestasis has been associated with chlorpromazine and prochlorperazine.

Primary biliary cirrhosis is a disease predominantly of middle-aged women in which there is a progressive destruction of interlobular bile ducts. The diagnosis is made by the presence of the antimitochondrial antibody that is found in 95% of patients.*Primary sclerosing cholangitis* (PSC) is characterized by the destruction and fibrosis of larger bile ducts. The disease may involve only the intrahepatic ducts and present as intrahepatic cholestasis. However, in 65% of patients with PSC, both intra- and extrahepatic ducts are involved. The diagnosis of PSC is made byERCP. The pathognomonic findings are multiple strictures of bile ducts with dilatations proximal to the strictures. Approximately 75% of patients with PSC have inflammatory bowel disease.

The *vanishing bile duct syndrome* and *adult bile ductopenia* are rare conditions in which there are a decreased number of bile ducts seen in liver biopsy specimens. The histologic picture is similar to that found in primary biliary cirrhosis. This picture is seen in patients who develop chronic rejection after liver transplantation and in those who develop graft-versus-host disease after bone marrow transplantation. Vanishing bile duct syndrome also occurs in rare cases of sarcoidosis, in patients taking certain drugs including chlorpromazine, and idiopathically. There are also familial forms of intrahepatic cholestasis, including the *familial intrahepatic cholestatic syndromes, I-III*. Benign recurrent cholestasis is an autosomal recessive disease that appears to be due to mutations in a P type ATPase, which probably acts as a bile acid transporter. The disease is marked by recurrent episodes of jaundice and pruritis; the episodes are self-limited but can be debilitating. *Cholestasis of pregnancy* occurs in the second and third trimesters and resolves after delivery. Its cause is unknown, but the condition is probably inherited and cholestasis can be triggered by estrogen administration.

Other causes of intrahepatic cholestasis include total parenteral nutrition (TPN), nonhepatobiliary sepsis, benign postoperative cholestasis, and a paraneoplastic syndrome associated with a number of different malignancies, including Hodgkin's disease, medullary thyroid cancer, hypernephroma, renal sarcoma, T cell lymphoma, prostate cancer, and several GI malignancies. In patients developing cholestasis in the intensive care unit, the major considerations should be sepsis, shock liver, and TPN jaundice. Jaundice occurring after bone marrow transplantation is most likely due to venoocclusive disease or graft-versus-host disease.

Causes of *extrahepatic cholestasis* can be split into malignant and benign (Table 45-3). Malignant causes include pancreatic, gallbladder, ampullary, and cholangiocarcinoma. The latter is most commonly associated withPSC and is exceptionally difficult to diagnose because its appearance is often identical to PSC. Pancreatic and gallbladder tumors, as well as cholangiocarcinoma, are rarely resectable and have poor prognoses. Ampullary carcinoma has the highest surgical cure rate of all the tumors that present as painless jaundice. Hilar lymphadenopathy due to metastases from other cancers may cause obstruction of the extrahepatic biliary tree.

Choledocholithiasis is the most common cause of extrahepatic cholestasis. The clinical presentation can range from mild right upper quadrant discomfort with only minimal elevations of the enzyme tests to ascending cholangitis with jaundice, sepsis, and circulatory collapse.PSC may occur with clinically important strictures limited to the extrahepatic biliary tree. In cases where there is a dominant stricture, patients can be effectively managed with serial endoscopic dilatations. Chronic pancreatitis rarely causes strictures of the distal common bile duct, where it passes through the head of the pancreas. AIDS cholangiopathy is a condition, usually due to infection of the bile duct epithelium withCMV or cryptosporidium, which has a cholangiographic appearance similar to PSC. These patients usually present with greatly elevated serum alkaline phosphatase levels, mean of 800 IU/L, but the bilirubin is often near normal. These patients do not typically present with jaundice.

SUMMARY

The goal of this chapter is not to provide an encyclopedic review of all of the conditions that can cause jaundice. Rather, it is intended to provide a framework that helps a physician to evaluate the patient with jaundice in a logical way (Fig. 45-1).

Simply stated, the initial step is to obtain appropriate blood tests to determine if the patient has an isolated elevation of serum bilirubin. If so, is the bilirubin elevation due to an increased unconjugated or conjugated fraction? If the hyperbilirubinemia is accompanied by other liver test abnormalities, is the disorder hepatocellular or cholestatic? If cholestatic, is it intra- or extrahepatic? All of these questions can be answered with a thoughtful history, physical examination, and interpretation of laboratory and radiologic tests and procedures.

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

46. ABDOMINAL SWELLING AND ASCITES - *Robert M. Glickman*

ABDOMINAL SWELLING

Abdominal swelling or distention is a common problem in clinical medicine and may be the initial manifestation of a systemic disease or of otherwise unsuspected abdominal disease. *Subjective* abdominal enlargement, often described as a sensation of fullness or bloating, is usually transient and is often related to a functional gastrointestinal disorder when it is not accompanied by objective physical findings of increased abdominal girth or local swelling. *Obesity* and lumbar lordosis, which may be associated with prominence of the abdomen, may usually be distinguished from true increases in the volume of the peritoneal cavity by history and careful physical examination.

Clinical History Abdominal swelling may first be noticed by the patient because of a progressive increase in belt or clothing size, the appearance of abdominal or inguinal hernias, or the development of a localized swelling. Often, considerable abdominal enlargement has gone unnoticed for weeks or months, either because of coexistent obesity or because the ascites formation has been insidious, without pain or localizing symptoms. Progressive abdominal distention may be associated with a sensation of "pulling" or "stretching" of the flanks or groins and vague low back pain. Localized pain usually results from involvement of an abdominal organ (e.g., a passively congested liver, large spleen, or colonic tumor). Pain is uncommon in cirrhosis with ascites, and when it is present, pancreatitis, hepatocellular carcinoma, or peritonitis should be considered. Tense ascites or abdominal tumors may produce increased intraabdominal pressure, resulting in indigestion and heartburn due to gastroesophageal reflux or dyspnea, orthopnea, and tachypnea from elevation of the diaphragm. A coexistent pleural effusion, more commonly on the right, presumably due to leakage of ascitic fluid through lymphatic channels in the diaphragm, also may contribute to respiratory embarrassment. The patient with diffuse abdominal swelling should be questioned about increased alcohol intake, a prior episode of jaundice or hematuria, or a change in bowel habits. Such historic information may provide the clues that will lead one to suspect an occult cirrhosis, a colonic tumor with peritoneal seeding, congestive heart failure, or nephrosis.

Physical Examination A carefully executed general physical examination can yield valuable clues concerning the etiology of abdominal swelling. Thus palmar erythema and spider angiomas suggest an underlying cirrhosis, while supraclavicular adenopathy (Virchow's node) should raise the question of an underlying gastrointestinal malignancy.

Inspection of the abdomen is important. By noting the abdominal contour, one may be able to distinguish localized from generalized swelling. The tensely distended abdomen with tightly stretched skin, bulging flanks, and everted umbilicus is characteristic of ascites. A prominent abdominal venous pattern with the direction of flow away from the umbilicus often is a reflection of portal hypertension; venous collaterals with flow from the lower part of the abdomen toward the umbilicus suggest obstruction of the inferior vena cava; flow downward toward the umbilicus suggests superior vena cava obstruction. "Doming" of the abdomen with visible ridges from underlying intestinal loops is usually due to intestinal obstruction or distention. An epigastric mass, with evident peristalsis proceeding from left to right, usually indicates underlying pyloric obstruction.

A liver with metastatic deposits may be visible as a nodular right upper quadrant mass moving with respiration.

Auscultation may reveal the high-pitched, rushing sounds of early intestinal obstruction or a succussion sound due to increased fluid and gas in a dilated hollow viscus. Careful auscultation over an enlarged liver occasionally reveals the harsh bruit of a vascular tumor, especially a hepatocellular carcinoma, or the leathery friction rub of a surface nodule. A venous hum at the umbilicus may signify portal hypertension and an increased collateral blood flow around the liver. A fluid wave and flank dullness that shifts with change in position of the patient are important signs that indicate the presence of peritoneal fluid. In obese patients, small amounts of fluid may be difficult to demonstrate; on occasion, the fluid may be detected by abdominal percussion with patients on their hands and knees. Small amounts of ascites often can only be detected by ultrasound examination of the abdomen, which can detect as little as 100 mL of fluid. Careful percussion should serve to distinguish generalized abdominal enlargement from localized swelling due to an enlarged uterus, ovarian cyst, or distended bladder. Percussion also can outline an abnormally small or large liver. Loss of normal liver dullness may result from massive hepatic necrosis; it also may be a clue to free gas in the peritoneal cavity, as from perforation of a hollow viscus.

Palpation is often difficult with massive ascites, and ballottement of overlying fluid may be the only method of palpating the liver or spleen. A slightly enlarged spleen in association with ascites may be the only evidence of an occult cirrhosis. When there is evidence of portal hypertension, a soft liver suggests that obstruction to portal flow is extrahepatic; a firm liver suggests cirrhosis as the likely cause of the portal hypertension. A very hard or nodular liver is a clue that the liver is infiltrated with tumor, and when accompanied by ascites, it suggests that the latter is due to peritoneal seeding. The presence of a hard periumbilical nodule (Sister Mary Joseph's nodule) suggests metastatic disease from a pelvic or gastrointestinal primary tumor. A pulsatile liver and ascites may be found in tricuspid insufficiency.

An attempt should be made to determine whether a mass is solid or cystic, smooth or irregular, and whether it moves with respiration. The liver, spleen, and gallbladder should descend with respiration unless they are fixed by adhesions or extension of tumor beyond the organ. A fixed mass not descending with respiration may indicate that it is retroperitoneal. Tenderness, especially if localized, may indicate an inflammatory process such as an abscess; it also may be due to stretching of the visceral peritoneum or tumor necrosis. Rectal and pelvic examinations are mandatory; they may reveal otherwise undetected masses due to tumor or infection.

Radiographic and laboratory examinations are essential for confirming or extending the impressions gained on physical examination. Upright and recumbent films of the abdomen may demonstrate the dilated loops of intestine with fluid levels characteristic of intestinal obstruction or the diffuse abdominal haziness and loss of psoas margins suggestive of ascites. Ultrasonography is often of value in detecting ascites, determining the presence of a mass, or evaluating the size of the liver and spleen. Computed tomography (CT) scanning provides similar information. CT scanning is often necessary to visualize the retroperitoneum, pancreas, and lymph nodes. A plain film of the abdomen may reveal the distended colon of otherwise unsuspected ulcerative colitis and give valuable information as to the size of the liver and spleen. An irregular and elevated right side of the diaphragm may be a clue to a liver abscess or hepatocellular carcinoma. Studies of the gastrointestinal tract with barium or other contrast media are usually necessary in the search for a primary tumor.

ASCITES

The evaluation of a patient with ascites requires that the cause of the ascites be established. In most cases ascites appears as part of a well-recognized illness, that is, cirrhosis, congestive heart failure, nephrosis, or disseminated carcinomatosis. In these situations, the physician should determine that the development of ascites is indeed a consequence of the basic underlying disease and not due to the presence of a separate or related disease process. This distinction is necessary even when the cause of ascites seems obvious. For example, when the patient with compensated cirrhosis and minimal ascites develops progressive ascites that is increasingly difficult to control with sodium restriction or diuretics, the temptation is to attribute the worsening of the clinical picture to progressive liver disease. However, an occult hepatocellular carcinoma, portal vein thrombosis, spontaneous bacterial peritonitis, or even tuberculosis may be responsible for the decompensation. The disappointingly low success in diagnosing tuberculous peritonitis or hepatocullar carcinoma in the patient with cirrhosis and ascites reflects the too-low index of suspicion for the development of such superimposed conditions. Similarly, the patient with congestive heart failure may develop ascites from a disseminated carcinoma with peritoneal seeding.

Diagnostic paracentesis (50 to 100 mL) should be part of the routine evaluation of the patient with ascites. The fluid should be examined for its gross appearance; protein content, cell count, and differential cell count should be determined; and Gram's and acid-fast stains and culture should be performed. Cytologic and cell-block examination may disclose an otherwise unsuspected carcinoma.Table 46-1 presents some of the features of ascitic fluid typically found in various disease states. In some disorders, such as cirrhosis, the fluid has the characteristics of a transudate (<25 g protein per liter and a specific gravity of<1.016); in others, such as peritonitis, the features are those of an exudate. Rather than the total protein content of ascites, many authors prefer the use of a *serum-ascites albumin gradient (SAG)* to characterize ascites. The gradient correlates directly with portal pressure. A gradient >1.1 g/dL (high gradient) is characteristic of uncomplicated cirrhotic ascites and differentiates ascites due to portal hypertension from ascites not due to portal hypertension>95% of the time. A gradient<1.1 g/dL (low gradient) suggests that the ascites is not due to portal hypertension with >95% accuracy and mandates a search for other causes (Table 46-1). Although there is variability of the ascitic fluid in any given disease state, some features are sufficiently characteristic to suggest certain diagnostic possibilities. For example, blood-stained fluid with >25 g protein per liter is unusual in uncomplicated cirrhosis but is consistent with tuberculous peritonitis or neoplasm. Cloudy fluid with a predominance of polymorphonuclear cells and a positive Gram's stain are characteristic of bacterial peritonitis; if most cells are lymphocytes, tuberculosis should be suspected. The complete examination of each fluid is most important, for occasionally only one finding may be abnormal. For example, if the fluid is a typical transudate but contains>250 white blood cells per microliter, the finding should be recognized as atypical for cirrhosis and should warrant a search for tumor or infection. This is especially true in the evaluation of cirrhotic ascites where occult peritoneal infection may be present with only minor elevations in the white blood cell count of the peritoneal fluid (300 to 500 cells per microliter). Since Gram's stain of the fluid may be negative in a high proportion of such cases, careful culture of the peritoneal fluid is mandatory. Bedside innoculation of blood culture flasks with ascitic fluid results in a dramatically increased incidence of positive cultures when bacterial infection is present (90 versus 40% positivity with conventional cultures done by the laboratory). Direct visualization of the peritoneum (laparoscopy) may disclose peritoneal deposits of tumor, tuberculosis, or metastatic disease of the liver. Biopsies are taken under direct vision, often adding to the diagnostic accuracy of the procedure.

Chylous ascites refers to a turbid, milky, or creamy peritoneal fluid due to the presence of thoracic or intestinal lymph. Such a fluid shows Sudan-staining fat globules microscopically and an increased triglyceride content by chemical examination. Opaque milky fluid usually has a triglyceride concentration of>1000 mg/dL. A turbid fluid due to leukocytes or tumor cells may be confused with chylous fluid (pseudochylous), and it is often helpful to carry out alkalinization and ether extraction of the specimen. Alkali tend to dissolve cellular proteins and thereby reduce turbidity; ether extraction leads to clearing if the turbidity of the fluid is due to lipid. Chylous ascites is most often the result of lymphatic obstruction from trauma, tumor, tuberculosis, filariasis (Chap. 221), or congenital abnormalities. It also may be seen in the nephrotic syndrome.

Rarely, ascitic fluid may be *mucinous* in character, suggesting either pseudomyxoma peritonei (Chap. 289) or rarely a colloid carcinoma of the stomach or colon with peritoneal implants.

On occasion, ascites may develop as a seemingly isolated finding in the absence of a clinically evident underlying disease. Then, a careful analysis of ascitic fluid may indicate the direction the evaluation should take. A useful framework for the workup starts with an analysis of whether the fluid is classified as a high (transudate) or low (exudate) gradient fluid. *High gradient (transudative) ascites* of unclear etiology is most often due to occult cirrhosis, right-sided venous hypertension raising hepatic sinusoidal pressure, or hypoalbuminemic states such as nephrosis or protein-losing enteropathy. Cirrhosis with well-preserved liver function (normal albumin) resulting in ascites invariably is associated with significant portal hypertension (Chap. 298). Evaluation should include liver function tests, liver-spleen scan, or other hepatic imaging procedure (i.e., CT or ultrasound) to detect nodular changes in the liver or a colloid shift of isotope to suggest portal hypertension. On occasion, a wedged hepatic venous pressure can be useful to document portal hypertension. Finally, if clinically indicated, a liver biopsy will confirm the diagnosis of cirrhosis and perhaps suggest its etiology. Other etiologies may result in hepatic venous congestion and resultant ascites. Right-sided cardiac valvular disease and particularly constrictive pericarditis should raise a high index of suspicion and may require cardiac imaging and cardiac catheterization for definitive diagnosis. Hepatic vein thrombosis is evaluated by visualizing the hepatic veins with imaging techniques (Doppler ultrasound, angiography, CT scans, magnetic resonance imaging) to demonstrate obliteration, thrombosis, or obstruction by tumor. Uncommonly, transudative ascites may be associated with benign tumors of the ovary, particularly fibroma (Meigs' syndrome) with ascites and hydrothorax.

Low gradient (exudative) ascites should initiate an evaluation for primary peritoneal processes, most importantly infection and tumor. Routine bacteriologic culture of ascitic fluid often yields a specific organism causing infectious peritonitis. Tuberculous peritonitis (Table 46-1) is best diagnosed by peritoneal biopsy, either percutaneously or via laparoscopy. Histologic examination invariably shows granulomata that may contain acid-fast bacilli. Since cultures of peritoneal fluid and biopsies for tuberculosis may require 6 weeks, characteristic histology with appropriate stains allows antituberculosis therapy to be started promptly. Similarly, the diagnosis of peritoneal seeding by tumor can usually be made by cytologic analysis of peritoneal fluid or by peritoneal biopsy if cytology is negative. Appropriate diagnostic studies can then be undertaken to determine the nature and site of the primary tumor. Pancreatic ascites (Table 46-1) is invariably associated with an extravasation of pancreatic fluid from the pancreatic ductal system, most commonly from a leaking pseudocyst. Ultrasound or CT examination of the pancreas followed by visualization of the pancreatic duct by direct cannulation [viz., endoscopic retrograde cholangiopancreatography (ERCP)] usually discloses the site of leakage and permits resective surgery to be carried out.

An analysis of the physiologic and metabolic factors involved in the production of ascites (detailed inChap. 298), coupled with a complete evaluation of the nature of the ascitic fluid, invariably discloses the etiology of the ascites and permits appropriate therapy to be instituted.

ACKNOWLEDGEMENT

Dr. Kurt J. Isselbacher was the co-author of this chapter in previous editions.

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 7 -ALTERATIONS IN RENAL AND URINARY TRACT FUNCTION

47. AZOTEMIA AND URINARY ABNORMALITIES - *Bradley M. Denker*, *Barry M. Brenner*

Body homeostasis is maintained predominantly through the cellular processes that together comprise normal kidney function. Disturbances to any of these functions can lead to a constellation of abnormalities that may be detrimental to survival. The clinical manifestations of these diseases will depend upon the pathophysiology of the renal injury and will often be initially identified as a complex of symptoms, abnormal physical findings, and laboratory changes that will allow the identification of specific syndromes. These renal syndromes (summarized in Table 47-1) may arise as the consequence of a systemic illness or can occur as a primary renal disease. Nephrologic syndromes usually consist of several elements that reflect the underlying pathologic processes and the duration of the disease and typically include one or more of the following features: (1) disturbances in urine volume (oliguria, anuria, polyuria); (2) abnormalities of urine sediment [red blood cells (RBC); white blood cells, casts, and crystals]; (3) abnormal excretion of serum proteins (proteinuria); (4) reduction in glomerular filtration rate (GFR) (azotemia); (5) presence of hypertension and/or expanded total body volume (edema); (6) electrolyte abnormalities, or (7) in some syndromes, fever/pain. The combination of these findings should permit identification of one of the major nephrologic syndromes (Table 47-1) and will allow the differential diagnoses to be narrowed and the appropriate diagnostic evaluation and therapeutic course to be determined. Each of these syndromes and their associated diseases are discussed in more detail in subsequent chapters. This chapter will focus on several aspects of renal abnormalities that are critically important to distinguishing these processes: (1) reduction in GFR leading to azotemia, (2) alterations of the urinary sediment and/or protein excretion, and (3) abnormalities of urinary volume.

AZOTEMIA

ASSESSMENT OF GLOMERULAR FILTRATION RATE

Monitoring the GFRis important in both the hospital and outpatient settings, and several different methodologies are available (discussed below). In most acute clinical circumstances a measured GFR is not available, and it is necessary to estimate the GFR from the serum creatinine level in order to provide appropriate doses of drugs that are excreted into the urine. Serum creatinine is the most widely used marker for GFR and is related directly to the urine creatinine excretion and inversely to the serum creatinine (U_{Cr}/P_{Cr}). Based upon this relationship and some important caveats (discussed below), the GFR will fall proportionately with the increase in P_{Cr}. Failure to account for GFR reductions in drug dosing can lead to significant morbidity and mortality from drug toxicities (e.g., digoxin, aminoglycosides). In the outpatient setting, serial determinations of GFR are helpful for following the progression of chronic renal insufficiency, but again, the serum creatinine is often used as a surrogate for GFR (although much less accurate; see below). In patients with chronic progressive renal insufficiency there is an approximately linear relationship between $1/P_{Cr}$and time. The slope of this line will remain constant for an individual patient, and when values are obtained that do not fall on this line, an investigation for a superimposed acute process (e.g., volume depletion, drug reaction) should be initiated. It should be emphasized that the signs and symptoms of uremia will develop at significantly different levels of serum creatinine depending upon the patient (size, age, and sex), the underlying renal disease, existence of concurrent diseases, and true GFR. In general, patients do not develop symptomatic uremia until renal insufficiency is usually quite severe (GFR< 15 mL/min) and in some patients it does not occur until the GFR < 5 mL/min.

A reducedGFRleads to retention of nitrogenous waste products (azotemia) such as serum urea nitrogen and creatinine. Azotemia may result from reduced renal perfusion, intrinsic renal disease, or postrenal processes (ureteral obstruction; see below and Fig. 47-1). Precise determination of GFR is problematic as both commonly used markers (urea and creatinine) have characteristics that affect their accuracy as markers of clearance. Urea clearance is generally an underestimate of GFR because of tubule urea reabsorption and may be as low as one-half of GFR measured by other techniques.

Creatinine is a small, freely filtered solute that varies little from day to day (since it is derived from muscle metabolism of creatine). However, serum creatinine can increase acutely from dietary ingestion of cooked meat. Creatinine can be secreted by the proximal tubule through an organic cation pathway. There are many clinical settings where a creatinine clearance is not available, and decisions concerning drug dosing must be made based on the serum creatinine. A formula that allows an estimate of creatinine clearance in men that accounts for age-related decreases in GFR, body weight, and sex has been derived by Cockcroft-Gault:

This value should be multiplied 0.85 for women, since a lower fraction of the body weight is composed of muscle. The gradual loss of muscle from chronic illness, chronic use of glucocorticoids, or malnutrition can mask significant changes inGFR with small or imperceptible changes in serum creatinine. More accurate determinations of GFR are available using inulin clearance or radionuclide-labeled markers such as$_{125}$I-iothalamate or EDTA. These methods are highly accurate due to precise quantitation and the absence of any renal reabsorption/secretion and should be used to follow GFR in patients in whom creatinine is not likely to be a reliable indicator (patients with decreased muscle mass secondary to age, malnutrition, concurrent illnesses).

Approach to the Patient

Once it has been established that GFRis reduced, the physician must decide if this represents acute or chronic renal failure. The clinical situation, history, and laboratory data often make this an easy distinction. However, the laboratory abnormalities characteristic of chronic renal failure, including anemia, hypocalcemia, and hyperphosphatemia, are often also present in patients presenting with acute renal failure. Radiographic evidence of renal osteodystrophy (Chap. 270) would be seen only in chronic renal failure but is a very late finding, and these patients are usually on dialysis. The urinalysis and renal ultrasound can occasionally facilitate distinguishing acute from chronic renal failure. An approach to the evaluation of azotemic patients is shown in Fig. 47-1. Patients with advanced chronic renal insufficiency often have some proteinuria, nonconcentrated urine (isosthenuria), and small kidneys on ultrasound characterized by increased echogenicity and cortical thinning. Treatment should be directed toward slowing the progression of renal disease and providing symptomatic relief for edema, acidosis, anemia, and hyperphosphatemia, as discussed inChap. 270. Acute renal failure (Chap. 269) can result from processes affecting renal blood flow (prerenal azotemia), intrinsic renal diseases (affecting vessels, glomeruli, or tubules), or postrenal processes (obstruction to urine flow in ureters, bladder, or urethra) (Chap. 281).

Prerenal Failure Decreased renal perfusion accounts for 40 to 80% of acute renal failure and, if appropriately treated, is readily reversible. The etiologies of prerenal azotemia include any cause of decreased circulating blood volume including volume loss (gastrointestinal hemorrhage, burns, diarrhea, diuretics), volume sequestration (pancreatitis, peritonitis, rhabdomyolysis), or decreased effective circulating volume (cardiogenic shock, sepsis). Renal perfusion can also be affected by reductions in cardiac output from peripheral vasodilatation (sepsis, drugs) or profound renal vasoconstriction [severe heart failure, hepatorenal syndrome, drugs (such as nonsteroidal anti-inflammatory drugs (NSAIDs)]. True, or "effective," hypovolemia leads to a fall in mean arterial pressure, which in turn triggers a series of neural and humoral responses that include activation of the sympathetic nervous and renin-angiotensin-aldosterone systems and ADH release.GFR is maintained by prostaglandin-mediated relaxation of afferent arterioles and angiotensin II-mediated constriction of efferent arterioles. Once the mean arterial pressure falls below 80 mmHg, there is a steep decline in GFR.

Blockade of prostaglandin production byNSAIDscan result in severe vasoconstriction and acute renal failure under these circumstances. Angiotensin-converting enzyme (ACE) inhibitors decrease efferent arteriolar tone and can decrease glomerular capillary perfusion pressure. Patients on NSAIDs and/or ACE inhibitors are most susceptible to hemodynamically mediated acute renal failure when blood volume is reduced for any reason. Patients with renal artery stenosis are dependent upon efferent arteriolar vasoconstriction for maintenance of glomerular filtration pressure and are particularly susceptible to precipitous decline inGFR when given ACE inhibitors.

Prolonged renal hypoperfusion can lead to acute tubular necrosis (ATN; an intrinsic renal disease discussed below). The urinalysis and urinary electrolytes can be useful in distinguishing prerenal azotemia from ATN (Table 47-2). The urine of patients with prerenal azotemia can be predicted from the stimulatory actions of norepinephrine, angiotensin II, ADH, and low tubule fluid flow on salt and water reabsorption from the urine. In prerenal conditions the tubules are intact, leading to a concentrated urine (>500 mosm), avid Na retention (urine Na concentration <20 m*M*/L; fractional excretion of Na <1%), and U_{Cr}/P_{Cr}> 40 (Table 47-2). The prerenal urine sediment is usually normal or has occasional hyaline and granular casts, while the sediment of ATN is usually filled with cellular debris and muddy brown granular casts.

Intrinsic Renal Disease When prerenal and postrenal azotemia have been excluded as etiologies of renal failure, an intrinsic parenchymal renal disease is present. Intrinsic renal disease can arise from processes involving large renal vessels, microvasculature and glomeruli, or tubulointerstitium. Ischemic and toxicATNaccount for about 90% of acute intrinsic renal failure. As outlined inFig. 47-1, the clinical setting and urinalysis are helpful in separating the possible etiologies of acute intrinsic renal failure. Prerenal azotemia and ATN are part of a spectrum of renal hypoperfusion; evidence of structural tubule injury is present in ATN, whereas prompt reversibility occurs with prerenal azotemia upon restoration of adequate renal perfusion. Thus, ATN can often be distinguished from prerenal azotemia by urinalysis and urine electrolyte composition (Table 47-2 andFig. 47-1). Ischemic ATN is observed most frequently in patients who have undergone major surgery, trauma, severe hypovolemia, overwhelming sepsis, or extensive burns. Nephrotoxic ATN complicates the administration of many common medications, usually by inducing a combination of intrarenal vasoconstriction, direct tubule toxicity, and/or tubular obstruction. The kidney is vulnerable to toxic injury by virtue of its rich blood supply (25% of cardiac output) and its ability to concentrate and metabolize toxins. A diligent search for hypotension and nephrotoxins will usually uncover the specific etiology of ATN. Discontinuation of nephrotoxins and stabilizing blood pressure will often suffice without the need for dialysis while the tubules recover.**An extensive list of potential drugs and toxins implicated in ATN can be found in Chap. 269.*

Processes that involve the tubules and interstitium can lead to acute renal failure. These include drug-induced interstitial nephritis (especially antibiotics,NSAIDs, and diuretics), severe infections (both bacterial and viral), systemic diseases (e.g., systemic lupus erythematosus), or infiltrative disorders (e.g., sarcoid, lymphoma, or leukemia). A list of drugs associated with allergic interstitial nephritis can be found in Chap. 277. The urinalysis usually shows mild to moderate proteinuria, hematuria, and pyuria (approximately 75% of cases) and occasionally white blood cell casts. The finding of RBC casts in interstitial nephritis has been reported but should prompt a search for glomerular diseases. Occasionally renal biopsy will be needed to distinguish among these possibilities. The finding of eosinophils in the urine is suggestive of allergic interstitial nephritis and is optimally observed by using a Hansel stain. The absence of eosinophiluria, however, does not exclude the possibility of acute interstitial nephritis.

Occlusion of large renal vessels including arteries and veins is an uncommon cause of acute renal failure. A significant reduction inGFR by this mechanism suggests bilateral processes or a unilateral process in a patient with a single functioning kidney. Renal arteries can be occluded with atheroemboli, thromboemboli, in situ thrombosis, aortic dissection, or vasculitis. Atheroembolic renal failure can occur spontaneously but is most often associated with recent aortic instrumentation. The emboli are cholesterol-rich and lodge in medium and small renal arteries leading to an eosinophil-rich inflammatory reaction. Atheroembolic acute renal failure often has a normal urinalysis but may contain eosinophils and casts. The diagnosis can be confirmed by renal biopsy, but this is often unnecessary when other stigmata of atheroemboli are present (livedo reticularis, distal peripheral infarcts, eosinophilia). Renal artery thrombosis may lead to mild proteinuria and hematuria, whereas renal vein thrombosis typically induces heavy proteinuria and hematuria.**These vascular catastrophes often require angiography for confirmation and are discussed in Chap. 278.*

Diseases of glomeruli (glomerulonephritis or vasculitis) and the renal microvasculature (hemolytic uremic syndromes, thrombotic thrombocytopenic purpura, or malignant hypertension) usually present with various combinations of glomerular injury: proteinuria, hematuria, reducedGFR, and alterations of Na excretion leading to hypertension, edema, and circulatory congestion (acute nephritic syndrome). These findings may occur as primary renal diseases or as renal manifestations of systemic diseases. The clinical setting and other laboratory data will help distinguish primary renal from systemic diseases. The finding ofRBCcasts in the urine is an indication for early renal biopsy (Fig. 47-1) as the pathologic pattern has important implications for diagnosis, prognosis, and treatment. Hematuria without RBC casts can also be an indication of glomerular disease, and this evaluation is summarized in Fig. 47-2.**A detailed discussion of glomerulonephritis and diseases of the microvasculature can be found in Chap. 274.*

Postrenal Azotemia Urinary tract obstruction accounts for fewer than 5% of cases of acute renal failure, but it is usually reversible and must be ruled out early in the evaluation (Fig. 47-1). Since a single kidney is capable of adequate clearance, acute renal failure from obstruction requires obstruction at the urethra or bladder outlet, bilateral ureteral obstruction, or unilateral obstruction in a patient with a single functioning kidney. Obstruction is usually diagnosed by the presence of ureteral dilatation on renal ultrasound. However, early in the course of obstruction or if the ureters are unable to dilate (such as encasement by pelvic tumors), the ultrasound examination may be negative.**The specific urologic conditions that cause obstruction are discussed in Chap. 281.*

Oliguria and Anuria *Oliguria* refers to a 24-h urine output of<500 mL, and *anuria* is the complete absence of urine formation. Anuria can be caused by total urinary tract obstruction, total renal artery or vein occlusion, and shock (manifested by severe hypotension and intense renal vasoconstriction). Cortical necrosis,ATN, and rapidly progressive glomerulonephritis can occasionally cause anuria. Oliguria can accompany any cause of acute renal failure and carries a more serious prognosis for renal recovery in all conditions except prerenal azotemia. *Nonoliguria* refers to urine output in excess of 500 mL/day in patients with acute or chronic azotemia. With nonoliguric ATN, disturbances of potassium and hydrogen balance are less severe than in oliguric patients and recovery to normal renal function is usually more rapid.

ABNORMALITIES OF THE URINE

PROTEINURIA

The evaluation of proteinuria is shown schematically inFig. 47-3 and is typically initiated after colorimetric detection of proteinuria by dipstick examination. Current methods for measuring proteinuria vary significantly. The dipstick measurement detects mostly albumin and gives false-positive results when pH > 7.0 and the urine is very concentrated or contaminated with blood. A very dilute urine may obscure significant proteinuria on dipstick examination, and proteinuria that is not predominantly albumin will be missed. This is particularly important for the detection of Bence Jones proteins in the urine of patients with multiple myeloma. Tests to measure total urine concentration accurately rely on precipitation with sulfosalicylic or trichloracetic acids. Currently, ultrasensitive dipsticks are available to measure microalbuminuria (30 to 300 mg/d), an early marker of glomerular disease that has been shown to predict glomerular injury in early diabetic nephropathy (Fig. 47-3).

The magnitude of proteinuria and the protein composition in the urine depend upon the mechanism of renal injury leading to protein losses. Large amounts of plasma proteins normally course through the glomerular capillaries but do not enter the urinary space. Both charge and size selectivity prevent virtually all of albumin, globulin, and other large-molecular-weight proteins from crossing the glomerular wall. However, if this barrier is disrupted, there can be leakage of plasma proteins into the urine (glomerular proteinuria;Fig. 47-3). Smaller proteins (<20 kDa) are freely filtered but are readily reabsorbed by the proximal tubule. Normal individuals excrete less than 150 mg/d of total protein and only about 30 mg/d of albumin. The remainder of the protein in the urine is secreted by the tubules (Tamm-Horsfall, IgA, and urokinase) or represents small amounts of filteredb$_2$-microglobulin, apoproteins, enzymes, and peptide hormones. Another mechanism of proteinuria occurs when there is excessive production of an abnormal protein that exceeds the capacity of the tubule for reabsorption. This most commonly occurs with plasma cell dyscrasias such as multiple myeloma and lymphomas that are associated with monoclonal production of immunoglobulin light chains.

The normal glomerular endothelial cell forms a barrier penetrated by pores of about 100 nm that holds back cells and other particles but offers little impediment to passage of most proteins. The glomerular basement membrane traps most large proteins (>100 kDa), while the foot processes of epithelial cells (podocytes) cover the urinary side of the glomerular basement membrane and produce a series of narrow channels (slit diaphragms) to allow molecular passage of small solutes and water (Fig. 47-4). The channels are coated with anionic glycoproteins that are rich in glutamate, aspartate, and sialic acid, which are negatively charged at physiologic pH. This negatively charged barrier impedes the passage of anionic molecules such as albumin. Some glomerular diseases, such as minimal change disease, cause fusion of glomerular epithelial cell foot processes, resulting in predominantly "selective" (Fig. 47-3) loss of albumin. Other glomerular diseases can present with disruption of the basement membrane and slit diaphragms (e.g., by immune complex deposition), resulting in large amounts of protein losses that include albumin and other plasma proteins. The fusion of foot processes causes increased pressure across the capillary basement membrane, resulting in areas with larger pore sizes. The combination of increased pressure and larger pores results in significant proteinuria ("nonselective";Fig. 47-3).

When the total daily excretion of protein exceeds 3.5 g, there is often associated hypoalbuminemia, hyperlipidemia, and edema (nephrotic syndrome;Table 47-1). However, total daily urinary protein excretion greater than 3.5 g can occur without the other features of the nephrotic syndrome in a variety of other renal diseases (Fig. 47-3). Plasma cell dyscrasias (multiple myeloma) can be associated with large amounts of excreted light chains in the urine, which may not be detected by dipstick (which detects mostly albumin). The light chains produced from these disorders are filtered by the glomerulus and overwhelm the reabsorptive capacity of the proximal tubule. A sulfosalicylic acid precipitate that is out of proportion to the dipstick estimate is suggestive of light chains (Bence Jones protein), and light chains typically redissolve upon warming of the precipitate. Renal failure from these disorders occurs through a variety of mechanisms including tubule obstruction (cast nephropathy) and light chain deposition (Chap. 275).

Hypoalbuminemia in nephrotic syndrome occurs through excessive urinary losses, increased renal catabolism, and inadequate hepatic synthesis. The resulting decrease in plasma oncotic pressure contributes to edema formation by altering the Starling forces and favoring fluid movement from capillaries to interstitium. The resulting homeostatic mechanisms designed to correct the decrease in effective intravascular volume contribute to edema formation in some patients. These mechanisms include activation of the renin-angiotensin system, antidiuretic hormone, and the sympathetic nervous system, which contribute to excessive renal salt and water reabsorption and can contribute to unrelenting edema.

The severity of edema correlates with the degree of hypoalbuminemia and is modified by other factors such as heart disease or peripheral vascular disease. The diminished plasma oncotic pressure and urinary losses of regulatory proteins appear to stimulate hepatic lipoprotein synthesis. The resulting hyperlipidemia results in lipid bodies (fatty casts, oval fat bodies) in the urine. Other proteins are lost in the urine, leading to a variety of metabolic disturbances. These include thyroxine-binding globulin, cholecalciferol-binding protein, transferrin, and metal-binding proteins. A hypercoagulable state frequently accompanies severe nephrotic syndrome due to urinary losses of antithrombin III, reduced serum levels of proteins S and C, hyperfibrinogenemia, and enhanced platelet aggregation. Some patients develop severe IgG deficiency with resulting defects in immunity. Many diseases (some listed inFig. 47-3) and drugs can cause the nephrotic syndrome, and a complete list can be found inChap. 274.

HEMATURIA, PYURIA, AND CASTS

Isolated hematuria without proteinuria, other cells, or casts is often indicative of bleeding from the urinary tract. Normal red blood cell excretion is up to 2 millionRBCs per day. Hematuria is defined as two to five RBCs per high-power field (HPF) and can be detected by dipstick. Common causes of isolated hematuria include stones, neoplasms, tuberculosis, trauma, and prostatitis. Gross hematuria with blood clots is almost never indicative of glomerular bleeding; rather, it suggests a postrenal source in the urinary collecting system. Evaluation of patients presenting with microscopic hematuria is outlined inFig. 47-2. A single urinalysis with hematuria is common and can result from menstruation, viral illness, allergy, exercise, or mild trauma. Annual urinalysis of servicemen over a 10-year period showed an incidence of 38%. However, persistent or significant hematuria (>three RBCs/HPF on three urinalyses, or single urinalysis with >100 RBCs, or gross hematuria) identified significant renal or urologic lesions in 9.1% of over 1000 patients. Even patients who are chronically anticoagulated should be investigated as outlined in Fig. 47-2. The suspicion for urogenital neoplasms in patients with isolated painless hematuria (nondysmorphic RBCs) increases with age. Neoplasms are rare in the pediatric population, and isolated hematuria is more likely to be "idiopathic" or associated with a congenital anomaly. Hematuria with pyuria and bacteriuria is typical of infection and should be treated with antibiotics after appropriate cultures. Acute cystitis or urethritis in women can cause gross hematuria. Hypercalciuria and hyperuricosuria are also risk factors for unexplained isolated hematuria in both children and adults. In some of these patients (50 to 60%), reducing calcium and uric acid excretion through dietary interventions can eliminate the microscopic hematuria.

Isolated microscopic hematuria can be a manifestation of glomerular diseases. TheRBCs of glomerular origin are often dysmorphic when examined by phase-contrast microscopy. Irregular shapes of RBCs may also occur due to pH and osmolarity changes found in the distal tubule. There is, however, significant observer variability in detecting dysmorphic RBCs, especially if a phase-contrast microscope is not available. The most common etiologies of isolated glomerular hematuria are IgA nephropathy, hereditary nephritis, and thin basement membrane disease. IgA nephropathy and hereditary nephritis can have episodic gross hematuria. A family history of renal failure is often present in patients with hereditary nephritis, and patients with thin basement membrane disease often have other family members with microscopic hematuria. A renal biopsy is needed for the definitive diagnosis of these disorders, which are discussed in more detail in Chap. 275. Hematuria with dysmorphic RBCs, RBC casts, and protein excretion >500 mg/d is virtually diagnostic of glomerulonephritis. RBC casts form as RBCs that enter the tubular fluid become trapped in a cylindrical mold of gelled Tamm-Horsfall protein. Even in the absence of azotemia, these patients should undergo serologic evaluation and renal biopsy as outlined in Fig. 47-2.

Isolated pyuria is unusual since inflammatory reactions in the kidney or collecting system are also associated with hematuria. The presence of bacteria suggests infection, and white blood cell casts with bacteria are indicative of pyelonephritis. White blood cells and/or white blood cell casts may also be seen in tubulointerstitial processes such as interstitial nephritis, systemic lupus erythematosus, and transplant rejection. In chronic renal diseases, degenerated cellular casts called *waxy casts* can be seen in the urine. *Broad casts* are thought to arise in the dilated tubules of enlarged nephrons that have undergone compensatory hypertrophy in response to reduced renal mass (i.e., chronic renal failure). A mixture of broad casts typically seen with chronic renal failure together with cellular casts andRBCs may be seen in smoldering processes such as chronic glomerulonephritis with active glomerulitis.

ABNORMALITIES OF URINE VOLUME

The volume of urine produced varies depending upon the fluid intake, renal function, and physiologic demands of the individual. See "Azotemia," above, for discussion of decreased (oliguria) or absent urine production (anuria).**The physiology of water formation and renal water conservation are discussed in Chap. 268.*

POLYURIA

By history, it is often difficult for patients to distinguish urinary frequency (often of small volumes) from polyuria, and a 24-h urine collection is needed for evaluation (Fig. 47-5). It is necessary to determine if the polyuria represents a solute or water diuresis and if the diuresis is appropriate for the clinical circumstances. The average person excretes between 600 and 800 mosmol of solutes per day, primarily as urea and electrolytes. The urine osmolality can help distinguish a solute from water diuresis. If the urine output is >3 L/d (arbitrarily defined as polyuria) and the urine is dilute (<250 mosmol/L), then total mosmol excretion is normal and a water diuresis is present. This circumstance could arise from polydipsia, inadequate secretion of vasopressin (central diabetes insipidus), or failure of renal tubules to respond to vasopressin (nephrogenic diabetes insipidus). If the urine volume is>3 L/d and urine osmolality is>300 mosmol/L, then a solute diuresis is clearly present and a search for the responsible solute(s) is mandatory.

Excessive filtration of a poorly reabsorbed solute such as glucose, mannitol, or urea can depress reabsorption of NaCl and water in the proximal tubule and lead to enhanced excretion in the urine. Poorly controlled diabetes mellitus is the most common cause of a solute diuresis, leading to volume depletion and serum hypertonicity. Since the urine Na concentration is less than that of blood, more water than Na is lost, causing hypernatremia and hypertonicity. Common iatrogenic solute diuresis occurs from mannitol administration, radiocontrast media, and high-protein feedings (enterally or parenterally), leading to increased urea production and excretion. Less commonly, excessive Na loss may occur from cystic renal diseases, Bartter's syndrome, or during the course of a tubulointerstitial process (such as resolvingATN). In these so-called salt-wasting disorders, the tubule damage results in direct impairment of Na reabsorption and indirectly reduces the responsiveness of the tubule to aldosterone. Usually, the Na losses are mild, and the obligatory urine output is less than 2 L/d (resolving ATN and postobstructive diuresis are exceptions and may be associated with significant natriuresis and polyuria.)

Formation of large volumes of dilute urine represent polydipsic states or diabetes insipidus. Primary polydipsia can result from habit, psychiatric disorders, neurologic lesions, or medications. During deliberate polydipsia, extracellular fluid volume is normal or expanded and vasopressin levels are reduced because serum osmolality tends to be near the lower limits of normal.

Central diabetes insipidus may be idiopathic in origin or secondary to a variety of hypothalamic conditions including posthypophysectomy or trauma or neoplastic, inflammatory, vascular, or infectious hypothalamic diseases. Idiopathic central diabetes insipidus is associated with selective destruction of the vasopressin-secreting neurons in the supraoptic and paraventricular nuclei and can be inherited as an autosomal dominant trait or occur spontaneously. Nephrogenic diabetes insipidus can occur in a variety of clinical situations as summarized inFig. 47-5.

A plasma vasopressin level is recommended as the best method for distinguishing between central and nephrogenic diabetes insipidus. Alternatively, a water deprivation test plus exogenous vasopressin may also distinguish primary polydipsia from central and nephrogenic diabetes insipidus.**For a detailed discussion, see Chap. 329.*

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

48. INCONTINENCE AND LOWER URINARY TRACT SYMPTOMS - *Philippe E. Zimmern*, *John D. McConnell*

PHYSIOLOGY OF VOIDING

Normal bladder filling depends on unique elastic properties of the bladder wall that allow it to increase in volume at a pressure lower than that of the bladder neck and urethra (otherwise incontinence would occur). Despite provocative maneuvers such as coughing, voluntary bladder contractions do not occur. Emptying is dependent on the integrity of a complex neuromuscular network that causes relaxation of the urethral sphincter a few milliseconds before the onset of the detrusor (bladder muscle) contraction. With normal, sustained detrusor contraction, the bladder empties completely. A bladder that can fill and empty in this manner has a normal detrusor muscle and is described as *stable* according to conventional terminology.

Since the voluntary control of micturition depends on the neural connections between the cerebral cortex and the brainstem, disruption of these pathways (brain tumor, stroke, head trauma, Parkinson's disease) impairs the ability to suppress and control bladder contractions. A bladder contraction without voluntary effort characterizes an unstable bladder. Bladder or detrusor instability of neurologic origin is termed *detrusor hyperreflexia*. Conversely, the detrusor muscle that cannot contract during voiding is called *noncontractile*; underactivity of the detrusor due to a lesion of the sacral cord or pelvic nerves is termed *detrusor areflexia*.

Contrary to common belief, the center that controls normal micturition is not in the spinal cord but in the brainstem. Proper coordination (*synergia*) between the detrusor and urethral sphincters requires an intact neural (autonomic and somatic nervous systems) communication between bladder and urethra. Injury to the upper spinal cord, for example, can cause dyssynergia between bladder and urethra that results in urge incontinence, residual urine retention, bladder wall changes (trabeculation and fibrosis), and possibly renal insufficiency.

A simple way to classify voiding dysfunction is to determine whether it is primarily a *storage failure* or an *emptying failure* by asking two questions:

Is the voiding dysfunction due to the bladder or outlet (bladder neck or urethra) (failure to store)?

Is there neurologic dysfunction (failure to empty)?

Bladder storage and emptying problems may coexist in the same individual and can cause similar lower urinary tract symptoms (LUTS).

LOWER URINARY TRACT SYMPTOMS IN MEN

The most common cause of LUTSin men of middle age and older is prostatic hyperplasia, which causes obstruction to urine flow by encroachment on the urethral lumen (Chap. 95). Histologically, 50 to 80% of the prostatic volume is composed of stromal tissue (smooth muscle), while the remainder is glandular. The transitional zone, which is responsible for benign prostatic growth, comprises 10 to 15% of the prostate at the end of puberty but increases in volume after age 40. However, prostatic enlargement is not always accompanied by symptoms because the direction of growth can be outward, so that little change may occur in urine flow. Alternatively, men with early histologic evidence of prostatic hyperplasia can experience significant voiding symptoms. In this circumstance, increased tone of the prostatic smooth muscle and enhanced prostatic tension within a nondistensible capsule can cause obstruction.

In response to obstruction, the bladder smooth-muscle cells hypertrophy to generate the higher pressures necessary for voiding, and the increase in bladder muscle mass leads to reduced elasticity, or compliance, and decreased bladder capacity. Detrusor dysfunction from bladder outlet obstruction can cause any combination of theLUTSdescribed above. When the obstruction progresses, infiltration of extracellular matrix between the smooth-muscle bundles of the bladder wall can result in a hypocontractile or acontractile bladder (bladder failure).

Other complications such as urinary tract infections or bladder stones secondary to the large postvoid residuals (stasis) and upper tract damage (hydronephrosis, reflux) can develop during the course of the obstructive process. Although prostatic hyperplasia is the most common cause of bladder outlet obstruction in men, other sources of obstruction include prostate cancer, urethral stricture, and lack of proper sphincteric relaxation (neurologic cause). Nonobstructive causes ofLUTSinclude diabetic neuropathy, which can affect the parasympathetic nerves of the bladder. Decreased sensation of bladder fullness leads to incomplete emptying and overdistention of the bladder and, in turn, to increased frequency and nocturia due to bladder overflow; these symptoms are frequently made worse by the polydipsia/polyuria of diabetes mellitus. At times, storage symptoms can be caused by other neurologic causes such as stroke, multiple sclerosis, or Parkinson's disease.

The International Prostate Symptom Score (IPSS) is used to assess the severity ofLUTS:

Decreased force of stream -- over the past month how often have you had a weak urinary stream?

Intermittency -- over the past month how often have you found you stopped and started again several times when urinating?

Incomplete emptying -- over the past month how often have you had a sensation of not emptying your bladder completely after finishing urination?

Straining -- over the past month how often have you had to push or strain to begin urination?TheIPSSalso assesses the impact of storage symptoms:

Frequency -- over the past month how often have you had to urinate again within 2 h after urinating?

Urgency -- over the past month how often have you found it difficult to postpone urination?

Nocturia -- over the past month how many times did you typically get up to urinate between going to bed and getting up in the morning? (Range: none to five or more times.)Except for nocturia, the answers range from 0 (not at all) to 5 (almost always). A total score of <8 indicates minimal voiding dysfunction; a total score of ³13 is usually required to enroll patients in drug studies for the management of benign prostatic hyperplasia (BPH); and symptom scores >23 suggest significant bladder outlet obstruction. Because similar symptoms can result from neurologic causes, theIPSSquestionnaire cannot be used to make the diagnosis of prostatic hyperplasia but is useful only as an index of severity and of the response to treatment.

LOWER URINARY TRACT SYMPTOMS IN WOMEN

Urethral obstruction is an uncommon cause ofLUTS in women. A careful bimanual examination and passage of a urethral catheter are sufficient to exclude urethral stenosis, which is usually secondary to prior instrumentation or operative procedures, and urethral cancer. Urinary tract infection (cystitis) is more prevalent in women and must be excluded by urinalysis. Multiple sclerosis should be considered in middle-aged women presenting with frequency, urgency, or incontinence. In addition to many of the same disorders that produce voiding symptoms in men, estrogen deficiency, frequency-urgency syndrome, and interstitial cystitis (IC) with minimal pain must be considered. Cystocele and pelvic prolapse can cause urinary frequency secondary to impairment of bladder emptying.

EVALUATION

Men and women withLUTS and concomitant neurologic disease should undergo a complete urodynamic evaluation. In the absence of neurologic disease, men with LUTS most commonly have prostatic hyperplasia. However, it is necessary to exclude prostate cancer, especially if there is a positive family history, an abnormal prostate examination, or an elevated level of prostate-specific antigen (PSA). In both sexes bladder cancer can also cause storage symptoms and is suggested by microscopic hematuria and/or abnormal urine cytology. Usually, a detailed genitourinary history, a symptom assessment, a careful neurologic examination including rectal examination and assessment of the bulbocavernosus reflex, measurements of urine flow and postvoid residual urine volume (by bladder ultrasound), and limited laboratory evaluation (urinalysis, urine culture, PSA levels, urine cytology, urea/creatinine levels, as indicated) should be sufficient to direct therapy. More complex investigations of the lower urinary tract (cystoscopy, voiding cystography, urodynamics) and upper urinary tract (pyelogram or ultrasonography) are sometimes indicated.**For therapy of BPH, see Chap. 95.*

INCONTINENCE

Incontinence is a condition where involuntary loss of urine is objectively demonstrated and is a social or hygienic problem. A common variant, *stress incontinence*, denotes involuntary loss of urine with physical exercise (coughing, sneezing, sports, sexual activity). *Urge incontinence* is an involuntary loss of urine associated with a strong desire to void, and *overflow incontinence* is an involuntary loss of urine when the elevation of intravesical pressure with bladder overfilling or distention exceeds the maximal urethral pressure. Loss of urine through channels other than the urethra is rare (ectopic ureter, fistulae) but causes total or continuous incontinence.

INCONTINENCE IN WOMEN

Among noninstitutionalized women 60 years of age and older, 25 to 30% have urinary incontinence daily or weekly, and approximately half of institutionalized women are incontinent more than once a day. The annual cost of caring for incontinent persons is very high and, if not well managed, can be associated with complications such as decubitus ulcers.

Stress urinary incontinence (SUI) is secondary to urethral hypermobility or, less commonly (<10%), to intrinsic sphincteric deficiency (ISD). In the continent woman the bladder neck and proximal urethra are supported by the anterior vaginal wall and its lateral attachment to the levator muscles. Anterior vaginal wall relaxation causes urethral hypermobility, usually due to aging and/or estrogen deficiency or a prior traumatic delivery or pelvic surgery. Paradoxically, women can have clinical evidence of urethral hypermobility but no stress urinary incontinence.

Some women have an anatomically normal urethra and bladder neck but still haveSUI due to damage to the internal sphincter (fixed, rigid, or "pipestem" urethra), due to prior anti-incontinence surgery, pelvic radiation or trauma, or neurologic disorders that cause denervation of the urethra. Urethral hypermobility andISD can coexist in some patients and cause persistence (or rapid recurrence) of incontinence after a simple bladder neck suspension procedure that fixes the hypermobility but leaves the sphincter untreated.

Urge incontinence can be present alone or in association withSUI(mixed incontinence). The cause of the unsuppressible or uninhibited bladder contractions is usually idiopathic, but bacterial cystitis, bladder tumor, bladder outlet obstruction, and neurogenic bladder must be excluded. Overflow incontinence is due either to bladder outlet obstruction (rare in women), an acontractile bladder (diabetic neuropathy, multiple sclerosis), excessive smooth-muscle relaxation from drugs (anticholinergic medications), or psychogenic retention.

INCONTINENCE IN MEN

In men incontinence is less common than obstruction, but urgency and urge incontinence can occur as the result of bladder outlet obstruction (as from prostatic hyperplasia) that impairs detrusor smooth-muscle function and leads to detrusor instability. Men with neurogenic bladders (diabetic neuropathy, multiple sclerosis, Parkinson's disease, stroke) can develop urge incontinence. Other causes such as bacterial cystitis or bladder tumor must be excluded.SUI in men is usually the result of distal sphincteric damage, for example, as the result of radical prostatectomy for prostate cancer.

INCONTINENCE IN THE ELDERLY

Transient urinary incontinence is common in the elderly. A mnemonic devised by

Resnick delineates its numerous causes, namely *d*elirium, *i*nfection, *a*trophic urethritis, *p*harmacologic, *p*sychological, *e*xcessive urine output (hyperglycemia, congestive heart failure), *r*estricted mobility, and *s*tool impaction (DIAPPERS). Urge incontinence is the next most common disorder in this age group and is attributed to the progressive loss of the modulating influence of the frontal lobes of the cortex on the micturition center in the brainstem.

EVALUATION

The evaluation of urinary incontinence in women should include history and quality-of-life assessment, voiding diary, physical examination including pelvic examination, urinalysis and urine culture, and measurement of postvoid residual urine volume. For patients with an unclear history or after prior pelvic or anti-incontinence procedures, evaluation may include cystoscopy, urodynamic evaluation, and imaging studies (lower and/or upper urinary tract). The history should define the onset, duration, evolution, and triggering events of leakage. Prior treatments with medications, frequent voiding schedules, and exercise regimens should be noted. Severity of incontinence is denoted by recording the type and number of pads used per day or at night and how the incontinence affects daily activities (incontinence-impact questionnaire). The amount and type of fluid consumed, sexual history (hormonal status, deliveries, venereal diseases), gastrointestinal function (fecal incontinence, constipation), and past urologic history (bed-wetting, surgeries) must also be documented. The physical examination should place special emphasis on the abdominal, genital, pelvic (associated prolapses), and neurologic systems.SUI must be demonstrated by asking the patient to cough, strain, or even stand or squat. While leakage during a cough confirms SUI, leakage after a cough is due to bladder instability (stress-induced instability). SUI in the absence of urethral hypermobility raises the suspicion of a sphincteric defect. More complex testing is needed to determine whether the urethral anatomy is normal (evaluation of urethral mobility, lateral view of the urethra on the voiding cystourethrogram, cystoscopy), whether urethral function is normal with adequate closure (leak point pressure, urethral profilometry, videourodynamics), or whether bladder function is normal (bladder volume based on home diary, filling cystometrogram).

TREATMENT

Mild stress incontinence can be treated nonoperatively with medications, estrogen replacement, or biofeedback techniques. Modalities such as urethral plugs and anterior vaginal wall prostheses are under investigation. Moderate to severe stress incontinence responds to surgical procedures aimed at supporting the anterior vaginal wall (vaginal, laparoscopic, or abdominal operations) or enhancing urethral closure when stress incontinence is secondary to internal sphincter deficiency (periurethral injection of fat or collagen, autologous or cadaveric fascial sling, synthetic sling, or insertion of an artificial urinary sphincter).

Urge incontinence responds to the management of its cause. When it is due to neurogenic or idiopathic causes, anticholinergic agents are partially effective, although side effects such as mouth dryness, blurring of vision, or constipation can limit their usefulness. Better tolerated medications are now available including slow-release oxybutinin (Ditropan XL), which is administered as 5- to 10-mg tablets once daily, and the more specific antimuscarinic agent, tolterodine (Detrol), which is usually given as 2 mg orally twice daily. Fluid restriction (which must be undertaken only with great caution) and bladder retraining with biofeedback may also be helpful. More aggressive intervention with bladder augmentation or urinary diversion are seldom necessary in the absence of neurologic disease.

BLADDER PAIN

Painful bladder disease is a general term for any bladder pathology that causes suprapubic, urethral, or pelvic pain.IC is the most common cause of bladder pain, but endometriosis, bacterial cystitis, and outlet obstruction that causes bladder instability can mimic the symptoms of IC.

INTERSTITIAL CYSTITIS

ICis a severe, chronic bladder disorder that causes frequency, nocturia, and suprapubic pain. The disorder usually affects women and is rare in blacks. Routine urine culture is uniformly negative, and the symptoms do not respond to antibiotic therapy. The etiology is probably multifactorial (Table 48-1). Current hypotheses as to etiology include autoimmune reaction against bladder antigens, deficiency in the glycosaminoglycan layer of the bladder surface allowing presumed toxins to penetrate the mucosa, mast cell infiltration and activation leading to the histamine release, and local bladder wall damage from bacteria.

The National Institutes of Health has established a series of criteria to defineICclinically (Table 48-2). The diagnosis is one of exclusion -- infection, radiation cystitis, urethral diverticula, herpes simplex, and malignancy must be excluded. Cystoscopy under anesthesia may be used for the following: (1) reveal glomerulations (submucosal vascular anomalies) or the infrequent Hunner's ulcer suggestive of IC; (2) make it possible to estimate bladder capacity (an important guide to treatment); (3) allow biopsy of the bladder wall when indicated; and (4) by bladder filling, sometimes provide therapeutic benefit with a reduction in pain level and urinary frequency up to 6 months or rarely longer.

EVALUATION

Chronic urinary frequency and bladder pain affect the quality of life to an extreme degree, though most patients experience a waxing and waning evolution; only 10% of patients have a consistent progression in symptoms. Evaluation should include a detailed history; physical examination designed to exclude neurologic and gynecologic pathology; voiding cystogram to exclude urethral defects; and urodynamic testing to eliminate a neurogenic bladder, bladder instability, or outlet obstruction and to document sensory instability. Referral to specialists may be indicated to exclude adnexal pathology, endometriosis, or bowel dysfunction or to utilize modern pain management techniques to prevent drug addiction.

TREATMENT

Empirical treatments that have been used include oral medications (amitriptyline, hydroxyzine, pentosanpolysulfate) and intravesical agents (dimethyl sulfoxide, chlorpactin, heparin). These measures may improve the urinary symptoms and occasionally reduce pain but do not modify the long-term course. Surgical intervention (augmentation cystoplasty, urinary diversion) is indicated in fewer than 5% of cases because this is a non-life-threatening, chronic disease with occasional spontaneous remissions. "Last-resort" interventions such as removal of the bladder and urethra are not a guarantee of success because some patients continue to experience pelvic pain afterwards.

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

49. FLUID AND ELECTROLYTE DISTURBANCES - *Gary G. Singer*, *Barry M. Brenner*

SODIUM AND WATER

COMPOSITION OF BODY FLUIDS

Water is the most abundant constituent in the body, comprising approximately 50% of body weight in women and 60% in men. This difference is attributable to differences in the relative proportions of adipose tissue in men and women. Total body water is distributed in two major compartments -- 55 to 75% is intracellular [intracellular fluid (ICF)], and 25 to 45% is extracellular [extracellular fluid (ECF)]. TheECF is further subdivided into intravascular (plasma water) and extravascular (interstitial) spaces in a ratio of 1:3.

The solute or particle concentration of a fluid is known as its *osmolality* and is expressed as milliosmoles per kilogram of water (mosmol/kg). Water crosses cell membranes to achieve osmotic equilibrium (ECFosmolality =ICFosmolality). The extracellular and intracellular solutes or osmoles are markedly different due to disparities in permeability and the presence of transporters and active pumps. The major ECF particles are Na^+ and its accompanying anions Cl^- and HCO_3^-, whereas K^+ and organic phosphate esters (ATP, creatine phosphate, and phospholipids) are the predominant ICF osmoles. Solutes that are restricted to the ECF or the ICF determine the *effective osmolality* (or *tonicity*) of that compartment. Since Na^+ is largely restricted to the extracellular compartment, total body Na^+content is a reflection of ECF volume. Likewise, K^+ and its attendant anions are predominantly limited to the ICF and are necessary for normal cell function. Therefore, the number of intracellular particles is relatively constant, and a change in ICF osmolality is usually due to a change in ICF water content. However, in certain situations, brain cells can vary the number of intracellular solutes in order to defend against large water shifts. This process of *osmotic adaptation* is important in the defense of cell volume and occurs in chronic hyponatremia and hypernatremia. This response is mediated initially by transcellular shifts of K^+ and Na^+, followed by synthesis, import, or export of organic solutes (so-called osmolytes) such as inositol, betaine, and glutamine. During chronic hyponatremia, brain cells lose solutes, thereby defending cell volume and diminishing neurologic symptoms. The converse occurs during chronic hypernatremia. Certain solutes, such as urea, do not contribute to water shift across cell membranes and are known as *ineffective osmoles*.

Fluid movement between the intravascular and interstitial spaces occurs across the capillary wall and is determined by the Starling forces -- capillary hydraulic pressure and colloid osmotic pressure. The transcapillary hydraulic pressure gradient exceeds the corresponding oncotic pressure gradient, thereby favoring the movement of plasma ultrafiltrate into the extravascular space. The return of fluid into the intravascular compartment occurs via lymphatic flow.

WATER BALANCE (See alsoChap. 268)

The normal plasma osmolality is 275 to 290 mosmol/kg and is kept within a narrow range by mechanisms capable of sensing a 1 to 2% change in tonicity. To maintain a steady state, water intake must equal water excretion. Disorders of water homeostasis result in hypo- or hypernatremia. Normal individuals have an obligate water loss consisting of urine, stool, and evaporation from the skin and respiratory tract. Gastrointestinal excretion is usually a minor component of total water output, except in patients with vomiting, diarrhea, or high enterostomy output states. Evaporative or insensitive water losses are important in the regulation of core body temperature. Obligatory renal water loss is mandated by the minimum solute excretion required to maintain a steady state. Normally, about 600 mosmols must be excreted per day, and since the maximal urine osmolality is 1200 mosmol/kg a minimum urine output of 500 mL/d is required for neutral solute balance.

Water Intake The primary stimulus for water ingestion is *thirst*, mediated either by an increase in effective osmolality or a decrease inECFvolume or blood pressure. *Osmoreceptors*, located in the anterolateral hypothalamus, are stimulated by a rise in tonicity. Ineffective osmoles, such as urea and glucose, do not play a role in stimulating thirst. The average osmotic threshold for thirst is approximately 295 mosmol/kg and varies among individuals. Under normal circumstances, daily water intake exceeds physiologic requirements.

Water Excretion In contrast to the ingestion of water, its excretion is tightly regulated by physiologic factors. The principal determinant of renal water excretion is *arginine vasopressin* (AVP; formerly antidiuretic hormone), a polypeptide synthesized in the supraoptic and paraventricular nuclei of the hypothalamus and secreted by the posterior pituitary gland. The binding of AVP to V_2receptors on the basolateral membrane of principal cells in the collecting duct activates adenylyl cyclase and initiates a sequence of events that leads to the insertion of water channels into the luminal membrane. These water channels that are specifically activated by AVP are encoded by the *aquaporin-2* gene (Chap. 329). The net effect is passive water reabsorption along an osmotic gradient from the lumen of the collecting duct to the hypertonic medullary interstitium. The major stimulus for AVP secretion is hypertonicity. Since the majorECFsolutes are Na^+salts, effective osmolality is primarily determined by the plasma Na^+concentration. An increase or decrease in tonicity is sensed by hypothalamic osmoreceptors as a decrease or increase in cell volume, respectively, leading to enhancement or suppression of AVP secretion. The osmotic threshold for AVP release is 280 to 290 mosmol/kg, and the system is sufficiently sensitive that plasma osmolality varies by no more than 1 to 2%.

Nonosmotic factors that regulateAVPsecretion include *effective circulating* (arterial) *volume*, nausea, pain, stress, hypoglycemia, pregnancy, and numerous drugs. The hemodynamic response is mediated by baroreceptors in the carotid sinus. The sensitivity of these receptors is significantly lower than that of the osmoreceptors. In fact, depletion of blood volume sufficient to result in a decreased mean arterial pressure is necessary to stimulate AVP release, whereas small changes in effective circulating volume have little effect. In the setting of hypovolemia, the osmotic regulation of AVP remains intact. However, the osmotic threshold, or set point, for AVP release is decreased, and the sensitivity is increased.

To maintain homeostasis and a normal plasma Na^+concentration, the ingestion of solute-free water must eventually lead to the loss of the same volume of electrolyte-free water. Three steps are required for the kidney to excrete a water load: (1) filtration and delivery of water (and electrolytes) to the diluting sites of the nephron; (2) active reabsorption of Na^+ and Cl^- without water in the thick ascending limb of the loop of Henle and, to a lesser extent, in the distal nephron; and (3) maintenance of a dilute urine due to impermeability of the collecting duct to water in the absence ofAVP. Abnormalities of any of these steps can result in impaired free water excretion, and eventual hyponatremia.

SODIUM BALANCE

Sodium is actively pumped out of cells by the Na^+,K^+-ATPase pump. As a result, 85 to 90% of all Na^+ is extracellular, and theECFvolume is a reflection of total body Na^+content. Normal volume regulatory mechanisms ensure that Na^+ loss balances Na^+gain. If this does not occur, conditions of Na^+excess or deficit ensue and are manifest as edematous or hypovolemic states, respectively. It is important to distinguish between disorders of osmoregulation and disorders of volume regulation since water and Na^+balance are regulated independently. Changes in Na^+concentration generally reflect disturbed water homeostasis, whereas alterations in Na^+content are manifest as ECF volume contraction or expansion and imply abnormal Na^+balance.

Sodium Intake Individuals eating a typical western diet consume approximately 150 mmol of NaCl daily. This normally exceeds basal requirements. As noted above, sodium is the principal extracellular cation. Therefore, dietary intake of Na^+results inECFvolume expansion, which in turn promotes enhanced renal Na^+excretion to maintain steady state Na^+balance.

Sodium Excretion (See alsoChap. 268) The regulation of Na^+excretion is multifactorial and is the major determinant of Na^+balance. A Na^+deficit or excess is manifest as a decreased or increased effective circulating volume, respectively. Changes in effective circulating volume tend to lead to parallel changes in glomerular filtration rate (GFR). However, tubule Na^+reabsorption, and not GFR, is the major regulatory mechanism controlling Na^+excretion. Almost two-thirds of filtered Na^+ is reabsorbed in the proximal convoluted tubule -- this process is electroneutral and isoosmotic. Further reabsorption (25 to 30%) occurs in the thick ascending limb of the loop of Henle via the apical *Na^+-K^+-2Cl^-cotransporter* -- this is an active process and is also electroneutral. Distal convoluted tubule reabsorption of Na^+ (5%) is mediated by the *thiazide-sensitive Na^+-Cl^-cotransporter*. Final Na^+reabsorption occurs in the cortical and medullary collecting ducts, the amount excreted being reasonably equivalent to the amount ingested per day (Chap. 268).

HYPOVOLEMIA

ETIOLOGY

True volume depletion, or hypovolemia, generally refers to a state of combined salt and water loss exceeding intake, leading toECFvolume contraction. The loss of Na^+ may be renal or extrarenal (Table 49-1).

Renal Many conditions are associated with excessive urinary NaCl and water losses, including diuretics. Pharmacologic diuretics inhibit specific pathways of Na^+reabsorption along the nephron with a consequent increase in urinary Na^+excretion. Enhanced filtration of non-reabsorbed solutes, such as glucose or urea, can also impair tubular reabsorption of Na^+ and water, leading to an osmotic or solute diuresis. This often occurs in poorly controlled diabetes mellitus and in patients receiving high-protein hyperalimentation. Mannitol is a diuretic that produces an osmotic diuresis because the renal tubule is impermeable to mannitol. Many tubule and interstitial renal disorders are associated with Na^+wasting. Excessive renal losses of Na^+ and water may also occur during the diuretic phase of acute tubular necrosis (Chap. 269) and following the relief of bilateral urinary tract obstruction. The natriuresis and water diuresis associated with these two conditions are often short-lived and an appropriate response to a state ofECFvolume expansion that ensued as a result of prior oliguria. However, ongoing losses in the absence of adequate replacement fluids may eventually lead to a state of hypovolemia. Chronic renal insufficiency is associated with a diminished ability to regulate renal salt and water excretion appropriately (Chap. 270). Therefore, patients with aGFR of less than 25 mL/min have an obligatory renal Na^+ loss that may result in progressive ECF volume depletion if Na^+intake is restricted. Finally, mineralocorticoid deficiency (hypoaldosteronism) causes salt wasting in the presence of normal intrinsic renal function.

Massive renal water excretion can also lead to hypovolemia. TheECFvolume contraction is usually less severe since two-thirds of the volume lost is intracellular. Conditions associated with excessive urinary water loss include *central diabetes insipidus* (CDI) and *nephrogenic diabetes insipidus* (NDI). These two disorders are due to impaired secretion of and renal unresponsiveness toAVP, respectively, and are discussed below.

Extrarenal Nonrenal causes of hypovolemia include fluid loss from the gastrointestinal tract, skin, and respiratory system and third space accumulations (burns, pancreatitis, peritonitis). Approximately 9 L of fluid enters the gastrointestinal tract daily, 2 L by ingestion and 7 L by secretion. Almost 98% of this volume is reabsorbed so that fecal fluid loss is only 100 to 200 mL/d. Impaired gastrointestinal reabsorption or enhanced secretion leads to volume depletion. Since gastric secretions have a low pH (high H^+concentration) and biliary, pancreatic, and intestinal secretions are alkaline (high HCO_3^-concentration), vomiting and diarrhea are often accompanied by metabolic alkalosis and acidosis, respectively.

Water evaporation from the skin and respiratory tract contributes to thermoregulation. These *insensible losses* amount to 500 mL/d. During febrile illnesses, prolonged heat exposure, or exercise, increased salt and water loss from skin, in the form of sweat, can be significant and lead to volume depletion. The Na^+concentration of sweat is normally 20 to 50 mmol/L and decreases with profuse sweating due to the action of aldosterone. Since sweat is hypotonic, the loss of water exceeds that of Na^+. The water deficit is minimized by enhanced thirst. Nevertheless, ongoing Na^+ loss is manifest as hypovolemia. Enhanced evaporative water loss from the respiratory tract may be associated with hyperventilation, especially in mechanically ventilated febrile patients.

Certain conditions lead to fluid sequestration in a *third space*. This compartment is extracellular but is not in equilibrium with either theECF or the ICF. The fluid is effectively lost from the ECF and can result in hypovolemia. Examples include the bowel lumen in gastrointestinal obstruction, subcutaneous tissues in severe burns, retroperitoneal space in acute pancreatitis, and peritoneal cavity in peritonitis. Finally, severe hemorrhage from any source can result in volume depletion.

PATHOPHYSIOLOGY

ECFvolume contraction is manifest as a decreased plasma volume and hypotension. Hypotension is due to decreased venous return (preload) and diminished cardiac output; it triggers baroreceptors in the carotid sinus and aortic arch and leads to activation of the sympathetic nervous system and the renin-angiotensin system. The net effect is to maintain mean arterial pressure and cerebral and coronary perfusion. In contrast to the cardiovascular response, the renal response is aimed at restoring the ECF volume by decreasing theGFR and filtered load of Na^+ and, most importantly, by promoting tubular reabsorption of Na^+. Increased sympathetic tone increases proximal tubular Na^+reabsorption and decreases GFR by causing preferential afferent arteriolar vasoconstriction. Sodium is also reabsorbed in the proximal convoluted tubule in response to increased angiotensin II and altered peritubular capillary hemodynamics (decreased hydraulic and increased oncotic pressure). Enhanced reabsorption of Na^+ by the collecting duct is an important component of the renal adaptation to ECF volume contraction. This occurs in response to increased *aldosterone* andAVPsecretion, and suppressed *atrial natriuretic peptide* secretion.

CLINICAL FEATURES

A careful history is often helpful in determining the etiology ofECFvolume contraction (e.g., vomiting, diarrhea, polyuria, diaphoresis). Most symptoms are nonspecific and secondary to electrolyte imbalances and tissue hypoperfusion and include fatigue, weakness, muscle cramps, thirst, and postural dizziness. More severe degrees of volume contraction can lead to end-organ ischemia manifest as oliguria, cyanosis, abdominal and chest pain, and confusion or obtundation. Diminished skin turgor and dry oral mucous membranes are poor markers of decreased interstitial fluid. Signs of intravascular volume contraction include decreased jugular venous pressure, postural hypotension, and postural tachycardia. Larger and more acute fluid losses lead to hypovolemic shock, manifest as hypotension, tachycardia, peripheral vasoconstriction, and hypoperfusion -- cyanosis, cold and clammy extremities, oliguria, and altered mental status.

DIAGNOSIS

A thorough history and physical examination are generally sufficient to diagnose the etiology of hypovolemia. Laboratory data usually confirm and support the clinical diagnosis. The blood urea nitrogen (BUN) and plasma creatinine concentrations tend to be elevated, reflecting a decreasedGFR. Normally, the BUN:creatinine ratio is about 10:1. However, in *prerenal azotemia*, hypovolemia leads to increased urea reabsorption and a proportionately greater elevation in BUN than plasma creatinine, and a BUN:creatinine ratio of 20:1 or higher. An increased BUN (relative to creatinine) may also be due to increased urea production that occurs with hyperalimentation (high-protein), glucocorticoid therapy, and gastrointestinal bleeding.

Volume depletion may be associated with hyponatremia, hypernatremia, or a normal plasma Na^+concentration, depending on the tonicity of the fluid lost, the presence of thirst, and the access to water. Hypokalemia is common in settings of increased renal or gastrointestinal K^+ loss, and hyperkalemia occurs in renal failure, adrenal insufficiency, and certain types of metabolic acidosis. Metabolic alkalosis occurs with diuretic-induced hypovolemia and in cases of vomiting or nasogastric suction. In contrast, metabolic acidosis is associated with renal failure, tubulointerstitial disorders, adrenal insufficiency, diarrhea, diabetic ketoacidosis, and lactic acidosis. Since albumin and erythrocytes are confined to the intravascular compartment,ECFvolume contraction often leads to a relative elevation in hematocrit (hemoconcentration) and plasma albumin concentration.

The appropriate response to hypovolemia is enhanced renal Na^+ and water reabsorption, which is reflected in the urine composition. Therefore, the urine Na^+concentration should usually be less than 20 mmol/L except in conditions associated with impaired Na^+reabsorption, as in acute tubular necrosis (Chap. 269). Another exception is hypovolemia due to vomiting, since the associated metabolic alkalosis and increased filtered HCO_3^-impair proximal Na^+reabsorption. In this case, the urine Cl^- is low (<20 mmol/L). The urine osmolality and specific gravity in hypovolemic subjects are generally greater than 450 mosmol/kg and 1.015, respectively, reflecting the presence of enhancedAVPsecretion. However, in hypovolemia due to diabetes insipidus, urine osmolality and specific gravity are indicative of inappropriately dilute urine.

TREATMENT

The therapeutic goals are to restore normovolemia with fluid similar in composition to that lost and to replace ongoing losses. Symptoms and signs, including weight loss, can help estimate the degree of volume contraction and should also be monitored to assess response to treatment. Mild volume contraction can usually be corrected via the oral route. More severe hypovolemia requires intravenous therapy. Isotonic or normal saline (0.9% NaCl or 154 mmol/L Na^+) is the solution of choice in normonatremic and mildly hyponatremic individuals and should be administered initially in patients with hypotension or shock. Severe hyponatremia may require hypertonic saline (3.0% NaCl or 513 mmol/L Na^+). Hypernatremia reflects a proportionally greater deficit of water than Na^+, and its correction will therefore require a hypotonic solution such as half-normal saline (0.45% NaCl or 77 mmol/L Na^+) or 5% dextrose in water. Patients with significant hemorrhage, anemia, or intravascular volume depletion may require blood transfusion or colloid-containing solutions (albumin, dextran). Hypokalemia may be present initially or may ensue as a result of increased urinary K^+excretion; it should be corrected by adding appropriate amounts of KCl to replacement solutions.

HYPONATREMIA

ETIOLOGY

A plasma Na^+concentration less than 135 mmol/L usually reflects a hypotonic state. However, plasma osmolality may be normal or increased in some cases of hyponatremia, referred to as *pseudohyponatremia*. Plasma is 93% water, the remaining 7% consisting of plasma proteins and lipids. Since Na^+ ions are dissolved in plasma water, increasing the nonaqueous phase artificially lowers the Na^+concentration measured per liter of plasma (except when Na^+-sensitive glass electrodes are used). The plasma osmolality and the Na^+concentration remain normal. This type of hyponatremia has little clinical significance, except to ascertain the cause of the hyperproteinemia or hyperlipidemia. Isotonic or slightly hypotonic hyponatremia may complicate transurethral resection of the prostate or bladder because large volumes of isoosmotic (mannitol) or hypoosmotic (sorbital or glycine) bladder irrigation solution can be absorbed and result in a dilutional hyponatremia. The metabolism of sorbitol and glycine to CO_2 and water may lead to hypotonicity if the accumulated fluid and solutes are not rapidly excreted. Hypertonic hyponatremia is usually due to hyperglycemia or, occasionally, intravenous administration of mannitol. Relative insulin deficiency causes myocytes to become impermeable to glucose. Therefore, during poorly controlled diabetes mellitus, glucose is an effective osmole and draws water from muscle cells, resulting in hyponatremia. Plasma Na^+concentration falls by 1.4 mmol/L for every 100 mg/dL rise in the plasma glucose concentration.

Most causes of hyponatremia are associated with a low plasma osmolality (Table 49-2). In general, hypotonic hyponatremia is due either to a primary water gain (and secondary Na^+loss) or a primary Na^+ loss (and secondary water gain). In the absence of water intake or hypotonic fluid replacement, hyponatremia is usually associated with hypovolemic shock due to a profound sodium deficit and transcellular water shift. Contraction of theECFvolume stimulates thirst andAVPsecretion. The increased water ingestion and impaired renal excretion result in hyponatremia. It is important to note that *diuretic-induced hyponatremia* is almost always due to thiazide diuretics. Loop diuretics decrease the tonicity of the medullary interstitium and impair maximal urinary concentrating capacity. This limits the ability of AVP to promote water retention. In contrast, thiazide diuretics lead to Na^+ and K^+depletion, and AVP-mediated water retention. In the presence of a large K^+deficit, transcellular ion exchange (K^+exits and Na^+enters cells) may contribute to hyponatremia. Hyponatremia can also occur by a process of *desalination*. This occurs when the urine tonicity (the sum of the concentrations of Na^+ and K^+) exceeds that of administered intravenous fluids (including isotonic saline). This accounts for some cases of acute postoperative hyponatremia and cerebral salt wasting after neurosurgery.

Hyponatremia in the setting ofECFvolume expansion is usually associated with edematous states, such as congestive heart failure, hepatic cirrhosis, and the nephrotic syndrome. These disorders all have in common a decreased effective circulating arterial volume, leading to increased thirst and increasedAVPlevels. Additional factors impairing the excretion of solute-free water include a reducedGFR, decreased delivery of ultrafiltrate to the diluting site (due to increased proximal fractional reabsorption of Na^+ and water), and diuretic therapy. The degree of hyponatremia often correlates with the severity of the underlying condition and is an important prognostic factor. Oliguric acute and chronic renal failure may be associated with hyponatremia if water intake exceeds the ability to excrete equivalent volumes.

Hyponatremia in the absence ofECFvolume contraction, decreased effective circulating arterial volume, or renal insufficiency is usually due to increasedAVPsecretion resulting in impaired water excretion. Ingestion or administration of water is also required since high levels of AVP alone are usually insufficient to produce hyponatremia. This disorder, commonly termed the *syndrome of inappropriate antidiuretic hormone secretion* (SIADH), is the most common cause of normovolemic hyponatremia and is due to the nonphysiologic release of AVP from the posterior pituitary or an ectopic source (Chap. 329). Renal free water excretion is impaired while the regulation of Na^+balance is unaffected. The most common causes of SIADH include neuropsychiatric and pulmonary diseases, malignant tumors, major surgery (postoperative pain), and pharmacologic agents. Severe pain and nausea are physiologic stimuli of AVP secretion; these stimuli are inappropriate in the absence of hypovolemia or hyperosmolality. A variety of central nervous system disorders may be associated with SIADH, such as meningitis, encephalitis, hemorrhage, stroke, psychosis, primary and metastatic tumors, and acute porphyria. Pneumonia, empyema, tuberculosis, and acute respiratory failure can be complicated by hyponatremia secondary to SIADH. Hypoxemia, hypercarbia, and positive-pressure ventilation are all nonosmotic stimuli for AVP release. Various tumors, notably oat cell carcinoma of the lung, have been demonstrated to secrete AVP ectopically. Many drugs either stimulate AVP release or potentiate its actions on the kidney. The pattern of AVP secretion can be used to classify SIADH into four subtypes: (1) erratic autonomous AVP secretion (ectopic production); (2) normal regulation of AVP release around a lower osmolality set point or *reset osmostat* (cachexia, malnutrition); (3) normal AVP response to hypertonicity with failure to suppress completely at low osmolality (incomplete pituitary stalk section); and (4) normal AVP secretion with increased sensitivity to its actions or secretion of some other antidiuretic factor (rare).

Hormonal excess or deficiency may cause hyponatremia. Adrenal insufficiency (Chap. 331) and hypothyroidism (Chap. 330) may present with hyponatremia and should not be confused withSIADH. Although decreased mineralocorticoids may contribute to the hyponatremia of adrenal insufficiency, it is the cortisol deficiency that leads to hypersecretion ofAVP both indirectly (secondary to volume depletion) and directly (cosecreted with corticotropin-releasing factor). The mechanisms by which hypothyroidism leads to hyponatremia include decreased cardiac output andGFR and increased AVP secretion in response to hemodynamic stimuli.

Finally, hyponatremia may occur in the absence ofAVP or renal failure if the kidney is unable to excrete the dietary water load. In psychogenic or primary polydipsia, compulsive water consumption may overwhelm the normally large renal excretory capacity of 12 L/d (Chap. 329). These patients often have psychiatric illnesses and may be taking medications, such as phenothiazines, that enhance the sensation of thirst by causing a dry mouth. The maximal urine output is a function of the minimum urine osmolality achievable and the mandatory solute excretion. Metabolism of a normal diet generates about 600 mosmol/d, and the minimum urine osmolality in humans is 50 mosmol/kg. Therefore, the maximum daily urine output will be about 12 L (600 ¸ 50= 12). A solute excretion rate of greater than ~750 mosmol/d is, by definition, an *osmotic diuresis*. A low-protein diet may yield as few as 250 mosmol/d, which translates into a maximal urine output of 5 L/d at a minimum urine tonicity of 50 mosmol/kg. Beer drinkers typically have a poor dietary intake of protein and electrolytes and consume large volumes (of beer), which may exceed the renal excretory capacity and result in hyponatremia. This phenomenon is referred to as *beer potomania*.

CLINICAL FEATURES

The clinical manifestations of hyponatremia are related to osmotic water shift leading to increasedICFvolume, specifically brain cell swelling or cerebral edema. Therefore, the symptoms are primarily neurologic, and their severity is dependent on the rapidity of onset and absolute decrease in plasma Na^+concentration. Patients may be asymptomatic or complain of nausea and malaise. As the plasma Na^+concentration falls, the symptoms progress to include headache, lethargy, confusion, and obtundation. Stupor, seizures, and coma do not usually occur unless the plasma Na^+concentration falls acutely below 120 mmol/L or decreases rapidly. As described above, adaptive mechanisms designed to protect cell volume occur in chronic hyponatremia. Loss of Na^+ and K^+, followed by organic osmolytes, from brain cells decreases brain swelling due to secondary transcellular water shifts (from ICF toECF). The net effect is to minimize cerebral edema and its symptoms. Hospitalized patients with hyponatremia have an increased mortality rate compared to normonatremic control subjects. However, the excess mortality is usually attributed to the underlying disorder rather than the electrolyte disturbance.

DIAGNOSIS

Hyponatremia is not a disease but a manifestation of a variety of disorders. The underlying cause can often be ascertained from an accurate history and physical examination, including an assessment ofECFvolume status and effective circulating arterial volume. The differential diagnosis of hyponatremia, an expanded ECF volume, and decreased effective circulating volume includes congestive heart failure, hepatic cirrhosis, and the nephrotic syndrome. Hypothyroidism and adrenal insufficiency tend to present with a near-normal ECF volume and decreased effective circulating arterial volume. All of these diseases have characteristic signs and symptoms. Patients withSIADHare usually euvolemic.

Four laboratory findings often provide useful information and can narrow the differential diagnosis of hyponatremia: (1) the plasma osmolality, (2) the urine osmolality, (3) the urine Na^+concentration, and (4) the urine K^+concentration. SinceECFtonicity is determined primarily by the Na^+concentration, most patients with hyponatremia have a decreased plasma osmolality. If the plasma osmolality is not low, pseudohyponatremia must be ruled out. The appropriate renal response to hypoosmolality is to excrete the maximum volume of dilute urine, i.e., urine osmolality and specific gravity of less than 100 mosmol/kg and 1.003, respectively. This occurs in patients with primary polydipsia. If this is not present, it suggests impaired free water excretion due to the action ofAVP on the kidney. The secretion of AVP may be a physiologic response to hemodynamic stimuli or it may be inappropriate in the presence of hyponatremia and euvolemia. Since Na^+ is the major ECF cation and is largely restricted to this compartment, ECF volume contraction represents a deficit in total body Na^+content. Therefore, volume depletion in patients with normal underlying renal function results in enhanced tubule Na^+reabsorption and a urine Na^+concentration less than 20 mmol/L. The finding of a urine Na^+concentration greater than 20 mmol/L in hypovolemic hyponatremia implies a salt-wasting nephropathy, diuretic therapy, hypoaldosteronism, or occasionally vomiting. Both the urine osmolality and the urine Na^+concentration can be followed serially when assessing response to therapy.

SIADHis characterized by hypoosmotic hyponatremia in the setting of an inappropriately concentrated urine (urine osmolality greater than 100 mosmol/kg). Patients are typically normovolemic and have normal Na^+balance. They tend to be mildly volume expanded secondary to water retention and have a urine Na^+excretion rate equal to intake (urine Na^+concentration usually greater than 40 mmol/L). By definition, they have normal renal, adrenal, and thyroid function and usually have normal K^+ and acid-base balance. SIADH is often associated with hypouricemia due to the uricosuric state induced by volume expansion. In contrast, hypovolemic patients tend to be hyperuricemic secondary to increased proximal urate reabsorption.

CLINICAL APPROACH

SeeFig. 49-1.

TREATMENT

The goals of therapy are twofold: (1) to raise the plasma Na^+concentration by restricting water intake and promoting water loss; and (2) to correct the underlying disorder. Mild asymptomatic hyponatremia is generally of little clinical significance and requires no treatment. The management of asymptomatic hyponatremia associated with ECFvolume contaction should include Na^+repletion, generally in the form of isotonic saline. The direct effect of the administered NaCl on the plasma Na^+concentration is trivial. However, restoration of euvolemia removes the hemodynamic stimulus forAVPrelease, allowing the excess free water to be excreted. The hyponatremia associated with edematous states tends to reflect the severity of the underlying disease and is usually asymptomatic. These patients have increased total body water that exceeds the increase in total body Na^+content. Treatment should include restriction of Na^+ and water intake, correction of hypokalemia, and promotion of water loss in excess of Na^+. The latter may require the use of loop diuretics with replacement of a proportion of the urinary Na^+ loss to ensure net free water excretion. Dietary water restriction should be less than the urine output. Correction of the K^+deficit may raise the plasma Na^+concentration by favoring a shift of Na^+ out of cells as K^+ moves in. Water restriction is also a component of the therapeutic approach to hyponatremia associated with primary polydipsia, renal failure, andSIADH(Chap. 329).

The rate of correction of hyponatremia depends on the absence or presence of neurologic dysfunction. This, in turn, is related to the rapidity of onset and magnitude of the fall in plasma Na^+concentration. In asymptomatic patients, the plasma Na^+concentration should be raised by no more than 0.5 to 1.0 mmol/L per hour and by less than 10 to 12 mmol/L over the first 24 h. Acute or severe hyponatremia (plasma Na^+concentration <110 to 115 mmol/L) tends to present with altered mental status and/or seizures and requires more rapid correction. Severe symptomatic hyponatremia should be treated with hypertonic saline, and the plasma Na^+concentration should be raised by 1 to 2 mmol/L per hour for the first 3 to 4 h or until the seizures subside. Once again, the plasma Na^+concentration should probably be raised by no more than 12 mmol/L during the first 24 h. The quantity of Na^+required to increase the plasma Na^+concentration by a given amount can be estimated by multiplying the deficit in plasma Na^+concentration by the total body water. Under normal conditions, total body water is 50 or 60% of lean body weight in women or men, respectively. Therefore, to raise the plasma Na^+concentration from 105 to 115 mmol/L in a 70-kg man requires 420 mmol [(115- 105) ´ 70´ 0.6] of Na^+. The risk of correcting hyponatremia too rapidly is the development of the *osmotic demyelination syndrome* (ODS). This is a neurologic disorder characterized by flaccid paralysis, dysarthria, and dysphagia. The diagnosis is usually suspected clinically and can be confirmed by appropriate neuroimaging studies. There is no specific treatment for the disorder, which is associated with significant morbidity and mortality. Patients with chronic hyponatremia are most susceptible to the development of ODS, since their brain cell volume has returned to near normal as a result of the osmotic adaptive mechanisms described above. Therefore, administration of hypertonic saline to these individuals can cause sudden osmotic shrinkage of brain cells. In addition to rapid or overcorrection of hyponatremia, risk factors for ODS include prior cerebral anoxic injury, hypokalemia, and malnutrition, especially secondary to alcoholism. Water restriction in primary polydipsia and intravenous saline therapy inECFvolume-contracted patients may also lead to overly rapid correction of hyponatremia as a result of AVP suppression and a brisk water diuresis. This can be prevented by administration of water or use of an AVP analogue to slow down the rate of free water excretion.**For further discussion, see Chap. 329.*

HYPERNATREMIA

ETIOLOGY

Hypernatremia is defined as a plasma Na^+concentration greater than 145 mmol/L. Since Na^+ and its accompanying anions are the major effectiveECFosmoles, hypernatremia is a state of hyperosmolality. As a result of the fixed number ofICFparticles, maintenance of osmotic equilibrium in hypernatremia results in ICF volume contraction. Hypernatremia may be due to primary Na^+ gain or water deficit. The two components of an appropriate response to hypernatremia are increased water intake stimulated by thirst and the excretion of the minimum volume of maximally concentrated urine reflectingAVPsecretion in response to an osmotic stimulus.

In practice, the majority of cases of hypernatremia result from the loss of water. Since water is distributed between the ICFand theECF in a 2:1 ratio, a given amount of solute-free water loss will result in a twofold greater reduction in the ICF compartment than the ECF compartment. For example, consider three scenarios: the loss of 1 L of water, isotonic NaCl, or half-isotonic NaCl. If 1 L of water is lost, the ICF volume will decrease by 667 mL, whereas the ECF volume will fall by only 333 mL. Due to the fact that Na^+ is largely restricted to the ECF, this compartment will decrease by 1 L if the fluid lost is isoosmotic. One liter of half-isotonic NaCl is equivalent to 500 mL of water (one-third ECF, two-thirds ICF) plus 500 mL of isotonic saline (all ECF). Therefore, the loss of 1 L of half-isotonic saline decreases the ECF and ICF volumes by 667 mL and 333 mL, respectively.

The degree of hyperosmolality is typically mild unless the thirst mechanism is abnormal or access to water is limited. The latter occurs in infants, the physically handicapped, patients with impaired mental status, in the postoperative state, and in intubated patients in the intensive care unit. On rare occasions, impaired thirst may be due to *primary hypodipsia*. This usually occurs as a result of damage to the hypothalamic osmoreceptors that control thirst and tends to be associated with abnormal osmotic regulation ofAVPsecretion. Primary hypodipsia may be due to a variety of pathologic changes including granulomatous disease, vascular occlusion, and tumors. A subset of hypodipsic hypernatremia, referred to as *essential hypernatremia*, does not respond to forced water intake. This appears to be due to a specific osmoreceptor defect resulting in nonosmotic regulation of AVP release. Thus, the hemodynamic effects of water loading lead to AVP suppression and excretion of dilute urine.

The source of free water loss is either renal or extrarenal. Nonrenal loss of water may be due to evaporation from the skin and respiratory tract (insensible losses) or loss from the gastrointestinal tract. Insensible losses are increased with fever, exercise, heat exposure, and severe burns and in mechanically ventilated patients. Furthermore, the Na^+concentration of sweat decreases with profuse perspiration, thereby increasing solute-free water loss. Diarrhea is the most common gastrointestinal cause of hypernatremia. Specifically, osmotic diarrheas (induced by lactulose, sorbitol, or malabsorption of carbohydrate) and viral gastroenteritides result in water loss exceeding that of Na^+ and K^+. In contrast, secretory diarrheas (e.g., cholera, carcinoid, VIPoma) have a fecal osmolality (twice the sum of the concentrations of Na^+ and K^+) similar to that of plasma and present withECFvolume contraction and a normal plasma Na^+concentration or hyponatremia.

Renal water loss is the most common cause of hypernatremia and is due to drug-induced or osmotic diuresis or diabetes insipidus (Chap. 329). Loop diuretics interfere with the countercurrent mechanism and produce an isoosmotic solute diuresis. This results in a decreased medullary interstitial tonicity and impaired renal concentrating ability. The presence of non-reabsorbed organic solutes in the tubule lumen impairs the osmotic reabsorption of water. This leads to water loss in excess of Na^+ and K^+, known as an osmotic diuresis. The most frequent cause of an osmotic diuresis is hyperglycemia and glucosuria in poorly controlled diabetes mellitus. Intravenous administration of mannitol and increased endogenous production of urea (high-protein diet) can also result in an osmotic diuresis. Hypernatremia secondary to nonosmotic urinary water loss is usually due to: (1)CDI or neurogenic diabetes insipidus characterized by impairedAVPsecretion, or (2)NDIresulting from end-organ (renal) resistance to the actions of AVP. The most common cause of CDI is destruction of the neurohypophysis. This may occur as a result of trauma, neurosurgery, granulomatous disease, neoplasms, vascular accidents, or infection. In many cases, CDI is idiopathic and may occasionally be hereditary. The familial form of the disease is inherited in an autosomal dominant fashion and has been attributed to mutations in the propressophysin (AVP precursor) gene. NDI may be either inherited or acquired. Congenital NDI is an X-linked recessive trait due to mutations in the V_2receptor gene. Mutations in the autosomal aquaporin-2 gene may also result in NDI. The aquaporin-2 gene encodes the water channel protein whose membrane insertion is stimulated by AVP. The causes of sporadic NDI are numerous and include drugs (especially lithium), hypercalcemia, hypokalemia, and conditions that impair medullary hypertonicity (e.g., papillary necrosis or osmotic diuresis). Pregnant women, in the second or third trimester, may develop NDI as a result of excessive elaboration of vasopressinase by the placenta.

Finally, although infrequent, a primary Na^+ gain may cause hypernatremia. For example, inadvertent administration of hypertonic NaCl or $NaHCO_3$or replacing sugar with salt in infant formula can produce this complication.

CLINICAL FEATURES

As a consequence of hypertonicity, water shifts out of cells, leading to a contractedICFvolume. A decreased brain cell volume is associated with an increased risk of subarachnoid or intracerebral hemorrhage. Hence, the major symptoms of hypernatremia are neurologic and include altered mental status, weakness, neuromuscular irritability, focal neurologic deficits, and occasionally coma or seizures. Patients may also complain of polyuria or thirst. For unknown reasons, patients with polydipsia fromCDI tend to prefer ice-cold water. The signs and symptoms of volume depletion are often present in patients with a history of excessive sweating, diarrhea, or an osmotic diuresis. The mortality rate associated with a plasma Na^+concentration greater than 180 mmol/L is very high. As with hyponatremia, the severity of the clinical manifestations is related to the acuity and magnitude of the rise in plasma Na^+concentration. Chronic hypernatremia is generally less symptomatic as a result of adaptive mechanisms designed to defend cell volume. Brain cells initially take up Na^+ and K^+salts, later followed by accumulation of organic osmolytes such as inositol. This serves to restore the brain ICF volume towards normal.

DIAGNOSIS

A complete history and physical examination will often provide clues as to the underlying cause of hypernatremia. Relevant symptoms and signs include the absence or presence of thirst, diaphoresis, diarrhea, polyuria, and the features of ECFvolume contraction. The history should include a list of current and recent medications, and the physical examination is incomplete without a thorough mental status and neurologic assessment. Measurement of urine volume and osmolality are essential in the evaluation of hyperosmolality. The appropriate renal response to hypernatremia is the excretion of the minimum volume (500 mL/d) of maximally concentrated urine (urine osmolality>800 mosmol/kg). These findings suggest extrarenal or remote renal water loss or administration of hypertonic Na^+ salt solutions. The presence of a primary Na^+excess can be confirmed by the presence of ECF volume expansion and natriuresis (urine Na^+concentration usually >100 mmol/L). Many causes of hypernatremia are associated with polyuria and a submaximal urine osmolality. The product of the urine volume and osmolality, i.e., the solute excretion rate, is helpful in determining the basis of the polyuria (see above). To maintain a steady state, total solute excretion must equal solute production. As stated above, individuals eating a normal diet generate ~600 mosmol/d. Therefore, daily solute excretion in excess of 750 mosmol defines an osmotic diuresis. This can be confirmed by measuring the urine glucose and urea. In general, bothCDI andNDIpresent with polyuria and hypotonic urine (urine osmolality<250 mosmol/kg). The degree of hypernatremia is usually mild unless there is an associated thirst abnormality. The clinical history, physical examination, and pertinent laboratory data can often rule out causes of acquired NDI. CDI and NDI can generally be distinguished by administering theAVPanalogue desmopressin (10 ug intranasally) after careful water restriction. The urine osmolality should increase by at least 50% in CDI and will not change in NDI. Unfortunately, the diagnosis may sometimes be difficult due to partial defects in AVP secretion and action.

CLINICAL APPROACH

SeeFig. 49-2.

TREATMENT

The therapeutic goals are to stop ongoing water loss by treating the underlying cause and to correct the water deficit. TheECFvolume should be restored in hypovolemic patients. The quantity of water required to correct the deficit can be calculated from the following equation:

In hypernatremia due to water loss, total body water is approximately 50 and 40% of lean body weight in men and women, respectively. For example, a 50-kg woman with a plasma Na$^+$concentration of 160 mmol/L has an estimated free water deficit of 2.9 L lcub;[(160 - 140) ¸140] ´ (0.4 ´50)rcub;. As in hyponatremia, rapid correction of hypernatremia is potentially dangerous. In this case, a sudden decrease in osmolality could potentially cause a rapid shift of water into cells that have undergone osmotic adaptation. This would result in swollen brain cells and increase the risk of seizures or permanent neurologic damage. Therefore, the water deficit should be corrected slowly over at least 48 to 72 h. When calculating the rate of water replacement, ongoing losses should be taken into account, and the plasma Na$^+$concentration should be lowered by 0.5 mmol/L per hour and by no more than 12 mmol/L over the first 24 h. The safest route of administration of water is by mouth or via a nasogastric tube (or other feeding tube). Alternatively, 5% dextrose in water or half-isotonic saline can be given intravenously. The appropriate treatment ofCDIconsists of administering desmopressin intranasally (Chap. 329). Other options for decreasing urine output include a low-salt diet in combination with low-dose thiazide diuretic therapy. In some patients with partial CDI, drugs that either stimulateAVPsecretion or enhance its action on the kidney have been useful. These include chlorpropamide, clofibrate, carbamazepine, andnonsteroidal anti-inflammatory drugs (NSAIDs). The concentrating defect inNDI may be reversible by treating the underlying disorder or eliminating the offending drug. Symptomatic polyuria due to NDI can be treated with a low-Na$^+$diet and thiazide diuretics as described above. This induces mild volume depletion, which leads to enhanced proximal reabsorption of salt and water and decreased delivery to the site of action of AVP, the collecting duct. By impairing renal prostaglandin synthesis, NSAIDs potentiate AVP action and thereby increase urine osmolality and decrease urine volume. Amiloride may be useful in patients with NDI who need to be on lithium. The nephrotoxicity of lithium requires the drug to be taken up into collecting duct cells via the amiloride-sensitive Na$^+$channel.

POTASSIUM

POTASSIUM BALANCE

Potassium is the major intracellular cation. The normal plasma K$^+$concentration is 3.5 to 5.0 mmol/L, whereas that inside cells is about 150 mmol/L. Therefore, the amount of K$^+$ in the ECF(30 to 70 mmol) constitutes less than 2% of the total body K$^+$content (2500 to 4500 mmol). The ratio ofICF to ECF K$^+$concentration (normally 38:1) is the principal result of the resting membrane potential and is crucial for normal neuromuscular function. The basolateral Na^+, K^+-ATPase pump actively transports K^+ in and Na^+out of the cell in a 2:3 ratio, and the passive outward diffusion of K^+ is quantitatively the most important factor that generates the resting membrane potential. The activity of the electrogenic Na^+, K^+-ATPase pump may be stimulated as a result of an increased intracellular Na^+concentration and inhibited in the setting of digoxin toxicity or chronic illness such as heart failure or renal failure.

The distribution of K^+ is also affected by several other factors, including hormones, acid-base balance, osmolality, and cell turnover. Insulin increases Na^+, K^+-ATPase activity indirectly and independent of its effect on glucose transport, leading to K^+shift into muscle and liver cells. Conversely, insulin deficiency results in K^+movement from theICF to the ECFcompartment. Catecholamines have variable effects on K^+distribution -- b_2-adrenergic agonists promote whereas a-adrenergic agonists impair K^+uptake by cells. The Na^+, K^+-ATPase pump as well as insulin secretion are stimulated byb_2-adrenergic agonists. In contrast, a-adrenergic agonists have the opposite effect. The major action of aldosterone is to increase K^+excretion (see below). The role of extracellular pH in K^+balance relates to the underlying acid-base disorder. In metabolic acidosis, 60% of the H^+ load is buffered inside cells. To maintain electroneutrality, the H^+ ion must either be accompanied by an anion or exchanged for intracellular K^+(leading to hyperkalemia). Organic acidoses are not usually associated with a pH-related K^+shift, since anions such as lactate andb-hydroxybutyrate can be readily taken up by the cell. The converse, movement of K^+ into cells, may be seen with metabolic alkalosis. However, this is less important due to diminished intracellular buffering. Primary respiratory disturbances in acid-base balance result in minimal transcellular K^+shifts. In hyperosmolal states, K^+diffuses out of cells along with water due to *solvent drag*. The concentration gradient favoring K^+movement out of cells is also increased as a result of ICF water loss. Tissue destruction or breakdown results in the release of intracellular K^+, whereas the production of new cells shifts K^+ out of the ECF. Finally, moderate to severe exercise may be associated with K^+release from muscle, leading to glycogenolysis and local vasodilatation. This is usually transient but may affect the plasma K^+concentration if patients repeatedly clench and unclench their fist prior to venipuncture.

The K^+intake of individuals on an average western diet is 40 to 120 mmol/d or approximately 1 mmol/kg per day, 90% of which is absorbed by the gastrointestinal tract. Maintenance of the steady state necessitates matching K^+ingestion with excretion. Initially, extrarenal adaptive mechanisms, followed later by urinary excretion, prevent a doubling of the plasma K^+concentration that would occur if the dietary K^+ load remained in theECFcompartment. Immediately following a meal, most of the absorbed K^+enters cells as a result of the initial elevation in the plasma K^+concentration and facilitated by insulin release and basal catecholamine levels. Eventually, however, the excess K^+ is excreted in the urine (see below). The regulation of gastrointestinal K^+handling is not well understood. The amount of K^+ lost in the stool can increase from 10 to 50 or 60% (of dietary intake) in chronic renal insufficiency. In addition, colonic secretion of K^+ is stimulated in patients with large volumes of diarrhea, resulting in potentially severe K^+depletion.

POTASSIUM EXCRETION (See alsoChap. 268)

Renal excretion is the major route of elimination of dietary and other sources of excess K^+. The filtered load of K^+(GFR´ plasma K^+concentration = 180 L/d´ 4 mmol/L = 720 mmol/d) is 10- to 20-fold greater than theECFK^+content. Some 90% of filtered K^+ is reabsorbed by the proximal convoluted tubule and loop of Henle. Proximally, K^+ is reabsorbed passively with Na^+ and water, whereas the luminal Na^+-K^+-2Cl^-cotransporter mediates K^+uptake in the thick ascending limb of the loop of Henle. Therefore, K^+delivery to the distal nephron [distal convoluted tubule and cortical collecting duct (CCD)] approximates dietary intake. Net distal K^+secretion or reabsorption occurs in the setting of K^+excess or depletion, respectively. The cell responsible for K^+secretion in the late distal convoluted tubule (or connecting tubule) and CCD is the principal cell. Virtually all regulation of renal K^+excretion and total body K^+balance occurs in the distal nephron. The driving force for K^+secretion is a favorable electrochemical gradient across the luminal membrane of the principal cell. As a result of the action of the basolateral Na^+, K^+-ATPase pump, the intracellular K^+concentration far exceeds that of the fluid in the lumen of the CCD. The electrical gradient is created by electrogenic Na^+reabsorption leading to a lumen-negative transepithelial potential difference (TEPD), favoring K^+secretion. The generation of a lumen-negative TEPD depends on the relative rates of reabsorption of Na^+ and its accompanying anion (primarily Cl^-). Equimolar reabsorption of Na^+ and Cl^- at equivalent rates is electroneutral, whereas reabsorption of Na^+ in excess of Cl^- is electrogenic. The cellular uptake of Na^+ by the principal cell occurs via an apical Na^+channel and is driven by a low intracellular Na^+concentration relative to that in the lumen of the CCD. The mechanism and regulation of distal nephron Cl^- transport is less clear. Obviously, factors that impact on either Na^+ or Cl^- reabsorption by the principal cell will influence the TEPD. Potassium secretion is regulated by two physiologic stimuli -- aldosterone and hyperkalemia. Aldosterone is secreted by the zona glomerulosa cells of the adrenal cortex in response to high renin and angiotensin II or hyperkalemia. The actions of aldosterone on the principal cell include enhanced apical membrane Na^+conductivity, stimulation of the basolateral Na^+, K^+-ATPase, and increased luminal K^+channels. The plasma K^+concentration, independent of aldosterone, can directly affect K^+secretion. In addition to the K^+concentration in the lumen of the CCD, renal K^+ loss depends on the urine flow rate, a function of daily solute excretion (see above). Since excretion is equal to the product of concentration and volume, increased distal flow rate can significantly enhance urinary K^+output. Finally, in severe K^+depletion, secretion of K^+ is reduced and reabsorption, via apical H^+, K^+-ATPase pumps in cortical and medullary collecting ducts, is upregulated.

HYPOKALEMIA

ETIOLOGY (See Table 49-3)

Hypokalemia, defined as a plasma K^+concentration < 3.5 mmol/L, may result from one (or more) of the following: decreased net intake, shift into cells, or increased net loss. Diminished intake is seldom the sole cause of K^+depletion since urinary excretion can be effectively decreased to less than 15 mmol/d as a result of net K^+reabsorption in the distal nephron. With the exception of the urban poor and certain cultural groups, the amount of K^+ in the diet almost always exceeds that excreted in the urine. However, dietary K^+restriction may exacerbate the hypokalemia secondary to increased gastrointestinal or renal loss. An unusual cause of decreased K^+intake is ingestion of clay (geophagia), which binds dietary K^+ and iron. This custom was previously common among African Americans in the American South.

Redistribution into Cells Movement of K^+ into cells may transiently decrease the plasma K^+concentration without altering total body K^+content. For any given cause, the magnitude of the change is relatively small, often less than 1 mmol/L. However, a combination of factors may lead to a significant fall in the plasma K^+concentration and may amplify the hypokalemia due to K^+wasting. Alkalosis, especially that due to a primary increase in plasma HCO_3^-(metabolic alkalosis), is often associated with hypokalemia. This occurs as a result of K^+redistribution as well as excessive renal K^+ loss. Treatment of diabetic ketoacidosis with insulin may lead to hypokalemia due to stimulation of the Na^+-H^+antiporter and (secondarily) the Na^+, K^+-ATPase pump. Furthermore, uncontrolled hyperglycemia often leads to K^+depletion from an osmotic diuresis (see below). Stress-induced catecholamine release and administration ofb_2-adrenergic agonists directly induce cellular uptake of K^+ and promote insulin secretion by pancreatic islet b cells. *Hypokalemic periodic paralysis* is a rare condition characterized by recurrent episodic weakness or paralysis (Chap. 381). Since K^+ is the majorICFcation, anabolic states can potentially result in hypokalemia due to a K^+ shift into cells. This may occur following rapid cell growth seen in patients with pernicious anemia treated with vitamin B_{12} or with neutropenia after treatment with granulocyte-macrophage colony stimulating factor. Massive transfusion with thawed washed red blood cells (RBCs) could cause hypokalemia since frozen RBCs lose up to half of their K^+during storage.

Nonrenal Loss of Potassium Excessive sweating may result in K^+depletion from increased integumentary and renal K^+ loss. Hyperaldosteronism, secondary toECFvolume contraction, enhances K^+excretion in the urine (Chap. 331). Normally, K^+ lost in the stool amounts to 5 to 10 mmol/d in a volume of 100 to 200 mL. Hypokalemia subsequent to increased gastrointestinal loss can occur in patients with profuse diarrhea (usually secretory), villous adenomas, VIPomas, or laxative abuse. However, the loss of gastric secretions does not account for the moderate to severe K^+depletion often associated with vomiting or nasogastric suction. Since the K^+concentration of gastric fluid is 5 to 10 mmol/L, it would take 30 to 80 L of vomitus to achieve a K^+deficit of 300 to 400 mmol typically seen in these patients. In fact, the hypokalemia is primarily due to increased renal K^+excretion. Loss of gastric contents results in volume depletion and metabolic alkalosis, both of which promote kaliuresis. Hypovolemia stimulates aldosterone release, which augments K^+secretion by the principal cells. In addition, the filtered load of HCO_3^-exceeds the reabsorptive capacity of the proximal convoluted tubule, thereby increasing distal delivery of $NaHCO_3$, which enhances the electrochemical gradient favoring K^+ loss in the urine.

Renal Loss of Potassium In general, most cases of chronic hypokalemia are due to renal K^+wasting. This may be due to factors that increase the K^+concentration in the lumen of theCCD or augment distal flow rate. As described above, distal nephron K^+secretion is driven by a lumen-negativeTEPD, affected by aldosterone and the relative rates of reabsorption of Na^+ and its accompanying anion(s). Mineralocorticoid excess commonly results in hypokalemia (Chap. 331). *Primary hyperaldosteronism* is due to dysregulated aldosterone secretion by an adrenal adenoma (Conn's syndrome) or carcinoma or to adrenocortical hyperplasia. In a rare subset of patients, the disorder is familial (autosomal dominant) and aldosterone levels can be suppressed by administering low doses of exogenous glucocorticoid. The molecular defect responsible for *glucocorticoid-remediable hyperaldosteronism* is a rearranged gene (due to a chromosomal crossover), containing the 5¢-regulatory region of the 11b-hydroxylase gene and the coding sequence of the aldosterone synthase gene. Consequently, mineralocorticoid is synthesized in the zona fasciculata and regulated by corticotropin. A number of conditions associated with hyperreninemia result in secondary hyperaldosteronism and renal K^+wasting. High renin levels are commonly seen in both renovascular and malignant hypertension. Renin-secreting tumors of the juxtaglomerular apparatus are a rare cause of hypokalemia. Other tumors that have been reported to produce renin include renal cell carcinoma, ovarian carcinoma, and Wilms' tumor. Hyperreninemia may also occur secondary to decreased effective circulating arterial volume.

In the absence of elevated renin or aldosterone levels, enhanced distal nephron secretion of K^+ may result from increased production of non-aldosterone mineralocorticoids in *congenital adrenal hyperplasia* (Chap. 331). Glucocorticoid-stimulated kaliuresis does not normally occur due to the conversion of cortisol to cortisone by 11b-hydroxysteroid dehydrogenase (11b-HSDH). Therefore, 11b-HSDH deficiency or suppression allows cortisol to bind to the aldosterone receptor and leads to the *syndrome of apparent mineralocorticoid excess*. Drugs that inhibit the activity of 11b-HSDH include glycyrrhetinic acid, present in licorice, chewing tobacco, and carbenoxolone. The presentation of Cushing's syndrome may include hypokalemia if the capacity of 11b-HSDH to inactivate cortisol is overwhelmed by persistently elevated glucocorticoid levels.

Liddle's syndrome is a rare familial (autosomal dominant) disease characterized by hypertension, hypokalemic metabolic alkalosis, renal K^+wasting, and suppressed renin and aldosterone secretion (Chap. 331). Increased distal delivery of Na^+ with a non-reabsorbable anion (not Cl^-) enhances the lumen-negativeTEPD and K^+secretion. Classically, this is seen with *proximal (type 2) renal tubular acidosis* (RTA) and vomiting, associated with bicarbonaturia. Diabetic ketoacidosis and toluene abuse (glue-sniffing) can lead to increased delivery of b-hydroxybutyrate and hippurate, respectively, to theCCD and to renal K^+ loss. High doses of penicillin derivatives administered to volume-depleted patients may likewise promote renal K^+secretion as well as an osmotic diuresis. *Classic distal (type 1) RTA* is associated with hypokalemia due to increased renal K^+ loss, the mechanism of which is uncertain. Amphotericin B causes hypokalemia due to increased distal nephron permeability to Na^+ and K^+ and to renal K^+wasting.

Bartter's syndrome is a disorder characterized by hypokalemia, metabolic alkalosis, hyperreninemic hyperaldosteronism secondary toECFvolume contraction, and juxtaglomerular apparatus hyperplasia (Chap. 331). Finally, diuretic use and abuse are common causes of K^+depletion. Carbonic anhydrase inhibitors, loop diuretics, and thiazides are all kaliuretic. The degree of hypokalemia tends to be greater with long-acting agents and is dose-dependent. Increased renal K^+excretion is due primarily to increased distal solute delivery and secondary hyperaldosteronism (due to volume depletion).

CLINICAL FEATURES

The clinical manifestations of K^+depletion vary greatly between individual patients, and their severity depends on the degree of hypokalemia. Symptoms seldom occur unless the plasma K^+concentration is less than 3 mmol/L. Fatigue, myalgia, and muscular weakness of the lower extremities are common complaints and are due to a lower (more negative) resting membrane potential. More severe hypokalemia may lead to progressive weakness, hypoventilation (due to respiratory muscle involvement), and eventually complete paralysis. Impaired muscle metabolism and the blunted hyperemic response to exercise associated with profound K^+depletion increase the risk of rhabdomyolysis. Smooth-muscle function may also be affected and manifest as paralytic ileus.

The electrocardiographic changes of hypokalemia (Fig. 226-19) are due to delayed ventricular repolarization and do not correlate well with the plasma K^+concentration. Early changes include flattening or inversion of the T wave, a prominent U wave, ST-segment depression, and a prolonged QU interval. Severe K^+depletion may result in a prolonged PR interval, decreased voltage and widening of the QRS complex, and an increased risk of ventricular arrhythmias, especially in patients with myocardial ischemia or left ventricular hypertrophy. Hypokalemia may also predispose to digitalis toxicity. Epidemiologic studies have linked a low-K^+diet with an increased prevalence of hypertension, particularly among African Americans. Furthermore, in patients with essential hypertension, systemic blood pressure may be lowered by K^+supplementation. The mechanism of the hypertensive effect of K^+depletion is not certain but may relate to enhanced distal NaCl reabsorption.

Hypokalemia is often associated with acid-base disturbances related to the underlying disorder. In addition, K^+depletion results in intracellular acidification and an increase in net acid excretion or new HCO_3^-production. This is a consequence of enhanced proximal HCO_3^-reabsorption, increased renal ammoniagenesis, and increased distal H^+secretion. This contributes to the generation of metabolic alkalosis frequently present in hypokalemic patients.NDI (see above) is not uncommonly seen in K^+depletion and is manifest as polydipsia and polyuria. Glucose intolerance may also occur with hypokalemia and has been attributed to either impaired insulin secretion or peripheral insulin resistance.

DIAGNOSIS

In most cases, the etiology of K^+depletion can be determined by a careful history. Diuretic and laxative abuse as well as surreptitious vomiting may be difficult to identify but should be excluded. Rarely, patients with a marked leukocytosis (e.g., acute myeloid leukemia) and normokalemia may have a low measured plasma K^+concentration due to white blood cell uptake of K^+ at room temperature. This *pseudohypokalemia* can be avoided by storing the blood sample on ice or rapidly separating the plasma (or serum) from the cells. After eliminating decreased intake and intracellular shift as potential causes of hypokalemia, examination of the renal response can help to clarify the source of K^+loss. The appropriate response to K^+depletion is to excrete less than 15 mmol/d of K^+ in the urine, due to increased reabsorption and decreased distal secretion. Hypokalemia with minimal renal K^+excretion suggests that K^+ was lost via the skin or gastrointestinal tract or that there is a remote history of vomiting or diuretic use. As described above, renal K^+wasting may be due to factors that either increase the K^+concentration in theCCD or increase the distal flow rate (or both). TheECFvolume status, blood pressure, and associated acid-base disorder may help to differentiate the causes of excessive renal K^+ loss. A rapid and simple test designed to evaluate the driving force for net K^+secretion is the *transtubular K^+concentration gradient* (TTKG). The TTKG is the ratio of the K^+concentration in the lumen of the CCD ($[K^+]_{CCD}$) to that in peritubular capillaries or plasma ($[K^+]_P$). The validity of this measurement depends on three assumptions: (1) few solutes are reabsorbed in the medullary collecting duct (MCD), (2) K^+ is neither secreted nor reabsorbed in the MCD, and (3) the osmolality of the fluid in the terminal CCD is known. Significant reabsorption or secretion of K^+ in the MCD seldom occurs, except in profound K^+depletion or excess, respectively. WhenAVP is acting ($OSM_U{}^3OSM_P$), the osmolality in the terminal CCD is the same as that of plasma, and the K^+concentration in the lumen of the distal nephron can be estimated by dividing the urine K^+concentration ($[K^+]_U$) by the ratio of the urine to plasma osmolality (OSM_U/OSM_P):

Hypokalemia with aTTKGgreater than 4 suggests renal K^+ loss due to increased distal K^+secretion. Plasma renin and aldosterone levels are often helpful in differentiating the various causes of hyperaldosteronism. Bicarbonaturia and the presence of other non-reabsorbed anions also increase the TTKG and lead to renal K^+-wasting.

CLINICAL APPROACH

SeeFig. 49-3.

TREATMENT

The therapeutic goals are to correct the K^+deficit and to minimize ongoing losses. With the exception of periodic paralysis, hypokalemia resulting from transcellular shifts rarely requires intravenous K^+supplementation, which can lead to rebound hyperkalemia. It is generally safer to correct hypokalemia via the oral route. The degree of K^+depletion does not correlate well with the plasma K^+concentration. A decrement of 1 mmol/L in the plasma K^+concentration (from 4.0 to 3.0 mmol/L) may represent a total body K^+deficit of 200 to 400 mmol, and patients with plasma levels under 3.0 mmol/L often require in excess of 600 mmol of K^+ to correct the deficit. Furthermore, factors promoting K^+ shift out of cells (e.g., insulin deficiency in diabetic ketoacidosis) may result in underestimation of the K^+deficit. Therefore, the plasma K^+concentration should be monitored frequently when assessing the response to treatment. Potassium chloride is usually the preparation of choice and will promote more rapid correction of hypokalemia and metabolic alkalosis. Potassium bicarbonate and citrate (metabolized to HCO_3^-) tend to alkalinize the patient and would be more appropriate for hypokalemia associated with chronic diarrhea orRTA.

Patients with severe hypokalemia or those unable to take anything by mouth require intravenous replacement therapy with KCl. The maximum concentration of administered K^+should be no more than 40 mmol/L via a peripheral vein or 60 mmol/L via a central vein. The rate of infusion should not exceed 20 mmol/h unless paralysis or malignant ventricular arrhythmias are present. Ideally, KCl should be mixed in normal saline since dextrose solutions may initially exacerbate hypokalemia due to insulin-mediated movement of K^+ into cells. Rapid intravenous administration of K^+should be used judiciously and requires close observation of the clinical manifestations of hypokalemia (electrocardiogram and neuromuscular examination).

HYPERKALEMIA

ETIOLOGY

Hyperkalemia, defined as a plasma K^+concentration >5.0 mmol/L, occurs as a result of either K^+release from cells or decreased renal loss. Increased K^+intake is rarely the sole cause of hyperkalemia since the phenomenon of *potassium adaptation* ensures rapid K^+excretion in response to increases in dietary consumption. Iatrogenic hyperkalemia may result from overzealous parenteral K^+replacement or in patients with renal insufficiency. *Pseudohyperkalemia* represents an artificially elevated plasma K^+concentration due to K^+movement out of cells immediately prior to or following venipuncture. Contributing factors include prolonged use of a tourniquet with or without repeated fist clenching, hemolysis, and marked leukocytosis or thrombocytosis. The latter two result in an elevated serum K^+concentration due to release of intracellular K^+following clot formation. Pseudohyperkalemia should be suspected in an otherwise asymptomatic patient with no obvious underlying cause. If proper venipuncture technique is used and a plasma (not serum) K^+concentration is measured, it should be normal. Intravascular hemolysis, tumor lysis syndrome, and rhabdomyolysis all lead to K^+release from cells as a result of tissue breakdown. Metabolic acidoses, with the exception of those due to the accumulation of organic anions, can be associated with mild hyperkalemia resulting from intracellular buffering of H^+ (see above). As previously described (p. 278), insulin deficiency and hypertonicity (e.g., hyperglycemia) promote K^+ shift from theICF to the ECF. The severity of exercise-induced hyperkalemia is related to the degree of exertion. It is due to release of K^+ from muscles and is usually rapidly reversible, often associated with rebound hypokalemia. Treatment with beta blockers rarely causes hyperkalemia but may contribute to the elevation in plasma K^+concentration seen with other conditions. *Hyperkalemic periodic paralysis* (Chap. 381) is a rare autosomal dominant disorder characterized by episodic weakness or paralysis, precipitated by stimuli that normally lead to mild hyperkalemia (e.g., exercise). The genetic defect appears to be a single amino acid substitution due to a mutation in the gene for the skeletal muscle Na^+channel. Hyperkalemia may occur with severe digitalis toxicity due to inhibition of the Na^+, K^+-ATPase pump. Depolarizing muscle relaxants such as succinylcholine can increase the plasma K^+concentration, especially in patients with massive trauma, burns, or neuromuscular disease.

Chronic hyperkalemia is virtually always associated with decreased renal K^+excretion due to either impaired secretion or diminished distal solute delivery (Table 49-4). The latter is seldom the only cause of impaired K^+excretion but may significantly contribute to hyperkalemia in protein-malnourished (low urea excretion) andECFvolume-contracted (decreased distal NaCl delivery) patients. Decreased K^+secretion by the principal cells results from either impaired Na^+reabsorption or increased Cl^- reabsorption, both of which give rise to a diminished (less lumen-negative)TEPD in the CCD. *Hyporeninemic hypoaldosteronism* is a syndrome characterized by euvolemia or ECF volume expansion and suppressed renin and aldosterone levels (Chaps. 331 and333). This disorder is commonly seen in mild renal insufficiency, diabetic nephropathy, or chronic tubulointerstitial disease. Patients frequently have an impaired kaliuretic response to exogenous mineralocorticoid administration, suggesting that enhanced distal Cl$^-$ reabsorption (electroneutral Na$^+$reabsorption) may account for many of the findings of hyporeninemic hypoaldosteronism.NSAIDsinhibit renin secretion and the synthesis of vasodilatory renal prostaglandins. The resultant decrease inGFR and K$^+$secretion is often manifest as hyperkalemia. As a rule, the degree of hyperkalemia due to hypoaldosteronism is mild in the absence of increased K$^+$intake or renal dysfunction. Angiotensin-converting enzyme (ACE) inhibitors block the conversion of angiotensin I to angiotensin II, resulting in impaired aldosterone release. Patients at increased risk of ACE inhibitor-induced hyperkalemia include those with diabetes mellitus, renal insufficiency, decreased effective circulating arterial volume, bilateral renal artery stenosis, or concurrent use of K$^+$-sparing diuretics or NSAIDs.

Decreased aldosterone synthesis may be due to *primary adrenal insufficiency* (Addison's disease) or congenital adrenal enzyme deficiency (Chap. 331). Heparin (including low-molecular-weight heparin) inhibits production of aldosterone by the cells of the zona glomerulosa and can lead to severe hyperkalemia in a subset of patients with underlying renal disease; diabetes mellitus; or those receiving K$^+$-sparing diuretics,ACEinhibitors, orNSAIDs. *Pseudohypoaldosteronism* is a rare familial disorder characterized by hyperkalemia, metabolic acidosis, renal Na$^+$wasting, hypotension, high renin and aldosterone levels, and end-organ resistance to aldosterone. The gene encoding the mineralocorticoid receptor is normal in these patients, and the electrolyte abnormalities can be reversed with suprapharmacologic doses of an exogenous mineralocorticoid (e.g., 9a-fludrocortisone) or an inhibitor of11b-HSDH(e.g., carbenoxolone). The kaliuretic response to aldosterone is impaired by K$^+$-sparing diuretics. Spironolactone is a competitive mineralocorticoid antagonist, whereas amiloride and triamterene block the apical Na$^+$channel of the principal cell. Two other drugs that impair K$^+$secretion by blocking distal nephron Na$^+$reabsorption are trimethoprim and pentamidine. These antimicrobial agents may contribute to the hyperkalemia often seen in patients infected with HIV who are being treated for *Pneumocystis carinii* pneumonia.

Hyperkalemia frequently complicates acute oliguric renal failure due to increased K$^+$release from cells (acidosis, catabolism) and decreased excretion. Increased distal flow rate and K$^+$secretion per nephron compensate for decreased renal mass in chronic renal insufficiency. However, these adaptive mechanisms eventually fail to maintain K$^+$balance when theGFRfalls below 10 to 15 mL/min or oliguria ensues. Otherwise asymptomatic urinary tract obstruction is an often overlooked cause of hyperkalemia. Other nephropathies associated with impaired K$^+$excretion include drug-induced interstitial nephritis, lupus nephritis, sickle cell disease, and diabetic nephropathy.

Gordon's syndrome is a rare condition characterized by hyperkalemia, metabolic acidosis, and a normalGFR. These patients are usually volume-expanded with suppressed renin and aldosterone levels as well as refractory to the kaliuretic effect of exogenous mineralocorticoids. It has been suggested that these findings could all be accounted for by increased distal Cl$^-$ reabsorption (electroneutral Na$^+$reabsorption), also referred to as a *Cl$^-$shunt*. A similar mechanism may be partially responsible for the hyperkalemia associated with cyclosporine nephrotoxicity. *Hyperkalemic distal (type 4)RTA*may be due to either hypoaldosteronism or a Cl^- shunt (aldosterone-resistant).

CLINICAL FEATURES

Since the resting membrane potential is related to the ratio of theICF toECFK^+concentration, hyperkalemia partially depolarizes the cell membrane. Prolonged depolarization impairs membrane excitability and is manifest as weakness, which may progress to flaccid paralysis and hypoventilation if the respiratory muscles are involved. Hyperkalemia also inhibits renal ammoniagenesis and reabsorption of NH_4^+in the thick ascending limb of the loop of Henle. Thus, net acid excretion is impaired and results in metabolic acidosis, which may further exacerbate the hyperkalemia due to K^+movement out of cells.

The most serious effect of hyperkalemia is cardiac toxicity, which does not correlate well with the plasma K^+concentration. The earliest electrocardiographic changes include increased T-wave amplitude, or peaked T waves. More severe degrees of hyperkalemia result in a prolonged PR interval and QRS duration, atrioventricular conduction delay, and loss of P waves. Progressive widening of the QRS complex and merging with the T wave produces a sinewave pattern. The terminal event is usually ventricular fibrillation or asystole.

DIAGNOSIS

With rare exceptions, chronic hyperkalemia is always due to impaired K^+excretion. If the etiology is not readily apparent and the patient is asymptomatic, pseudohyperkalemia should be excluded, as described above. Oliguric acute renal failure and severe chronic renal insufficiency should also be ruled out. The history should focus on medications that impair K^+handling and potential sources of K^+intake. Evaluation of theECFcompartment, effective circulating volume, and urine output are essential components of the physical examination. The severity of hyperkalemia is determined by the symptoms, plasma K^+concentration, and electrocardiographic abnormalities.

The appropriate renal response to hyperkalemia is to excrete at least 200 mmol of K^+daily. In most cases, diminished renal K^+ loss is due to impaired K^+secretion, which can be assessed by measuring theTTKG(see above). A TTKG <10 implies a decreased driving force for K^+secretion due to either hypoaldosteronism or resistance to the renal effects of mineralocorticoid. This can be determined by evaluating the kaliuretic response to administration of mineralocorticoid (e.g., 9a-fludrocortisone). Primary adrenal insufficiency can be differentiated from hyporeninemic hypoaldosteronism by examining the renin-aldosterone axis. Renin and aldosterone levels should be measured in the supine and upright positions, following three days of Na^+restriction (Na^+intake <10 mmol/d) in combination with a loop diuretic to induce mild volume contraction. Aldosterone-resistant hyperkalemia can result from the various causes of impaired distal Na^+reabsorption or from a Cl^- shunt. The former leads to salt wasting,ECFvolume contraction, and high renin and aldosterone levels. In contrast, enhanced distal Cl^-reabsorption is associated with volume expansion and suppressed renin and aldosterone secretion. As mentioned above, hypoaldosteronism seldom causes severe hypokalemia in the absence of increased dietary K^+intake, renal insufficiency, transcellular K^+shifts, or antikaliuretic drugs.

CLINICAL APPROACH

SeeFig. 49-4.

TREATMENT

The approach to therapy depends on the degree of hyperkalemia as determined by the plasma K^+concentration, associated muscular weakness, and changes on the electrocardiogram. Potentially fatal hyperkalemia rarely occurs unless the plasma K^+concentration exceeds 7.5 mmol/L and is usually associated with profound weakness and absent P waves, QRS widening, or ventricular arrhythmias on the electrocardiogram.

Severe hyperkalemia requires emergent treatment directed at minimizing membrane depolarization, shifting K^+ into cells, and promoting K^+ loss. In addition, exogenous K^+intake and antikaliuretic drugs should be discontinued. Administration of calcium gluconate decreases membrane excitability. The usual dose is 10 mL of a 10% solution infused over 2 to 3 min. The effect begins within minutes but is short-lived (30 to 60 min), and the dose can be repeated if no change in the electrocardiogram is seen after 5 to 10 min. Insulin causes K^+ to shift into cells by mechanisms described previously and will temporarily lower the plasma K^+concentration. Although glucose alone will stimulate insulin release from normal pancreatic b cells, a more rapid response generally occurs when exogenous insulin is administered (with glucose to prevent hypoglycemia). A commonly recommended combination is 10 to 20 units of regular insulin and 25 to 50 g of glucose. Obviously, hyperglycemic patients should not be given glucose. If effective, the plasma K^+concentration will fall by 0.5 to 1.5 mmol/L in 15 to 30 min and the effect will last for several hours. Alkali therapy with intravenous $NaHCO_3$can also shift K^+ into cells. This is safest when administered as an isotonic solution of 3 ampules per liter (134 mmol/L $NaHCO_3$) and ideally should be reserved for severe hyperkalemia associated with metabolic acidosis. Patients with end-stage renal disease seldom respond to this intervention and may not tolerate the Na^+ load and resultant volume expansion. When administered parenterally or in nebulized form,b_2-adrenergic agonists promote cellular uptake of K^+ (see above). The onset of action is 30 min, lowering the plasma K^+concentration by 0.5 to 1.5 mmol/L, and the effect lasts 2 to 4 h.

Removal of K^+ can be achieved using diuretics, cation-exchange resin, or dialysis. Loop and thiazide diuretics, often in combination, may enhance K^+excretion if renal function is adequate. Sodium polystyrene sulfonate is a cation-exchange resin that promotes the exchange of Na^+ for K^+ in the gastrointestinal tract. Each gram binds 1 mmol of K^+ and releases 2 to 3 mmol of Na^+. When given by mouth, the usual dose is 25 to 50 g mixed with 100 mL of 20% sorbitol to prevent constipation. This will generally lower the plasma K^+concentration by 0.5 to 1.0 mmol/L within 1 to 2 h and last for 4 to 6 h. Sodium polystyrene sulfonate can also be administered as a retention enema consisting of 50 g of resin and 50 mL of 70% sorbitol mixed in 150 mL of tap water. The sorbitol should be omitted from the enema in postoperative patients due to the increased incidence of sorbitol-induced colonic necrosis, especially following renal transplantation. The most rapid and effective way of lowering the plasma K^+concentration is hemodialysis. This should be reserved for patients with renal failure and those with severe life-threatening hyperkalemia unresponsive to more conservative measures. Peritoneal dialysis also removes K$^+$ but is only 15 to 20% as effective as hemodialysis. Finally, the underlying cause of the hyperkalemia should be treated. This may involve dietary modification, correction of metabolic acidosis, cautious volume expansion, and administration of exogenous mineralocorticoid.

(Bibliography omitted in Palm version)

Back to Table of Contents

(Bibliography omitted in Palm version)

Back to Table of Contents

50. ACIDOSIS AND ALKALOSIS - *Thomas D. DuBose, Jr.*

NORMAL ACID-BASE HOMEOSTASIS

Systemic arterial pH is maintained between 7.35 and 7.45 by extracellular and intracellular chemical buffering together with respiratory and renal regulatory mechanisms. The control of arterial CO_2 tension (Pa_{CO_2}) by the central nervous system and respiratory systems and the control of the plasma bicarbonate by the kidneys stabilize the arterial pH by excretion or retention of acid or alkali. The metabolic and respiratory components that regulate systemic pH are described by the Henderson-Hasselbalch equation:

Under most circumstances, CO_2 production and excretion are matched, and the usual steady-state Pa_{CO_2} is maintained at 40 mmHg. Underexcretion of CO_2 produces hypercapnia, and overexcretion causes hypocapnia. Nevertheless, production and excretion are again matched at a new steady-state Pa_{CO_2}. Therefore, the Pa_{CO_2} is regulated primarily by neural respiratory factors (Chap. 263) and is not subject to regulation by the rate of CO_2 production. Hypercapnia is usually the result of hypoventilation rather than of increased CO_2 production. Increases or decreases in Pa_{CO_2} represent derangements of neural respiratory control or are due to compensatory changes in response to a primary alteration in the plasma [HCO_3^-].

Primary changes in Pa_{CO_2} can cause acidosis or alkalosis, depending on whether Pa_{CO_2} is above or below the normal value of 40 mmHg (respiratory acidosis or alkalosis, respectively). Primary alteration of Pa_{CO_2} evokes cellular buffering and renal adaptation, a slow process that becomes more efficient with time. A primary change in the plasma [HCO_3^-] as a result of metabolic or renal factors results in compensatory changes in ventilation that blunt the changes in blood pH that would occur otherwise. Such respiratory alterations are referred to as *secondary*, or compensatory, changes, since they occur in response to primary metabolic changes.

The kidneys regulate plasma [HCO_3^-] through three main processes: (1) "reabsorption" of filtered HCO_3^-, (2) formation of titratable acid, and (3) excretion of NH_4^+ in the urine. The kidney filters approximately 4000 mmol of HCO_3^- per day. To reabsorb the filtered load of HCO_3^-, the renal tubules must therefore secrete 4000 mmol of hydrogen ions. Between 80 and 90% of HCO_3^- is reabsorbed in the proximal tubule. The distal nephron reabsorbs the remainder and secretes protons, as generated from metabolism, to defend systemic pH. While this quantity of protons, 40 to 60 mmol/d, is small, it must be secreted to prevent chronic positive H^+ balance and metabolic acidosis. This quantity of secreted protons is represented in the urine as titratable acid and NH_4^+. Metabolic acidosis in the face of normal renal function increases NH_4^+ production and excretion. NH_4^+ production and excretion are impaired in chronic renal failure, hyperkalemia, and renal tubular acidosis.

In sum, these regulatory responses, including chemical buffering, the regulation of Pa_{CO_2} by the respiratory system, and of [HCO_3^-] by the kidneys, act in concert to maintain a systemic arterial pH between 7.35 and 7.45.

DIAGNOSIS OF GENERAL TYPES OF DISTURBANCES

The most common clinical disturbances are simple acid-base disorders, i.e., metabolic acidosis or alkalosis or respiratory acidosis or alkalosis. Since compensation is not complete, the pH is abnormal in simple disturbances. More complicated clinical situations can give rise to mixed acid-base disturbances.

SIMPLE ACID-BASE DISORDERS

Primary respiratory disturbances (primary changes in Pa_{CO2}) invoke compensatory metabolic responses (secondary changes in $[HCO_{3-}]$), and primary metabolic disturbances elicit predictable compensatory respiratory responses. Physiologic compensation can be predicted from the relationships displayed in Table 50-1. Primary changes in Pa_{CO2} or $[HCO_{3-}]$ alter systemic pH and cause acidosis or alkalosis. To illustrate, metabolic acidosis due to an increase in endogenous acids (e.g., ketoacidosis) lowers extracellular fluid $[HCO_{3-}]$ and decreases extracellular pH. This stimulates the medullary chemoreceptors to increase ventilation and to return the ratio of $[HCO_{3-}]$ to Pa_{CO2}, and thus pH, toward normal, although not to normal. The degree of respiratory compensation expected in a simple form of metabolic acidosis can be predicted from the relationship: Pa_{CO2}= (1.5 ´[HCO_{3-}]) + 8, i.e., the Pa_{CO2} is expected to decrease 1.25 mmHg for each mmol per liter decrease in $[HCO_{3-}]$. Thus, a patient with metabolic acidosis and $[HCO_{3-}]$ of 12 mmol/L would be expected to have a Pa_{CO2}between 24 and 28 mmHg. Values for Pa_{CO2}below 24 or greater than 28 mmHg define a mixed disturbance (metabolic acidosis and respiratory alkalosis or metabolic alkalosis and respiratory acidosis, respectively). Another way to judge the appropriateness of the response in $[HCO_{3-}]$ or $Pa_{CO2}$is to use an acid-base nomogram (Fig. 50-1). While the shaded areas of the nomogram show the 95% confidence limits for normal compensation in simple disturbances, finding acid-base values within the shaded area does not necessarily rule out a mixed disturbance. Imposition of one disorder over another may result in values lying within the area of a third. Thus, the nomogram, while convenient, is not a substitute for the equations in Table 50-1.

MIXED ACID-BASE DISORDERS

Mixed acid-base disorders -- defined as independently coexisting disorders, not merely compensatory responses -- are often seen in patients in critical care units and can lead to dangerous extremes of pH. A patient with diabetic ketoacidosis (metabolic acidosis) may develop an independent respiratory problem leading to respiratory acidosis or alkalosis. Patients with underlying pulmonary disease may not respond to metabolic acidosis with an appropriate ventilatory response because of insufficient respiratory reserve. Such imposition of respiratory acidosis on metabolic acidosis can lead to severe acidemia and a poor outcome. When metabolic acidosis and metabolic alkalosis coexist in the same patient, the pH may be normal or near normal. When the pH is normal, an elevated anion gap (see below) denotes the presence of a metabolic acidosis. A diabetic patient with ketoacidosis may have renal dysfunction resulting in simultaneous metabolic acidosis. Patients who have ingested an overdose of drug combinations such as sedatives and salicylates may have mixed disturbances as a result of the acid-base response to the individual drugs (metabolic acidosis mixed with respiratory acidosis or respiratory alkalosis, respectively). Even more complex are triple acid-base disturbances. For example, patients with metabolic acidosis due to alcoholic ketoacidosis may develop metabolic alkalosis due to vomiting and superimposed respiratory alkalosis due to the hyperventilation of hepatic dysfunction or alcohol withdrawal.

DIAGNOSIS OF ACID-BASE DISORDERS

Care should be taken when measuring blood gases to obtain the arterial blood sample without using excessive heparin. In the determination of arterial blood gases by the clinical laboratory, both pH and $Pa_{CO2}$are measured, and the [HCO_{3-}] is calculated from the Henderson-Hasselbalch equation. This calculated value should be compared with the measured [HCO_{3-}] (total CO_2) on the electrolyte panel. These two values should agree within 2 mmol/L. If they do not, the values may not have been drawn simultaneously, a laboratory error may be present, or an error could have been made in calculating the [HCO_{3-}]. After verifying the blood acid-base values, one can then identify the precise acid-base disorder.

The most common causes of acid-base disorders should be kept in mind while probing the history for clues about the etiology. For example, established chronic renal failure is expected to cause a metabolic acidosis, and chronic vomiting frequently causes metabolic alkalosis. Patients with pneumonia, sepsis, or cardiac failure frequently have respiratory alkalosis, and patients with chronic obstructive pulmonary disease or a sedative drug overdose often display a respiratory acidosis. The drug history is important since loop or thiazide diuretics may cause metabolic alkalosis, and the carbonic anhydrase inhibitor, acetazolamide, can result in metabolic acidosis.

Blood for electrolytes and arterial blood gases should be drawn simultaneously prior to therapy, since an increase in [HCO_{3-}] occurs with metabolic alkalosis and respiratory acidosis. Conversely, a decrease in [HCO_{3-}] occurs in metabolic acidosis and respiratory alkalosis.

Metabolic acidosis leads to hyperkalemia as a result of cellular shifts in which H_+ is exchanged for K_+ or Na_+. For each decrease in blood pH of 0.10, the plasma [K_+] should rise by 0.6 mmol/L. This relationship is not invariable. Diabetic ketoacidosis, lactic acidosis, diarrhea, and renal tubular acidosis (RTA) are often associated with potassium depletion because of urinary K_+wasting.

Anion Gap All evaluations of acid-base disorders should include a simple calculation of the anion gap (AG); it represents those unmeasured anions in plasma (normally 10 to 12 mmol/L) and is calculated as follows: AG =Na_+-(Cl_- + HCO_{3-}). The unmeasured anions include anionic proteins, phosphate, sulfate, and organic anions. When acid anions, such as acetoacetate and lactate, accumulate in extracellular fluid, the AG increases, causing a high-AG acidosis. An increase in the AG is most often due to an increase in unmeasured anions and less commonly is due to a decrease in unmeasured cations (calcium, magnesium, potassium). In addition, the AG may increase with an increase in anionic albumin, either because of increased albumin concentration or alkalosis, which alters albumin charge. A decrease in the AG can be due to: (1) an increase in unmeasured cations; (2) the addition to the blood of abnormal cations, such as lithium (lithium intoxication) or cationic immunoglobulins (plasma cell dyscrasias); (3) a reduction in the major plasma anion albumin concentration (nephrotic syndrome); (4) a decrease in the effective anionic charge on albumin by acidosis; or (5) hyperviscosity and severe hyperlipidemia, which can lead to an underestimation of sodium and chloride concentrations.

In the face of a normal serum albumin, a highAG is usually due to non-chloride-containing acids that contain inorganic (phosphate, sulfate), organic (ketoacids, lactate, uremic organic anions), exogenous (salicylate or ingested toxins with organic acid production), or unidentified anions. By definition, therefore, a high-AG acidosis has two identifying features: a low [HCO_3-] and an elevated AG. The latter is present even if an additional acid-base disorder is superimposed to modify the [HCO_3-] independently. Simultaneous metabolic acidosis of the high-AG variety plus either chronic respiratory acidosis or metabolic alkalosis represents such a situation in which [HCO_3-] may be normal or even high. However, the AG is elevated, and [Cl-] is depressed.

Similarly, normal values for [HCO_3-], Pa_{CO_2}, and pH do not ensure the absence of an acid-base disturbance. For instance, an alcoholic who has been vomiting may develop a metabolic alkalosis with a pH of 7.55, Pa_{CO_2} of 48 mmHg, [HCO_3-] of 40 mmol/L, [Na^+] of 135, [Cl-] of 80, and [K^+] of 2.8. If such a patient were then to develop a superimposed alcoholic ketoacidosis with a b-hydroxybutyrate concentration of 15 m*M*, arterial pH would fall to 7.40, [HCO_3-] to 25 mmol/L, and the Pa_{CO_2} to 40 mmHg. Although these blood gases are normal, theAG is elevated at 30 mmol/L, indicating a mixed metabolic alkalosis and metabolic acidosis.

METABOLIC ACIDOSIS

Metabolic acidosis can occur because of an increase in endogenous acid production (such as lactate and ketoacids), loss of bicarbonate (as in diarrhea), or accumulation of endogenous acids (as in renal failure). Metabolic acidosis has profound effects on the respiratory, cardiac, and nervous systems. The fall in blood pH is accompanied by a characteristic increase in ventilation, especially the tidal volume (Kussmaul respiration). Intrinsic cardiac contractility may be depressed, but inotropic function can be normal because of catecholamine release. Both peripheral arterial vasodilation and central venoconstriction can be present; the decrease in central and pulmonary vascular compliance predisposes to pulmonary edema with even minimal volume overload. Central nervous system function is depressed, with headache, lethargy, stupor, and, in some cases, even coma. Glucose intolerance may also occur.

There are two major categories of clinical metabolic acidosis: high-AGand normal-AG, or hyperchloremic acidosis (Table 50-2 andTable 50-3).

TREATMENT

Treatment of metabolic acidosis with alkali should be reserved for severe acidemia except when the patient has no "potential HCO_3-" in plasma. Potential [HCO_3-] can be estimated from the increment (D) in the anion gap (DAG= patient's AG - 10). It must be determined if the acid anion in plasma is metabolizable (i.e.,b-hydroxybutyrate, acetoacetate, and lactate) or nonmetabolizable (anions that accumulate in chronic renal failure and after toxin ingestion). The latter requires return of renal function to replenish the [HCO_3^-] deficit, a slow and often unpredictable process. Consequently, patients with a normal AG acidosis (hyperchloremic acidosis), a slightly elevated AG (mixed hyperchloremic and AG acidosis), or an AG attributable to a nonmetabolizable anion in the face of renal failure should receive alkali therapy, either orally ($NaHCO_3$or Shohl's solution) or intravenously ($NaHCO_3$), in an amount necessary to slowly increase the plasma [HCO_3^-] into the 20 to 22 mmol/L range.

Controversy exists, however, in regard to the use of alkali in patients with a pureAGacidosis owing to accumulation of a metabolizable organic acid anion (ketoacidosis or lactic acidosis). In general, severe acidosis (pH< 7.20) warrants the intravenous administration of 50 to 100 meq of $NaHCO_3$, over 30 to 45 min, during the initial 1 to 2 h of therapy. Provision of such modest quantities of alkali in this situation seems to provide an added measure of safety, but it is essential to monitor plasma electrolytes during the course of therapy, since the [K^+] may decline as pH rises. The goal is to increase the [HCO_3^-] to 10 meq/L and the pH to 7.25, not to increase these values to normal.

HIGH-ANION-GAP ACIDOSES

There are four principal causes of a high-AGacidosis: (1) lactic acidosis, (2) ketoacidosis, (3) ingested toxins (Table 50-2), and (4) acute and chronic renal failure. Initial screening to differentiate the high-AG acidoses should include: (1) a probe of the history for evidence of drug and toxin ingestion and measurement of arterial blood gas to detect coexistent respiratory alkalosis (salicylates); (2) determination of whether diabetes mellitus is present (diabetic ketoacidosis); (3) a search for evidence of alcoholism or increased levels ofb-hydroxybutyrate (alcoholic ketoacidosis); (4) observation for clinical signs of uremia and determination of the blood urea nitrogen (BUN) and creatinine (uremic acidosis); (5) inspection of the urine for oxalate crystals (ethylene glycol); and (6) recognition of the numerous clinical settings in which lactate levels may be increased (hypotension, shock, cardiac failure, leukemia, cancer, and drug or toxin ingestion).

Lactic Acidosis An increase in plasma L-lactate may be secondary to poor tissue perfusion (type A) -- circulatory insufficiency (shock, circulatory failure), severe anemia, mitochondrial enzyme defects, and inhibitors (carbon monoxide, cyanide) -- or to aerobic disorders (type B) -- malignancies, diabetes mellitus, renal or hepatic failure, severe infections (cholera, malaria), seizures, AIDS, or drugs/toxins (biguanides, ethanol, methanol, isoniazid, AZT analogues, and fructose). Unrecognized bowel ischemia or infarction in a patient with severe atherosclerosis or cardiac decompensation receiving vasopressors is a common cause of lactic acidosis. D-Lactic acid acidosis, which may be associated with jejunoileal bypass or intestinal obstruction and is due to formation of D-lactate by gut bacteria, may cause both an increasedAG and hyperchloremia.

TREATMENT

The underlying condition that disrupts lactate metabolism must first be corrected; tissue perfusion must be restored when it is inadequate. Vasoconstrictors should be avoided, if possible, since they may worsen tissue perfusion. Alkali therapy is generally advocated for acute, severe acidemia (pH< 7.1) to improve cardiac function and lactate utilization. However, $NaHCO_3$therapy may paradoxically depress cardiac performance and exacerbate acidosis by enhancing lactate production (HCO_3^-stimulates phosphofructokinase). While the use of alkali in moderate lactic acidosis is controversial, it is generally agreed that attempts to return the pH or [HCO_3^-] to normal by administration of exogenous $NaHCO_3$are deleterious. A reasonable approach is to infuse sufficient $NaHCO_3$to raise the arterial pH to no more than 7.2 over 30 to 40 min.

$NaHCO_3$therapy can cause fluid overload and hypertension because the amount required can be massive when accumulation of lactic acid is relentless. Fluid administration is poorly tolerated because of central venoconstriction, especially in the oliguric patient. If the underlying cause of the lactic acidosis can be remedied, blood lactate will be converted to $HCO_3^-$and may result in an overshoot alkalosis.

Ketoacidosis

Diabetic Ketoacidosis This condition is caused by increased fatty acid metabolism and the accumulation of ketoacids (acetoacetate andb-hydroxybutyrate). Diabetic ketoacidosis usually occurs in insulin-dependent diabetes mellitus in association with cessation of insulin or an intercurrent illness, such as an infection, gastroenteritis, pancreatitis, or myocardial infarction, which increases insulin requirements temporarily and acutely. The accumulation of ketoacids accounts for the increment in theAG and is accompanied most often by hyperglycemia [glucose> 17 mmol/L (300 mg/dL)]. It should be noted that since insulin prevents production of ketones, bicarbonate therapy is rarely needed except with extreme acidemia (pH < 7.1), and then in only limited amounts (see "Treatment" for lactic acidosis).**The management of this condition is described in Chap. 333.*

Alcoholic Ketoacidosis Chronic alcoholics can develop ketoacidosis when alcohol consumption is abruptly curtailed; it is usually associated with binge drinking, vomiting, abdominal pain, starvation, and volume depletion. The glucose concentration is low or normal, and acidosis may be severe because of elevated ketones, predominantlyb-hydroxybutyrate. Mild lactic acidosis may coexist because of alteration in the redox state. The nitroprusside ketone reaction (Acetest) can detect acetoacetic acid but notb-hydroxybutyrate, so that the degree of ketosis and ketonuria can be underestimated. Typically, insulin levels are low, and concentrations of triglyceride, cortisol, glucagon, and growth hormone are increased.

TREATMENT

Extracellular fluid deficits should be repleted by intravenous administration of saline and glucose (5% dextrose in 0.9% NaCl). Hypophosphatemia, hypokalemia, and hypomagnesemia may coexist and should be corrected. Hypophosphatemia usually emerges 12 to 24 h after admission, may be exacerbated by glucose infusion, and, if severe, may induce rhabdomyolysis. Upper gastrointestinal hemorrhage, pancreatitis, and pneumonia may accompany this disorder.

Drug- and Toxin-Induced Acidosis

Salicylates (See alsoChap. 396) Salicylate intoxication in adults usually causes respiratory alkalosis, mixed metabolic acidosis-respiratory alkalosis, or a pure high-AGmetabolic acidosis. In the latter example, which is less common, only a portion of the AG is due to the salicylates. Lactic acid production is also often increased.

TREATMENT

This should begin with vigorous gastric lavage with isotonic saline (not $NaHCO_3$) followed by administration of activated charcoal. In the acidotic patient, to facilitate removal of salicylate, intravenous $NaHCO_3$is administered in amounts adequate to alkalinize the urine and to maintain urine output (urine pH > 7.5). While this form of therapy is straightforward in acidotic patients, a coexisting respiratory alkalosis may make this approach hazardous. Acetazolamide may be administered when an alkaline diuresis cannot be achieved, but this drug can cause systemic metabolic acidosis if $HCO_3^-$is not replaced. Hypokalemia may occur with an alkaline diuresis from $NaHCO_3$and should be treated promptly and aggressively. Glucose-containing fluids should be administered because of the danger of hypoglycemia. Excessive insensible fluid losses may cause severe volume depletion and hypernatremia. If renal failure prevents rapid clearance of salicylate, hemodialysis can be performed against a bicarbonate dialysate.

Alcohols Under most physiologic conditions, sodium, urea, and glucose generate the osmotic pressure of blood. Plasma osmolality is calculated according to the following expression: $P_{osm}= 2Na^+ + Glu +BUN$ (all in mmol/L), or, using conventional laboratory values in which glucose and BUN are expressed in milligrams per deciliter: $P_{osm}= 2Na^+ + Glu/18 + BUN/2.8$. The calculated and determined osmolality should agree within 10 to 15 mmol/kg H_2O. When the measured osmolality exceeds the calculated osmolality by more than 15 to 20 mmol/kg H_2O, one of two circumstances prevails. Either the serum sodium is spuriously low, as with hyperlipidemia or hyperproteinemia (pseudohyponatremia), or osmolytes other than sodium salts, glucose, or urea have accumulated in plasma. Examples include mannitol, radiocontrast media, isopropyl alcohol, ethylene glycol, ethanol, methanol, and acetone. In this situation, the difference between the calculated osmolality and the measured osmolality (*osmolar gap*) is proportional to the concentration of the unmeasured solute. With an appropriate clinical history and index of suspicion, identification of an osmolar gap is helpful in identifying the presence of poison-associatedAGacidosis.

Ethylene Glycol (See alsoChap. 396) Ingestion of ethylene glycol (commonly used in antifreeze) leads to a metabolic acidosis and severe damage to the central nervous system, heart, lungs, and kidneys. The increasedAG and osmolar gap are attributable to ethylene glycol and its metabolites, oxalic acid, glycolic acid, and other organic acids. Lactic acid production increases secondary to inhibition of the tricarboxylic acid cycle and altered intracellular redox state. Diagnosis is facilitated by recognizing oxalate crystals in the urine, the presence of an osmolar gap in serum, and a high-AG acidosis. Treatment should not be delayed while awaiting measurement of ethylene glycol levels in this setting.

TREATMENT

This includes the prompt institution of a saline or osmotic diuresis, thiamine and pyridoxine supplements, fomepizole or ethanol, and hemodialysis. The intravenous administration of the new alcohol dehydrogenase inhibitor, fomepizole (4-methylpyrazole; 7 mg/kg as a loading dose), or ethanol intravenously to achieve a level of 22 mmol/L (100 mg/dL) serves to lessen toxicity because they compete with ethylene glycol for metabolism by alcohol dehydrogenase. Fomepizole, although expensive, offers the advantages of a predictable decline in ethylene glycol levels without the adverse effects, such as excessive obtundation, associated with ethyl alcohol infusion.

Methanol (See alsoChap. 396) The ingestion of methanol (wood alcohol) causes metabolic acidosis, and its metabolites formaldehyde and formic acid cause severe optic nerve and cental nervous system damage. Lactic acid, ketoacids, and other unidentified organic acids may contribute to the acidosis. Due to its low molecular weight (32 Da), an osmolar gap is usually present.

TREATMENT

This is similar to that for ethylene glycol intoxication, including general supportive measures, fomepizole or ethanol administration, and hemodialysis.

Renal Failure (See alsoChaps. 269 and270) The hyperchloremic acidosis of moderate renal insufficiency is eventually converted to the high-AGacidosis of advanced renal failure. Poor filtration and reabsorption of organic anions contribute to the pathogenesis. As renal disease progresses, the number of functioning nephrons eventually becomes insufficient to keep pace with net acid production. Uremic acidosis is characterized, therefore, by a reduced rate of NH_4^+production and excretion, primarily due to decreased renal mass. [HCO_3^-] rarely falls below 15 mmol/L, and the AG rarely exceeds 20 mmol/L. The acid retained in chronic renal disease is buffered by alkaline salts from bone. Despite significant retention of acid (up to 20 mmol/d), the serum [HCO_3^-] does not decrease further, indicating participation of buffers outside the extracellular compartment. Chronic metabolic acidosis results in significant loss of bone mass due to reduction in bone calcium carbonate. Chronic acidosis also increases urinary calcium excretion, proportional to cumulative acid retention.

TREATMENT

Both uremic acidosis and the hyperchloremic acidosis of renal failure require oral alkali replacement to maintain the [HCO_3^-] between 20 and 24 mmol/L. This can be accomplished with relatively modest amounts of alkali (1.0 to 1.5 mmol/kg body weight per day). It is assumed that alkali replacement prevents the harmful effects of H^+balance on bone and prevents or retards muscle catabolism. Sodium citrate (Shohl's solution) or $NaHCO_3$tablets are equally effective alkalinizing salts. Citrate enhances the absorption of aluminum from the gastrointestinal tract and should never be given together with aluminum-containing antacids because of the risk of aluminum intoxication. When hyperkalemia is present, furosemide (60 to 80 mg/d) should be added.

HYPERCHLOREMIC METABOLIC ACIDOSES

Alkali can be lost from the gastrointestinal tract in diarrhea or from the kidneys (renal tubular acidosis,RTA). In these disorders (Table 50-3), reciprocal changes in [Cl^-] and [HCO_3^-] result in a normalAG. In pure hyperchloremic acidosis, therefore, the increase in [Cl^-] above the normal value approximates the decrease in [HCO_3^-]. The absence of such a relationship suggests a mixed disturbance.

In diarrhea, stools contain a higher [HCO_3^-] and decomposed HCO_3^-than plasma so that metabolic acidosis develops along with volume depletion. Instead of an acid urine pH (as anticipated with systemic acidosis), urine pH is usually around 6 because metabolic acidosis and hypokalemia increase renal synthesis and excretion of NH_4^+, thus providing a urinary buffer that increases urine pH. Metabolic acidosis due to gastrointestinal losses with a high urine pH can be differentiated fromRTA(Chap. 276) because urinary NH_4^+excretion is typically low in RTA and high with diarrhea. Urinary NH_4^+levels can be estimated by calculating the urine anion gap (UAG): UAG=[Na^+ + K^+]$_u$-[Cl^-]$_u$. When [Cl^-]$_u$> [Na^+ + K^+], the urine ammonium level is appropriately increased, suggesting an extrarenal cause of the acidosis.

Loss of functioning renal parenchyma by progressive renal disease leads to hyperchloremic acidosis when the glomerular filtration rate (GFR) is between 20 and 50 mL/min and to uremic acidosis with a highAG when the GFR falls to <20 mL/min. Such a progression occurs commonly with tubulointerstitial forms of renal disease, but hyperchloremic metabolic acidosis can persist with advanced glomerular disease. In advanced renal failure, ammoniagenesis is reduced in proportion to the loss of functional renal mass, and ammonium accumulation and trapping in the outer medullary collecting tubule may also be impaired. Because of adaptive increases in K^+secretion by the collecting duct and colon, the acidosis of chronic renal insufficiency is typically normokalemic.

ProximalRTA(type 2 RTA) is most often due to generalized proximal tubular dysfunction manifested by glycosuria, generalized aminoaciduria, and phosphaturia (Fanconi syndrome). With a low plasma [HCO_3^-], the urine pH is acid (pH < 5.5). The fractional excretion of [HCO_3^-] may exceed 10 to 15% when the serum HCO_3^-> 20 mmol/L. Since $HCO_3^-$is not reabsorbed normally in the proximal tubule, therapy with $NaHCO_3$will enhance renal potassium wasting and hypokalemia.

The typical findings in classic distalRTA(type 1 RTA) (Chap. 276) include hypokalemia, hyperchloremic acidosis, low urinary NH_4^+excretion (positiveUAG, low urine [NH_4^+]), and inappropriately high urine pH (pH > 5.5). Such patients are unable to acidify the urine below a pH of 5.5. Most patients have hypocitraturia and hypercalciuria, so that nephrolithiasis, nephrocalcinosis, and bone disease are common. In type 4 RTA, hyperkalemia is disproportionate to the reduction inGFRbecause of coexisting dysfunction of potassium and acid secretion. Urinary ammonium excretion is invariably depressed, and renal function may be compromised, for example, due to diabetic nephropathy, amyloidosis, or tubulointerstital disease.**See Chap. 276 for the pathophysiology, diagnosis, and treatment of RTA.*

Hyporeninemic Hypoaldosteronism (See also Chap. 331) This condition typically causes hyperchloremic metabolic acidosis, most commonly in older adults with diabetes mellitus or tubulointerstitial disease and renal insufficiency. Patients usually have mild to moderate renal insufficiency and acidosis, with elevation in serum [K^+] (5.2 to 6.0 mmol/L), concurrent hypertension, and congestive heart failure. Both the metabolic acidosis and the hyperkalemia are out of proportion to impairment in GFR. Nonsteroidal anti-inflammatory drugs -- trimethoprim, pentamidine, and ACE-inhibitors -- can also cause hyperkalemia with hyperchloremic metabolic acidosis in patients with renal insufficiency (Table 50-3).

METABOLIC ALKALOSIS

Metabolic alkalosis is manifested by an elevated arterial pH, an increase in the serum [HCO_3^-], and an increase in Pa_{CO_2} as a result of compensatory alveolar hypoventilation. It is often accompanied by hypochloremia and hypokalemia. The patient with a high [HCO_3^-] and a low [Cl^-] has either metabolic alkalosis or chronic respiratory acidosis. As shown inTable 50-1, the Pa_{CO_2}increases 6 mmHg for each 10-mmol/L increase in the [HCO_3^-] above normal. Stated differently, in the range of [HCO_3^-] from 10 to 40 mmol/L, the predicted Pa_{CO_2} is approximately equal to the [HCO_3^-] + 15. The arterial pH establishes the diagnosis, since it is increased in metabolic alkalosis and decreased or normal in respiratory acidosis. Metabolic alkalosis frequently occurs in association with other disorders such as respiratory acidosis or alkalosis or metabolic acidosis.

PATHOGENESIS

Metabolic alkalosis occurs as a result of net gain of [HCO_3^-] or loss of nonvolatile acid (usually HCl by vomiting) from the extracellular fluid. Since it is unusual for alkali to be added to the body, the disorder involves a generative stage, in which the loss of acid usually causes alkalosis, and a maintenance stage, in which the kidneys fail to compensate by excreting HCO_3^-because of volume contraction, a lowGFR, or depletion of Cl^- or K^+.

Under normal circumstances, the kidneys have an impressive capacity to excrete HCO_3^-. Continuation of metabolic alkalosis represents a failure of the kidneys to eliminate HCO_3^-in the usual manner. For $HCO_3^-$to be added to the extracellular fluid, it must be administered exogenously or synthesized endogenously, in part or entirely by the kidneys. The kidneys will retain, rather than excrete, the excess alkali and maintain the alkalosis if (1) volume deficiency, chloride deficiency, and K^+deficiency exist in combination with a reducedGFR, which augments distal tubule H^+secretion; or (2) hypokalemia exists because of autonomous hyperaldosteronism. In the first example, alkalosis is corrected by administration of NaCl and KCl, while in the latter it is necessary to repair the alkalosis by pharmacologic or surgical intervention, not with saline administration.

DIFFERENTIAL DIAGNOSIS

To establish the cause of metabolic alkalosis (Table 50-4), it is necessary to assess the status of the extracellular fluid volume (ECFV), the recumbent and upright blood pressure, the serum [K^+], and the renin-aldosterone system. For example, the presence of chronic hypertension and chronic hypokalemia in an alkalotic patient suggests either mineralocorticoid excess or that the hypertensive patient is receiving diuretics. Low plasma renin activity and normal urine [Na^+] and [Cl^-] in a patient who is not taking diuretics indicate a primary mineralocorticoid excess syndrome. The combination of hypokalemia and alkalosis in a normotensive, nonedematous patient can be due to Bartter's or Gitelman's syndrome, magnesium deficiency, vomiting, exogenous alkali, or diuretic ingestion. Determination of urine electrolytes (especially the urine [Cl^-]) and screening of the urine for diuretics may be helpful. If the urine is alkaline, with an elevated [Na^+] and [K^+] but low [Cl^-], the diagnosis is usually either vomiting (overt or surreptitious) or alkali ingestion. If the urine is relatively acid and has low concentrations of Na^+, K^+, and Cl^-, the most likely possibilities are prior vomiting, the posthypercapnic state, or prior diuretic ingestion. If, on the other hand, neither the urine sodium, potassium, nor chloride concentrations are depressed, magnesium deficiency, Bartter's or Gitelman's syndrome, or current diuretic ingestion should be considered. Bartter's syndrome is distinguished from Gitelman's syndrome because of hypocalciuria and hypomagnesemia in the latter disorder. The genetic and molecular basis of these two disorders has been elucidated recently (Chap. 276).

Alkali Administration Chronic administration of alkali to individuals with normal renal function rarely, if ever causes alkalosis. However, in patients with coexistent hemodynamic disturbances, alkalosis can develop because the normal capacity to excrete HCO_3^-may be exceeded or there may be enhanced reabsorption of HCO_3^-. Such patients include those who receive oral or intravenous HCO_3^-, acetate loads (parenteral hyperalimentation solutions), citrate loads (transfusions), or antacids plus cation-exchange resins (aluminum hydroxide and sodium polystyrene sulfonate).

METABOLIC ALKALOSIS ASSOCIATED WITHECFVCONTRACTION, K^+DEPLETION, AND SECONDARY HYPERRENINEMIC HYPERALDOSTERONISM

Gastrointestinal Origin Gastrointestinal loss of H^+ from vomiting or gastric aspiration results in retention of HCO_3^-. The loss of fluid and NaCl in vomitus or nasogastric suction results in contraction of theECFV and an increase in the secretion of renin and aldosterone. Volume contraction causes a reduction inGFR and an enhanced capacity of the renal tubule to reabsorb HCO_3^-. During active vomiting, there is continued addition of $HCO_3^-$to plasma in exchange for Cl^-, and the plasma [HCO_3^-] exceeds the reabsorptive capacity of the proximal tubule. The excess $NaHCO_3$reaches the distal tubule, where secretion is enhanced by an aldosterone and the delivery of the poorly reabsorbed anion, HCO_3^-. Because of contraction of the ECFV and hypochloremia, Cl^- is avidly conserved by the kidney. Correction of the contracted ECFV with NaCl and repair of K^+deficits corrects the acid-base disorder.

Renal Origin

Diuretics (See alsoChap. 232) Drugs that induce chloruresis, such as thiazides and loop diuretics (furosemide, bumetanide, torsemide, and ethracrynic acid), acutely diminish theECFVwithout altering the total body bicarbonate content. The serum [HCO_3^-] increases. The chronic administration of diuretics tends to generate an alkalosis by increasing distal salt delivery, so that K^+ and H^+secretion are stimulated. The alkalosis is maintained by persistence of the contraction of the ECFV, secondary hyperaldosteronism, K^+deficiency, and the direct effect of the diuretic (as long as diuretic administration continues). Repair of the alkalosis is achieved by providing isotonic saline to correct the ECFV deficit.

Bartter's Syndrome and Gitelman's Syndrome See Chap. 276.

Nonreabsorbable Anions and Magnesium Deficiency Administration of large quantities of nonreabsorbable anions, such as penicillin or carbenicillin, can enhance distal acidification and K^+secretion by increasing the transepithelial potential difference (lumen negative). Mg^{2+}deficiency results in hypokalemic alkalosis by enhancing distal acidification through stimulation of renin and hence aldosterone secretion.

Potassium Depletion Chronic K^+depletion may cause metabolic alkalosis by increasing urinary acid excretion. Both NH_4^+production and absorption are enhanced and HCO_3^-reabsorption is stimulated. Chronic K^+deficiency upregulates the renal H^+, K^+-ATPase to increase K^+absorption at the expense of enhanced H^+secretion. Alkalosis associated with severe K^+depletion is resistant to salt administration, but repair of the K^+deficiency corrects the alkalosis.

After Treatment of Lactic Acidosis or Ketoacidosis When an underlying stimulus for the generation of lactic acid or ketoacid is removed rapidly, as with repair of circulatory insufficiency or with insulin therapy, the lactate or ketones are metabolized to yield an equivalent amount of HCO_3^-. Other sources of new $HCO_3^-$are additive with the original amount generated by organic anion metabolism to create a surfeit of HCO_3^-. Such sources include (1) new HCO_3^-added to the blood by the kidneys as a result of enhanced acid excretion during the preexisting period of acidosis, and (2) alkali therapy during the treatment phase of the acidosis. Acidosis-induced contraction of theECFV and K^+deficiency act to sustain the alkalosis.

Posthypercapnia Prolonged CO_2retention with chronic respiratory acidosis enhances renal HCO_3^-absorption and the generation of new HCO_3^-(increased net acid excretion). If the Pa_{CO2} is returned to normal, metabolic alkalosis results from the persistently elevated [HCO_3^-]. Alkalosis develops if the elevated Pa_{CO2} is abruptly returned toward normal by a change in mechanically controlled ventilation. AssociatedECFVcontraction does not allow complete repair of the alkalosis by correction of the Pa_{CO2}alone, and alkalosis persists until Cl^- supplementation is provided.

METABOLIC ALKALOSIS ASSOCIATED WITH ECFV EXPANSION, HYPERTENSION, AND HYPERALDOSTERONISM

Mineralocorticoid administration or excess production [primary aldosteronism of Cushing's syndrome and adrenal cortical enzyme defects (Chap. 331)] increases net acid excretion and may result in metabolic alkalosis, which may be worsened by associated K^+deficiency.ECFVexpansion from salt retention causes hypertension and antagonizes the reduction in GFRand/or increases tubule acidification induced by aldosterone and by K^+deficiency. The kaliuresis persists and causes continued K^+depletion with polydipsia, inability to concentrate the urine, and polyuria. Increased aldosterone levels may be the result of autonomous primary adrenal overproduction or of secondary aldosterone release due to renal overproduction of renin. In both situations, the normal feedback of ECFV on net aldosterone production is disrupted, and hypertension from volume retention can result.

Liddle's syndrome (Chap. 276) results from increased activity of collecting duct Na^+channel (ENaC) and is a rare inherited disorder associated with hypertension due to volume expansion manifested as hypokalemic alkalosis and normal aldosterone levels.

Symptoms With metabolic alkalosis, changes in central and peripheral nervous system function are similar to those of hypocalcemia (Chap. 340); symptoms include mental confusion, obtundation, and a predisposition to seizures, paresthesia, muscular cramping, tetany, aggravation of arrhythmias, and hypoxemia in chronic obstructive pulmonary disease. Related electrolyte abnormalities include hypokalemia and hypophosphatemia.

TREATMENT

This is primarily directed at correcting the underlying stimulus for HCO_3^-generation. If primary aldosteronism is present, correction of the underlying cause will reverse the alkalosis. [H^+] loss by the stomach or kidneys can be mitigated by the use of H_2receptor blockers, H^+, K^+-ATPase inhibitors, or the discontinuation of diuretics. The second aspect of treatment is to remove the factors that sustain HCO_3^-reabsorption, such asECFVcontraction or K^+deficiency. Although K^+deficits should be repaired, NaCl therapy is usually sufficient to reverse the alkalosis if ECFV contraction is present, as indicated by a low urine [Cl^-].

If associated conditions preclude infusion of saline, renal HCO_3^-loss can be accelerated by administration of acetazolamide, a carbonic anhydrase inhibitor, which is usually effective in patients with adequate renal function but can worsen K^+losses. Dilute hydrochloric acid (0.1 *N* HCl) is also effective but can cause hemolysis. Alternatively, acidification can also be achieved with oral NH_4Cl, which should be avoided in the presence of liver disease. Hemodialysis against a dialysate low in [HCO_3^-] and high in [Cl^-] can be effective when renal function is impaired.

RESPIRATORY ACIDOSIS

Respiratory acidosis can be due to severe pulmonary disease, respiratory muscle fatigue, or abnormalities in ventilatory control and is recognized by an increase in $Pa_{CO_2}$and decrease in pH (Table 50-5). In acute respiratory acidosis, there is an immediate compensatory elevation (due to cellular buffering mechanisms) in HCO_3^-, which increases 1 mmol/L for every 10-mmHg increase in Pa_{CO_2}. In chronic respiratory acidosis (>24 h), renal adaptation increases the [HCO_3^-] by 4 mmol/L for every 10-mmHg increase in Pa_{CO_2}. The serum HCO_3^-usually does not increase above 38 mmol/L.

The clinical features vary according to the severity and duration of the respiratory acidosis, the underlying disease, and whether there is accompanying hypoxemia. A rapid increase in Pa_{CO_2}may cause anxiety, dyspnea, confusion, psychosis, and hallucinations and may progress to coma. Lesser degrees of dysfunction in chronic hypercapnia include sleep disturbances, loss of memory, daytime somnolence, personality changes, impairment of coordination, and motor disturbances such as tremor, myoclonic jerks, and asterixis. Headaches and other signs that mimic raised intracranial pressure, such as papilledema, abnormal reflexes, and focal muscle weakness, are due to vasoconstriction secondary to loss of the vasodilator effects of CO_2.

Depression of the respiratory center by a variety of drugs, injury, or disease can produce respiratory acidosis. This may occur acutely with general anesthetics, sedatives, and head trauma or chronically with sedatives, alcohol, intracranial tumors, and the syndromes of sleep-disordered breathing, including the primary alveolar and obesity-hypoventilation syndromes (Chaps. 263 and264). Abnormalities or disease in the motor neurons, neuromuscular junction, and skeletal muscle can cause hypoventilation via respiratory muscle fatigue. Mechanical ventilation, when not properly adjusted and supervised, may result in respiratory acidosis, particularly if CO_2production suddenly rises (because of fever, agitation, sepsis, or overfeeding) or alveolar ventilation falls because of worsening pulmonary function. High levels of positive end-expiratory pressure in the presence of reduced cardiac output may cause hypercapnia as a result of large increases in alveolar dead space (Chap. 266). Permissive hypercapnia is being used with increasing frequency because of studies suggesting lower mortality rates than with conventional mechanical ventilation, especially with severe central nervous system or heart disease. Although the potential beneficial effects of permissive hypercapnia may be mitigated by correction of the acidemia, it seems prudent, nevertheless, to keep the pH in the range of 7.2 to 7.3 by administration of $NaHCO_3$.

Acute hypercapnia follows sudden occlusion of the upper airway or generalized bronchospasm as in severe asthma, anaphylaxis, inhalational burn, or toxin injury. Chronic hypercapnia and respiratory acidosis occur in end-stage obstructive lung disease. Restrictive disorders involving both the chest wall and the lungs can cause respiratory acidosis because the high metabolic cost of respiration causes ventilatory muscle fatigue. Advanced stages of intrapulmonary and extrapulmonary restrictive defects present as chronic respiratory acidosis.

The diagnosis of respiratory acidosis requires, by definition, the measurement of $Pa_{CO_2}$and arterial pH. A detailed history and physical examination often indicate the cause. Pulmonary function studies (Chap. 250), including spirometry, diffusion capacity for carbon monoxide, lung volumes, and arterial $Pa_{CO_2}$and O_2saturation, usually make it possible to determine if respiratory acidosis is secondary to lung disease. The workup for nonpulmonary causes should include a detailed drug history, measurement of hematocrit, and assessment of upper airway, chest wall, pleura, and neuromuscular function.

TREATMENT

The management of respiratory acidosis depends on its severity and rate of onset. Acute respiratory acidosis can be life-threatening, and measures to reverse the underlying cause should be undertaken simultaneously with restoration of adequate alveolar ventilation. This may necessitate tracheal intubation and assisted mechanical ventilation. Oxygen administration should be titrated carefully in patients with severe obstructive pulmonary disease and chronic CO_2retention who are breathing spontaneously (Chap. 258). When oxygen is used injudiciously, these patients may experience progression of the respiratory acidosis. Aggressive and rapid correction of hypercapnia should be avoided, because the falling Pa_{CO_2}may provoke the same complications noted with acute respiratory alkalosis (i.e., cardiac arrhythmias, reduced cerebral perfusion, and seizures). The Pa_{CO_2}should be lowered gradually in chronic respiratory acidosis, aiming to restore the Pa_{CO_2} to baseline levels and to provide sufficient Cl^- and K^+ to enhance the renal excretion of HCO_3^-.

Chronic respiratory acidosis is frequently difficult to correct, but measures aimed at improving lung function (Chap. 258) can help some patients and forestall further deterioration in most.

RESPIRATORY ALKALOSIS

Alveolar hyperventilation decreases $Pa_{CO_2}$and increases the HCO_3^-/Pa_{CO_2}ratio, thus increasing pH (Table 50-5). Nonbicarbonate cellular buffers respond by consuming HCO_3^-. Hypocapnia develops when a sufficiently strong ventilatory stimulus causes CO_2output in the lungs to exceed its metabolic production by tissues. Plasma pH and $[HCO_3^-]$ appear to vary proportionately with Pa_{CO_2}over a range from 40 to 15 mmHg. The relationship between arterial $[H^+]$ concentration and Pa_{CO_2} is about 0.7 mmol/L per mmHg (or 0.01 pH unit/mmHg), and that for plasma $[HCO_3^-]$ is 0.2 mmol/L per mmHg. Hypocapnia sustained longer than 2 to 6 h is further compensated by a decrease in renal ammonium and titrable acid excretion and a reduction in filtered HCO_3^-reabsorption. Full renal adaptation to respiratory alkalosis may take several days and requires normal volume status and renal function. The kidneys appear to respond directly to the lowered Pa_{CO_2}rather than to alkalosis per se. In chronic respiratory alkalosis a 1-mmHg fall in Pa_{CO_2}causes a 0.4- to 0.5-mmol/L drop in $[HCO_3^-]$ and a 0.3-mmol/L fall (or 0.003 rise in pH) in $[H^+]$.

The effects of respiratory alkalosis vary according to duration and severity but are primarily those of the underlying disease. Reduced cerebral blood flow as a consequence of a rapid decline in Pa_{CO_2}may cause dizziness, mental confusion, and seizures, even in the absence of hypoxemia. The cardiovascular effects of acute hypocapnia in the conscious human are generally minimal, but in the anesthetized or mechanically ventilated patient, cardiac output and blood pressure may fall because of the depressant effects of anesthesia and positive-pressure ventilation on heart rate, systemic resistance, and venous return. Cardiac arrhythmias may occur in patients with heart disease as a result of changes in oxygen unloading by blood from a left shift in the hemoglobin-oxygen dissociation curve (Bohr effect). Acute respiratory alkalosis causes intracellular shifts of Na^+, K^+, and $PO_4^-$and reduces free $[Ca^{2+}]$ by increasing the protein-bound fraction. Hypocapnia-induced hypokalemia is usually minor.

Chronic respiratory alkalosis is the most common acid-base disturbance in critically ill patients and, when severe, portends a poor prognosis. Many cardiopulmonary disorders manifest respiratory alkalosis in their early to intermediate stages, and the finding of normocapnia and hypoxemia in a patient with hyperventilation may herald the onset of rapid respiratory failure and should prompt an assessment to determine if the patient is becoming fatigued. Respiratory alkalosis is common during mechanical ventilation.

The hyperventilation syndrome may be disabling. Paresthesia, circumoral numbness, chest wall tightness or pain, dizziness, inability to take an adequate breath, and, rarely, tetany may themselves be sufficiently stressful to perpetuate the disorder. Arterial blood-gas analysis demonstrates an acute or chronic respiratory alkalosis, often with hypocapnia in the range of 15 to 30 mmHg and no hypoxemia. Central nervous system diseases or injury can produce several patterns of hyperventilation and sustained Pa_{CO2}levels of 20 to 30 mmHg. Hyperthyroidism, high caloric loads, and exercise raise the basal metabolic rate, but ventilation usually rises in proportion so that arterial blood gases are unchanged and respiratory alkalosis does not develop. Salicylates are the most common cause of drug-induced respiratory alkalosis as a result of direct stimulation of the medullary chemoreceptor (Chap. 396). The methylxanthines, theophylline, and aminophylline stimulate ventilation and increase the ventilatory response to CO_2. Progesterone increases ventilation and lowers arterial Pa_{CO2} by as much as 5 to 10 mmHg. Therefore, chronic respiratory alkalosis is a common feature of pregnancy. Respiratory alkalosis is also prominent in liver failure, and the severity correlates with the degree of hepatic insufficiency. Respiratory alkalosis is often an early finding in gram-negative septicemia, before fever, hypoxemia, or hypotension develop.

The diagnosis of respiratory alkalosis depends on measurement of arterial pH and Pa_{CO2}. The plasma $[K^+]$ is often reduced and the $[Cl^-]$ increased. In the acute phase, respiratory alkalosis is not associated with increased renal HCO_3^-excretion, but within hours net acid excretion is reduced. In general, the HCO_3^-concentration falls by 2.0 mmol/L for each 10-mmHg decrease in Pa_{CO2}. Chronic hypocapnia reduces the serum $[HCO_3^-]$ by 5.0 mmol/L for each 10-mmHg decrease in Pa_{CO2}. It is unusual to observe a plasma $HCO_3^- < 12$ mmol/L as a result of a pure respiratory alkalosis.

When a diagnosis of respiratory alkalosis is made, its cause should be investigated. The diagnosis of hyperventilation syndrome is made by exclusion. In difficult cases, it may be important to rule out other conditions such as pulmonary embolism, coronary artery disease, and hyperthyroidism.

TREATMENT

The management of respiratory alkalosis is directed toward alleviation of the underlying disorder. If respiratory alkalosis complicates ventilator management, changes in dead space, tidal volume, and frequency can minimize the hypocapnia. Patients with the hyperventilation syndrome may benefit from reassurance, rebreathing from a paper bag during symptomatic attacks, and attention to underlying psychological stress. Antidepressants and sedatives are not recommended. b-Adrenergic blockers may ameliorate peripheral manifestations of the hyperadrenergic state.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 8 -ALTERATIONS IN SEXUAL FUNCTION AND REPRODUCTION

51. ERECTILE DYSFUNCTION - *Kevin T. McVary*

Erectile dysfunction (ED) affects 10 to 25% of middle-aged and elderly men. Demographic changes, the popularity of newer treatments, and greater acceptance of ED by patients and society have led to increased diagnosis and associated health care expenditures for the management of this common disorder. Impairment of erectile function has a profound impact on the well-being of affected men. Because many patients are reluctant to initiate discussion of sexual function, the physician should address this topic directly to elicit a history of ED.

PHYSIOLOGIC CONTROL OF ERECTION AND MALE SEXUAL FUNCTION

Normal male sexual function requires (1) an intact libido, (2) the ability to achieve and maintain penile erection, (3) ejaculation, and (4) detumescence. *Libido* refers to sexual desire and is influenced by a variety of visual, olfactory, tactile, auditory, imaginative and hormonal stimuli. Sex steroids, particularly testosterone, act to increase libido. Libido can be diminished by hormonal or psychiatric disorders or by medications.

The major anatomic structures of the penis that are involved in erectile function include the three corpora, which consist of the paired cavernosa and a single spongiosum that encloses the urethra. A collagenous sheath, called the *tunica albuginea*, individually surrounds each corpora. The micro-architecture of the corpora is composed of a mass of smooth muscle (trabecula) which contains a network of endothelial-lined vessels (lacunar spaces).

Penile tumescence leading to erection depends on the increased flow of blood into the lacunar network after complete relaxation of the arteries and corporal smooth muscle. Subsequent compression of the trabecular smooth muscle against the fibroelastic tunica albuginea causes a passive closure of the emissary veins and accumulation of blood in the corpora. In the presence of a full erection and a competent valve mechanism, the corpora become noncompressible cylinders from which blood does not escape.

The central nervous system exerts an important influence by either stimulating or antagonizing spinal pathways that mediate erectile function and ejaculation. The erectile response is mediated by a combination of central (psychogenic) and peripheral (reflexogenic) innervation. Sensory nerves that originate from receptors in the penile skin and glans converge to form the dorsal nerve of the penis, which travels to the S2-S4 dorsal root ganglia via the pudendal nerve. Parasympathetic nerve fibers to the penis arise from neurons in the intermediolateral columns of S2-S4 sacral spinal segments. Sympathetic innervation originates from the T-11 to the L-2 spinal segments and descends through the hypogastric plexus.

Neural input to smooth muscle tone is crucial to the initiation and maintenance of an erection. There is also an intricate interaction between the corporal smooth muscle cell and its overlying endothelial cell lining (Fig. 51-1*A*). Nitric oxide, which induces vascular relaxation, promotes erection and is opposed by endothelin-1 (ET-1), which mediates vascular contraction. Nitric oxide is synthesized from L-arginine by nitric oxide synthase, and is released from the nonadrenergic, noncholinergic (NANC) autonomic nerve supply to act postjunctionally on smooth muscle cells. Nitric oxide increases the production of cyclic 3¢,5¢-guanosine monophosphate (cyclic GMP), which interacts with protein kinase G and decreases intracellular calcium, causing relaxation of the smooth muscle (Fig. 51-1*B*). Cyclic GMP is gradually broken down by phosphodiesterase type 5 (PDE-5). Inhibitors of PDE-5, such as the oral medication sildenafil, maintain erections by reducing the breakdown of cyclic GMP. However, if nitric oxide is not produced at some level, the addition of PDE-5 inhibitor is not effective, as the drug facilitates but does not initiate the initial enzyme cascade. In addition to nitric oxide, vasoactive prostaglandins (PGE_1, PGF_{2a}) are synthesized within the cavernosal tissue and increase cyclic AMP levels, also leading to relaxation of cavernosal smooth muscle cells.

Ejaculation is stimulated by the sympathetic nervous system, which results in contraction of the epididymis, vas deferens, seminal vesicles, and prostate, causing seminal fluid to enter the urethra. Seminal fluid emission is followed by rhythmic contractions of the bulbocavernosus and ischiocavernosus muscles, leading to ejaculation. *Premature ejaculation* is usually related to anxiety or a learned behavior and is amenable to behavioral therapy or treatment with medications such as selective serotonin reuptake inhibitors (SSRIs). *Retrograde ejaculation* results when the internal urethral sphincter does not close, and it may occur in men with diabetes or after surgery involving the bladder neck.

Detumescence is mediated by released norepinephrine from the sympathetic nerves, release of endothelin from the vascular surface, and contraction of smooth muscle induced by activation of postsynaptic a-adrenergic receptors. These events increase venous outflow and restore the flaccid state. Venous leak can cause premature detumescence and is thought to be caused by insufficient relaxation of the corporal smooth muscle rather than a specific anatomic defect. *Priapism* refers to a persistent and painful erection and may be associated with sickle cell anemia, hypercoagulable states, spinal cord injury, or injection of vasodilator agents into the penis.

ERECTILE DYSFUNCTION

EPIDEMIOLOGY

In the Massachusetts Male Aging Study (MMAS), a community-based survey of men between the ages of 40 and 70, 52% of responders reported some degree ofED. Complete ED occurred in 10% of respondents, moderate ED occurred in 25%, and minimal ED in 17%. The incidence of moderate or severe ED more than doubled between the ages of 40 and 70. In the National Health and Social Life Survey (NHSLS), which was a nationally representative sample of men and women age 18 to 59 years, 10% of men reported being unable to maintain an erection (corresponding to the proportion of men in the MMAS reporting severe ED). Incidence was highest among men in the 50 to 59 age group (21%) and among men who were poor (14%), divorced (14%), and less educated (13%).

The incidence ofED is also higher among men with certain medical disorders. In theMMAS, ED correlated with the presence of diabetes mellitus, heart disease, hypertension, and decreased HDL levels. Medications used to treat diabetes or cardiovascular disease are additional risk factors (see below). There is a higher incidence of ED among men who have undergone radiation or surgery for cancer of the prostate and in those with a lower spinal cord injury. Psychological causes of ED include depression and anger. TheNHSLSfound a higher incidence of ED among men who reported fair-to-poor health or experienced stress from unemployment or other causes. ED is not considered a normal part of the aging process. Nonetheless, it is associated with certain physiologic and psychological changes related to age.

PATHOPHYSIOLOGY

EDmay result from three basic mechanisms: (1) failure to initiate (psychogenic, endocrinologic, or neurogenic); (2) failure to fill (arteriogenic); or (3) failure to store (venoocclusive dysfunction) adequate blood volume within the lacunar network. The inability to initiate an erection may have psychogenic, endocrinologic, or neurogenic etiologies. These categories are not mutually exclusive, and multiple factors contribute to ED in many patients. For example, diminished filling pressure can lead secondarily to venous leak. Psychogenic factor frequently co-exist with other etiologic factors and should be considered in all cases. Diabetic, atherosclerotic, and drug-related causes account for>80% of cases of ED in older men.

Vasculogenic The most frequent organic cause ofED is a disturbance of blood flow to and from the penis. Atherosclerotic or traumatic arterial disease can decrease flow to the lacunar spaces, resulting in decreased rigidity and an increased time to full erection. Excessive outflow through the veins, despite adequate inflow, may also contribute to ED. In this case, the achieved perfusion pressures cannot compensate for the unrestricted outflow needed to ensure adequate erection. This situation may be due to insufficient relaxation of trabecular smooth muscle and may occur in anxious individuals with excessive adrenergic tone or in those with damaged parasympathetic outflow. Structural alterations to the fibroelastic components of the corpora may cause a loss of compliance and an inability to compress the tunical veins. This condition may result from aging, increased cross-leaking of collagen fibers induced by nonenzymatic glycosylation, hypoxia, or altered synthesis of collagen associated with hypercholesterolemia. Fibroelastic structures can also be damaged by surgery, radiation, or trauma to the penis.

Neurogenic Disorders that affect the sacral spinal cord or the autonomic fibers to the penis preclude nervous system relaxation of penile smooth muscle, thus leading toED. In patients with spinal cord injury, the degree of ED depends on the completeness and level of the lesion. Patients with incomplete lesions or injuries to the upper part of the spinal cord are more likely to retain erectile capabilities than those with complete lesions or injuries to the lower part. Although 75% of patients with spinal cord injuries have some erectile capability, only 25% have erections sufficient for penetration. Other neurologic disorders commonly associated with ED include multiple sclerosis and peripheral neuropathy. The latter is often due to either diabetes or alcoholism. Pelvic surgery may cause ED through disruption of the autonomic nerve supply.

Endocrinologic Androgens increase libido, but their exact role in erectile function remains unclear. Individuals with castrate levels of testosterone can achieve erections from visual or sexual stimuli. Nonetheless, normal levels of testosterone appear to be important for erectile function, particularly in older males. Androgen replacement therapy can improve depressed erectile function when it is secondary to hypogonadism; it is not useful for ED when endogenous testosterone levels are normal. Increased prolactin may decrease libido by suppressing gonadotropin-releasing hormone (GnRH), and it also leads to decreased testosterone levels. Treatment of hyperprolactinemia with dopamine agonists can restore libido and testosterone.

Diabetic EDoccurs in 35 to 75% of men with diabetes mellitus. Pathologic mechanisms are primarily related to diabetes-associated vascular and neurologic complications. Diabetic macrovascular complications are mainly related to age, whereas microvascular complications correlate with the duration of diabetes and the degree of glycemic control (Chap. 333). Individuals with diabetes also have reduced amounts of nitric oxide synthase in both endothelial and neural tissues.

Psychogenic Two mechanisms contribute to the inhibition of erections in psychogenicED. First, psychogenic stimuli to the sacral cord may inhibit reflexogenic responses, thereby blocking activation of vasodilator outflow to the penis. Second, excess sympathetic stimulation in an anxious man may increase penile smooth muscle tone. The most common causes of psychogenic ED are performance anxiety, depression, relationship conflict, loss of attraction, sexual inhibition, conflicts over sexual preference, sexual abuse in childhood, and fear of pregnancy or sexually transmitted disease. Almost all patients with ED, even when it has a clear-cut organic basis, develop a psychogenic component as a reaction to ED.

Medication-Related Medication-inducedED(Table 51-1) is estimated to occur in 25% of men seen in general medical outpatient clinics. Among the antihypertensive agents, the thiazide diuretics and beta blockers have been implicated most frequently. Calcium channel blockers and angiotensin-converting enzyme inhibitors are less frequently cited. These drugs may act directly at the corporal level (e.g., calcium channel blockers) or indirectly by reducing pelvic blood pressure, which is important in the development of penile rigidity. Alpha adrenergic blockers are less likely to cause ED. Estrogens,GnRHagonists, H_2antagonists, and spironolactone cause ED by suppressing gonadotropin production or by blocking androgen action. Antidepressant and antipsychotic agents -- particularly neuroleptics, tricyclics, andSSRIs -- are associated with erectile, ejaculatory, orgasmic, and sexual desire difficulties. Digoxin induces ED via blockade of the Na^+,K^+-ATPase pump, resulting in a net increase in intracellular calcium and increased corporal smooth muscle tone.

Although many medications can causeED, patients frequently have concomitant risk factors that confound the clinical picture. If there is a strong association between the institution of a drug and the onset of ED, alternative medications should be considered. Otherwise, it is often practical to treat the ED without attempting multiple changes in medications, as it may be difficult to establish a causal role for the drug.

CLINICAL EVALUATION

A good physician-patient relationship helps to unravel the possible causes ofED, many of which require discussion of personal and sometimes embarrassing topics. For this reason, a primary care provider is often ideally suited to initiate the evaluation. A complete medical and sexual history should be taken in an effort to assess whether the cause of ED is organic, psychogenic, or multifactorial (Fig. 51-2). Initial questions should focus on the onset of symptoms, the presence and duration of partial erections, and the progression of ED. A history of nocturnal or early morning erections is useful for distinguishing physiologic from psychogenic ED. Nocturnal erections occur during rapid eye movement (REM) sleep and require intact neurologic and circulatory systems. Organic causes of ED are generally characterized by a gradual and persistent change in rigidity or the inability to sustain nocturnal, coital, or self-stimulated erections. The patient should also be questioned about the presence of penile curvature or pain with coitus. It is also important to address libido, as decreased sexual drive and ED are sometimes the earliest signs of endocrine abnormalities (e.g., increased prolactin, decreased testosterone levels). It is useful to ask whether the problem is confined to coitus with one or other partners; ED arises not uncommonly in association with new or extramarital sexual relationships. Situational ED, as opposed to consistent ED, suggests psychogenic causes. Ejaculation is much less commonly affected than erection, but questions should be asked about whether ejaculation is normal, premature, delayed, or absent. Relevant risk factors should be identified, such as diabetes mellitus, coronary artery disease, lipid disorders, hypertension, peripheral vascular disease, smoking, alcoholism, and endocrine or neurologic disorders. The patient's surgical history should be explored with an emphasis on bowel, bladder, prostate, or vascular procedures. A complete drug history is also important, as medications constitute a major source of reversible ED. Social changes that may precipitate ED are also crucial to the evaluation, including health worries, spousal death, divorce, relationship difficulties, and financial concerns.

The physical examination is an essential element in the assessment ofED. Signs of hypertension as well as evidence of thyroid, hepatic, hematologic, cardiovascular, or renal diseases should be sought. An assessment should be made of the endocrine and vascular systems, the external genitalia, and the prostate gland. The penis should be carefully palpated along the corpora to detect fibrotic plaques. Reduced testicular size and loss of secondary sexual characteristics are suggestive of hypogonadism. Neurologic examination should include assessment of anal sphincter tone, the bulbocavernosus reflex, and testing for peripheral neuropathy.

Selected laboratory testing is recommended in all cases. Although hyperprolactinemia is uncommon, a serum prolactin level should be measured, as decreased libido and/or erectile dysfunction may be the presenting symptoms of a prolactinoma or other mass lesions of the sella (Chap. 328). The serum testosterone level should be measured and, if low, gonadotropins should be measured to determine whether hypogonadism is primary (testicular) or secondary (hypothalamic-pituitary) in origin (Chap. 335). Serum chemistries, CBC, and lipid profiles may be of value, if not performed recently, as they can yield evidence of anemia, diabetes, hyperlipidemia, or other systemic diseases associated withED. Determination of serum PSA should be conducted according to recommended clinical guidelines (Chap. 95).

Additional diagnostic testing is rarely necessary in the evaluation of ED. However, in selected patients, specialized testing may provide insight into pathologic mechanisms of ED and aid in the selection of treatment options. Optional specialized testing includes: (1) studies of nocturnal penile tumescence and rigidity; (2) vascular testing (in-office injection of vasoactive substances, penile Doppler ultrasound, penile angiography, dynamic infusion cavernosography/cavernosometry); (3) neurologic testing (biothesiometry-graded vibratory perception; somatosensory evoked potentials); and (4) psychological diagnostic tests. The information potentially gained from these procedures must be balanced against their invasiveness and cost.

TREATMENT

Patient Education Patient and partner education is essential in the treatment ofED. In goal-directed therapy, education facilitates understanding of the disease, results of the tests, and selection of treatment. Discussion of treatment options helps to clarify how treatment is best offered, and to stratify first- and second-line therapies. Patients with high-risk lifestyle issues, such as smoking, alcohol abuse, or recreational drug use, should be counseled on the role these factors play in the development of ED.

Oral Agents Sildenafil is the only approved and effective oral agent for the treatment ofED. Sildenafil has markedly improved the management of ED because it is effective for the treatment of a broad range of causes of ED, including psychogenic, diabetic, vasculogenic, post-radical prostatectomy (nerve-sparing procedures), and spinal cord injury. Sildenafil is a selective and potent inhibitor of PDE-5, the predominant phosphodiesterase isoform found in the penis. It is administered in doses of 25, 50, or 100 mg, and enhances erections after sexual stimulation. The onset of action is approximately 60 to 90 min. Reduced initial doses should be considered for patients who are elderly, have renal insufficiency, or are taking medications that inhibit the CYP3A4 metabolic pathway in the liver (e.g., erythromycin, cimetidine, ketoconazole, and, possibly, itraconazole and mibefradil), as they may increase the serum concentration of sildenafil. The drug does not affect ejaculation, orgasm, or sexual drive. Side effects associated with sildenafil include headaches (19%), facial flushing (9%), dyspepsia (6%) and nasal congestion (4%). Approximately 7% of men may experience transient altered color vision (blue halo effect). Sildenafil is contraindicated in men receiving nitrate therapy for cardiovascular disease, including agents delivered by oral, sublingual, transnasal, or topical routes. These agents can potentiate its hypotensive effect and may result in profound shock. Likewise, amyl/butyl nitrates (poppers) may have a fatal synergistic effect on blood pressure. Sildenafil should also be avoided in patients with congestive heart failure and cardiomyopathy because of the risk of vascular collapse. Because sexual activity leads to an increase in physiologic expenditure [5 to 6 metabolic equivalents (METS)], physicians have been advised to exercise caution in prescribing any drug for sexual activity to those with active coronary disease, heart failure, borderline hypotension, hypovolemia, and to those on complex antihypertensive regimens.

Androgen Therapy Testosterone replacement is used to treat both primary and secondary causes of hypogonadism (Chap. 335). Androgen supplementation in the setting of normal testosterone is rarely efficacious and is discouraged. Methods of androgen replacement include parenteral administration of long-acting testosterone esters (enanthate and cypionate), oral preparations (17a-alkylated derivatives), and transdermal patches (Chap. 335). The long-acting 17b-hydroxy esters of testosterone are the safest, most cost-effective, and practical preparations available. The administration of 200 to 300 mg intramuscularly every 2 to 3 weeks provides a practical option but is far from an ideal physiologic replacement. Oral androgen preparations have the potential for hepatotoxicity and should be avoided. Transdermal delivery of testosterone more closely mimics physiologic testosterone levels, but it is unclear whether this translates into improved sexual function. Because testosterone gradually decreases into the hypogonadal range by 24 hours, patches need to replaced daily. Testosterone therapy is contraindicated in men with androgen-sensitive cancers and may be inappropriate for men with bladder neck obstruction. It is generally advisable to measure PSA before giving androgen. Hepatic function should be tested before and during testosterone therapy.

Vacuum Constriction Devices Vacuum constriction devices (VCD) are a well-established, noninvasive therapy. They are a reasonable treatment alternative for select patients who cannot take sildenafil or do not desire other interventions. VCD draw venous blood into the penis and use a constriction ring to restrict venous return and maintain tumescence. Adverse events with VCD include pain, numbness, bruising, and altered ejaculation. Additionally, many patients complain that the devices are cumbersome and that the induced erections have a non-physiologic appearance.

Intraurethral Alprostadil If a patient fails to respond to oral agents, a reasonable next choice is intraurethral or self-injection or vasoactive substances. Intraurethral prostaglandin E_1(alprostadil), in the form of a semisolid pellet (doses of 125 to 1000 ug), is delivered with an applicator. Approximately 65% of men receiving intraurethral alprostadil respond with an erection when tested in the office, but only 50% of those achieve successful coitus at home. Intraurethral insertion is associated with a markedly reduced incidence of priapism in comparison to intracavernosal injection.

Intracavernosal Self-Injection Injection of synthetic formulations of alprostadil is effective in 70 to 80% of patients withED, but discontinuation rates are high because of the invasive nature of administration. Doses range between 1 and 40 ug. Injection therapy is contraindicated in men with a history of hypersensitivity to the drug and in men at risk for priapism (hypercoagulable states, sickle cell disease). Side effects include local adverse events, prolonged erections, pain, and fibrosis with chronic use. Various combinations of alprostadil, phentolamine, and/or papaverine are sometimes used.

Surgery A less frequently used form of therapy forEDinvolves the surgical implantation of a semi-rigid or inflatable penile prosthesis. These surgical treatments are invasive, associated with potential complications, and generally reserved for treatment of refractory ED. Despite their high cost and invasiveness, penile prostheses are associated with high rates of patient satisfaction.

Sex Therapy A course of sex therapy may be useful for addressing specific interpersonal factors that may affect sexual functioning. Sex therapy generally consists of in-session discussion and at-home exercises specific to the person and the relationship. It is preferable if therapy includes both partners, provided the patient is involved in an ongoing relationship.

(Bibliography omitted in Palm version)

Back to Table of Contents

52. DISTURBANCES OF MENSTRUATION AND OTHER COMMON GYNECOLOGIC COMPLAINTS IN WOMEN - *Bruce R. Carr*, *Karen D. Bradshaw*

Complaints related to the female reproductive tract can be categorized as disorders of menstruation, pelvic pain, disturbances in sexual function, or infertility. However, a single disorder, e.g., leiomyoma of the uterus, can present with symptoms referable to any one or more of these categories. Furthermore, sexual dysfunction can interdigitate with other problems in several ways. On the one hand, in women with complaints related to other reproductive tract functions, the underlying problem may actually be severe sexual dysfunction or marital conflict. Alternatively, women with severe organic disorders of the pelvis, e.g., pelvic inflammatory disease or endometriosis, may present with sexual dysfunction such as dyspareunia that in fact is only a minor manifestation of the underlying disease.

Since normal reproductive function depends on the integrated action of the central nervous system, the endocrine glands, and the reproductive organs, menstrual cycle abnormalities, sexual dysfunction, and infertility may be the result of systemic and psychological disorders as well as of primary defects in the endocrine and reproductive organs. The endocrine and physiologic control -- normal and abnormal -- of puberty, reproductive life, and menopause are discussed in Chap. 336. The focus of this chapter is on the initial evaluation of women with disturbances of the reproductive tract.

DISTURBANCES IN MENSTRUATION

Disorders of menstruation can be divided into abnormal uterine bleeding and amenorrhea.

Abnormal Uterine Bleeding The menstrual cycle is defined as the interval between the onset of one bleeding episode and the onset of the next. In normal women the cycle averages 28± 3 days, the mean duration of menstrual flow is 4 ± 2 days, and the average blood loss is 35 to 80 mL. Between menarche and menopause most women experience one or more episodes of abnormal uterine bleeding, here defined as any bleeding pattern outside the normal ranges of frequency, duration, and/or amount of blood loss. The decision to evaluate a patient depends on the severity and frequency of the abnormal bleeding pattern.

When vaginal bleeding occurs, it should first be determined whether the blood is derived from the uterine endometrium. Rectal, bladder, cervical, and vaginal sources of bleeding must be excluded. Once the bleeding is established to be uterine in origin, a pregnancy-related disorder (such as threatened or incomplete abortion or ectopic pregnancy) must be ruled out by physical examination and appropriate laboratory tests. It should also be remembered that uterine bleeding may also be the initial or principal manifestation of a generalized bleeding diathesis. The remaining causes of abnormal uterine bleeding can be divided into those associated with ovulatory or anovulatory cycles.

Ovulatory Cycles Menstrual bleeding with ovulatory cycles is spontaneous, regular in onset, predictable in duration and amount of flow, and frequently associated with discomfort; it is the consequence of progesterone withdrawal at the end of the luteal (postovulatory) phase and requires prior estrogen priming of the endometrium during the follicular (preovulatory) phase of the cycle. When deviations from an established pattern of menstrual flow occur but the cycles are still regular, the usual cause is disease of the outflow tract. For example, regular, prolonged, excessive bleeding episodes can result from abnormalities of the uterus such as submucous leiomyomas, adenomyosis, or endometrial polyps. On the other hand, cyclic, predictable menstruation characterized by spotting or light bleeding suggests obstruction of the outflow tract as with uterine synechiae or scarring of the cervix. Intermittent bleeding between cyclic ovulatory menses is often due to cervical or endometrial lesions.

Anovulatory Cycles Uterine bleeding that is irregular in occurrence, unpredictable as to amount and duration of flow, and usually painless is called *dysfunctional or anovulatory uterine bleeding*. This type of bleeding is the result of a failure of normal follicular maturation with consequent anovulation and may be either transient or chronic. Transient disruption of ovulatory cycles occurs most often in the early menarcheal years, during the perimenopausal period, or as the consequence of a variety of stresses and intercurrent illnesses. Persistent dysfunctional uterine bleeding during the reproductive years can occur in several organic diseases that affect ovarian function and is most often due to estrogen breakthrough bleeding. Estrogen breakthrough bleeding occurs when estrogen stimulation of the endometrium is continuous and is not interrupted by cyclic progesterone withdrawal, as can occur in polycystic ovarian disease.

Amenorrhea *Amenorrhea* is defined either as failure of menarche by age 16, regardless of the presence or absence of secondary sexual characteristics, or as the absence of menstruation for 6 months in a woman with previous periodic menses. Amenorrhea in a woman who has never menstruated is termed *primary amenorrhea*; cessation of menses is termed *secondary amenorrhea*. Because some disorders can cause both primary and secondary amenorrhea, we prefer a functional classification based on the nature of the underlying defect, namely, anatomic defects of the outflow tract (uterus, cervix, or vagina), ovarian failure, and chronic anovulation.

Anatomic defects of the outflow tract include congenital defects of the vagina, imperforate hymen, transverse vaginal septa, cervical stenosis, intrauterine adhesions (synechiae), absence of the vagina or uterus, and uterine maldevelopment. The diagnosis of an anatomic defect is usually made by physical examination but may be confirmed by demonstrating failure of bleeding following administration of estrogen plus a progestogen for 21 days. Pelvic ultrasonography, magnetic resonance imaging, hysterosalpingogram, or hysteroscopy may be helpful in defining the defect.

Causes of *ovarian failure* include gonadal dysgenesis, deficiency of 17a-hydroxylase, resistant ovary syndrome, and premature ovarian failure. Ovarian failure encompasses disorders in which the ovary is deficient in germ cells and those in which the germ cells are resistant to follicle-stimulating hormone (FSH). The diagnosis of ovarian failure as the cause of amenorrhea is confirmed by an elevated plasma FSH level.

Women with *chronic anovulation* fail to ovulate spontaneously but have the capability of ovulating with appropriate therapy. In some women with chronic anovulation, total estrogen production is adequate, but it is not secreted in a cyclic fashion. In others, estrogen production is deficient.

Women who have adequate estrogen production and demonstrate withdrawal bleeding after progestogen challenge often have polycystic ovarian disease (seeFig. 336-8). Other causes include hormone-secreting ovarian and adrenal tumors. Women with deficient or absent estrogen production, and therefore with absence of withdrawal bleeding after progestogen administration, usually have hypogonadotropic hypogonadism due to organic or functional disorders of the pituitary or central nervous system such as brain tumors, pituitary tumors (especially prolactin-secreting adenomas), primary hypopituitarism, or Sheehan's syndrome.

PELVIC PAIN

Pelvic pain may originate in the pelvis or be referred from another region of the body. A pelvic source is suggested by the history (e.g., dysmenorrhea and dyspareunia) and physical findings, but a high index of suspicion must be entertained for extrapelvic disorders that refer to the pelvis, such as appendicitis, diverticulitis, cholecystitis, intestinal obstruction, and urinary tract infections (Chap. 14).

"Physiologic" Pelvic Pain

Pain Associated with Ovulation ("Mittelschmerz") Many women experience low abdominal discomfort with ovulation, typically a dull aching pain at midcycle in one lower quadrant lasting from minutes to hours. It is rarely severe or incapacitating. The pain may result from peritoneal irritation by follicular fluid released into the peritoneal cavity at ovulation. The onset at midcycle and short duration of pain suggest this diagnosis.

Premenstrual or Menstrual Pain In normal ovulatory women, somatic symptoms during the few days prior to menses may be insignificant or disabling. Such symptoms include edema, breast engorgement, and abdominal bloating or discomfort. A symptom complex of cyclic irritability, depression, and lethargy is known as the *premenstrual syndrome* (PMS). PMS appears to be caused by changes in gonadal steroid levels. Although there is no consensus about therapy, randomized, controlled trials suggest significant improvement with the daily use of serotonin-reuptake inhibitors.

Severe or incapacitating uterine cramping during ovulatory menses and in the absence of demonstrable disorders of the pelvis is termed *primary dysmenorrhea*. Primary dysmenorrhea is caused by prostaglandin-induced uterine ischemia and is treated with prostaglandin synthetase inhibitors and/or oral contraceptive agents.

Pelvic Pain due to Organic Causes Severe dysmenorrhea associated with disease of the pelvis is termed *secondary dysmenorrhea.* Organic causes of pelvic pain can be classified as (1) uterine, (2) adnexal, (3) vulvar or vaginal, and (4) pregnancy-associated.

Uterine Pain Pain of uterine etiology is often chronic and continuous and increases in intensity during menstruation and intercourse. Causes include leiomyomas of the uterus (particularly submucous and degenerating leiomyomas), adenomyosis, and cervical stenosis. Infections of the uterus associated with intrauterine manipulation following dilatation and curettage or with the insertion of intrauterine devices can also cause pelvic pain (Chap. 336). Pelvic pain due to endometrial or cervical cancer is usually a late manifestation (Chap. 336).

Adnexal Pain The most common cause of pain in the adnexae (fallopian tubes and ovaries) is infection (Chap. 133). Acute salpingo-oophoritis presents as low abdominal pain, fever, and chills; begins a few days after a menstrual period; and is usually due to chlamydial or gonococcal disease with or without a superimposed pyogenic infection. Chronic pelvic inflammatory disease results from either a single episode or multiple episodes of infection and may present as infertility associated with chronic pelvic pain that increases in intensity with menses and intercourse. On physical examination, cervical motion tenderness, adnexal tenderness, and adnexal thickening and/or masses may be present. Pelvic inflammatory disease may become a surgical emergency if peritonitis results from rupture of a tuboovarian abscess. Ovarian cysts or neoplasms may cause pelvic pain that becomes more severe with torsion or rupture of the mass, and ectopic pregnancy must be considered in the differential diagnosis (see below). Endometriosis involving fallopian tubes, ovaries, or peritoneum may cause both chronic low abdominal pain and infertility; the magnitude of tissue involvement does not always correlate with the severity of symptoms. Endometriosis pain typically increases with menstruation and, if the posterior ligaments of the uterus are involved, with intercourse.

Vulvar or Vaginal Pain Pain in these areas is most often due to infectious vaginitis caused by *Monilia*, *Trichomonas*, or bacteria and is characteristically associated with vaginal discharge and pruritus. Herpetic vulvitis, other dermatologic conditions of the vulva, condyloma acuminatum, and cysts or abscesses of Bartholin's glands may also cause vulvar pain.

Pregnancy-Associated Disorders Pregnancy must be considered in the differential diagnosis of pelvic pain during the reproductive years. Threatened abortion or incomplete abortion often presents with uterine cramping, bleeding, or passage of tissue following a period of amenorrhea. Ectopic pregnancy may be insidious in presentation or result in abrupt intraperitoneal hemorrhage and maternal death.

Evaluation of Pelvic Pain The evaluation of pelvic pain requires a careful history and pelvic examination. This often leads to the correct diagnosis and institution of appropriate treatment. If the pain is severe and the diagnosis is unclear, the workup should follow that outlined for the acute abdomen (Chap. 14). A culdocentesis may be indicated if a ruptured ectopic pregnancy is suspected. If there is a question of an adnexal mass or if the patient is so obese as to preclude a thorough pelvic examination, abdominal or vaginal sonography may be useful. Serial human chorionic gonadotropin (hCG) measurements may help in establishing a diagnosis of tubal pregnancy and are useful in determining if an intrauterine pregnancy is viable. Finally, diagnostic laparoscopy and laparotomy may be indicated with pain of undetermined etiology.

SEXUAL DYSFUNCTION

Some women with sexual dysfunction describe minor complaints related to the reproductive tract as a means of bringing sexual problems to the attention of the physician. Alternatively, sexual dysfunction may be thought to be the cause of low abdominal discomfort or dyspareunia when the actual etiology is organic. However, more and more women seek medical advice because of sexual problems that interface in provenance between medicine, psychiatry, and sociology.

The normal sexual response begins with sexual arousal, which causes genital vasocongestion that results in vaginal lubrication in preparation for intromission. The lubrication is due to the formation of a transudate in the vagina and in conjunction with genital congestion produces the so-called orgasmic platform prior to orgasm. Sexual stimuli (visual, tactile, auditory, and olfactory) as well as healthy vaginal tissue are prerequisites for genital vasocongestion and vaginal lubrication. During the second stage of the sexual response, involuntary contractions of the muscles of the pelvis result in a pleasurable cortical sensory phenomenon known as orgasm. Direct or indirect stimulation of the clitoris is important in the production of the female orgasm. In simple terms, sexual dysfunction can be due to interference with the arousal or orgasmic phases of the sexual response. Either disorder can be due to an organic or functional cause or both.

Illnesses that impair neurologic function such as diabetes mellitus or multiple sclerosis can prevent normal sexual arousal. Local pelvic diseases such as vaginitis, endometriosis, and salpingo-oophoritis may preclude normal sexual response because of resulting dyspareunia. Debilitating systemic diseases such as cancer and cardiovascular diseases may inhibit normal sexual response indirectly.

More commonly, failure of a normal sexual response is due to psychological factors that impair sexual arousal. Such problems include misinformation, e.g., the perception of sexual satisfaction as bad, or feelings of guilt about previous psychologically traumatic events such as incest, rape, or unwanted pregnancy. In addition, women who have had previous hysterectomy or mastectomy may perceive themselves as "incomplete." Stresses such as anxiety, depression, fatigue, and marital or interpersonal conflicts may lead to failure of the vasocongestive response and prevent normal vaginal lubrication. Women with such experiences may be unable to achieve normal sexual response unless they receive professional counseling. Such problems are approached by attempting to identify and reduce the causative stresses.

Failure to achieve orgasm is a specific form of sexual dysfunction. In the absence of orgasm many women enjoy sexual encounters to variable degrees because of the pleasure derived from closeness in a cherished relationship, particularly with a loving partner. However, for other women sexual relations with rare or absent orgasms are frustrating and unsatisfying. In many instances, failure of orgasm is due to insufficient clitoral stimulation and may be rectified by appropriate counseling and patient education.

A specific entity, "vaginismus," painful, involuntary contractions of the musculature surrounding the entrance to the vagina, is a rare cause of dyspareunia. It is a conditioned response to a previous real or imagined frightening or traumatic sexual experience. Treatment is directed to elimination of the conditioned response by progressive vaginal dilation by the patient in conjunction with marital therapy.

REPRODUCTION

Infertility is discussed in detail in Chap. 54. The approach to infertile couples always involves evaluation of both the man and woman. The history should address the frequency of intercourse, the sexual responses of both, the use of contraceptives or lubricants, prior pregnancies, interval to conception and outcome of pregnancy, previous or past medical illnesses, and all medications taken.

Male-associated factors account for a third of infertility problems. Therefore, one of the first procedures in the workup of infertile couples should be a semen analysis. The initial evaluation of the woman includes documentation of normal ovulatory cycles. A history of regular, cyclic, predictable, spontaneous menses usually indicates ovulatory cycles, which may be confirmed by basal body temperature graphs, properly timed endometrial biopsies, or plasma progesterone measurements during the luteal phase of the cycle. Also, the diagnosis of luteal-phase dysfunction (low progesterone secretion during the luteal phase) can be established by these methods. Transvaginal ultrasonography is useful for evaluating follicular development.

The most common cause of infertility in women is tubal disease, usually due to infection (pelvic inflammatory disease) or endometriosis. Tubal disease can be evaluated by obtaining a hysterosalpingogram or by diagnostic laparoscopy. Tubal diseases can usually be treated by laparoscopic tuboplasty and lysis of adhesions.

In many instances of infertility, it is now possible to use assisted reproductive technologies including in vitro fertilization and embryo transfer, gamete intrafallopian tube transfer, transfer of cryopreserved ova and embryos, donor oocytes or donor sperm, and ovarian hyperstimulation with clomiphene citrate or gonadotropins followed by intrauterine insemination.

The desire for contraception is also a frequent cause for women to seek medical treatment or evaluation. The most widely used methods for fertility control include (1) rhythm and withdrawal techniques, (2) barrier methods, (3) intrauterine devices, (4) oral steroid contraceptives, (5) sterilization, and (6) abortion.**These methods and their complications are discussed in Chap. 54.*

(Bibliography omitted in Palm version)

Back to Table of Contents

53. HIRSUTISM AND VIRILIZATION - *David A. Ehrmann*

Hirsutism, defined as excessive male-pattern hair growth, affects approximately 10% of women of reproductive age. Hirsutism may be mild, essentially representing a variation of normal hair growth, or rarely it may be the harbinger of a serious underlying condition. It is often idiopathic but may be caused by several conditions associated with androgen excess, such as polycystic ovarian syndrome (PCOS) or congenital adrenal hyperplasia (CAH) (Table 53-1). Cutaneous manifestations commonly associated with hirsutism include acne and male-pattern balding (androgenic alopecia). *Virilization*, on the other hand, refers to the state in which androgen levels are sufficiently high to cause additional signs and symptoms such as deepening of the voice, breast atrophy, increased muscle bulk, clitoromegaly, and increased libido; virilization is an ominous sign that suggests the possibility of an ovarian or adrenal neoplasm.

HAIR FOLLICLE GROWTH AND DIFFERENTIATION

Hair can be categorized as either *vellus* (fine, soft, and not pigmented) or *terminal* (long, coarse, and pigmented). The number of hair follicles does not change over an individual's lifetime, but the follicle size and type of hair can change in response to numerous factors, particularly androgens. Androgens are necessary for terminal hair and sebaceous gland development and mediate differentiation of pilosebaceous units (PSUs) into either a terminal hair follicle or a sebaceous gland. In the former case, androgens transform the vellus hair into a terminal hair; in the latter, the sebaceous component proliferates and the hair remains vellus.

There are three phases in the cycle of hair growth: (1) *anagen* (growth phase), (2) *catagen* (involution phase), and (3) *telogen* (rest phase). Depending on the body site, hormonal regulation may play an important role in the hair growth cycle. For example, the eyebrows, eyelashes, and vellus hairs are androgen-insensitive, whereas the axillary and pubic areas are sensitive to low doses of androgens. Hair growth on the face, chest, upper abdomen, and back requires greater levels of androgens and is therefore more characteristic of the pattern typically seen in males. Androgen excess in women leads to increased hair growth in most androgen-sensitive sites but will manifest with loss of hair in the scalp region, in part by reducing the time hairs spend in anagen phase.

Although androgen excess underlies most cases of hirsutism, there is only a modest correlation between androgen levels and the quantity of hair growth. This is due to the fact that hair growth from the follicle depends on local factors and variability in end-organ sensitivity, as well as circulating androgen concentrations. Genetic factors and ethnic background also influence hair growth. In general, dark-haired individuals tend to be more hirsute than blonde or fair individuals. Asians and Native Americans have relatively sparse hair in regions sensitive to high androgen levels, whereas people of Mediterranean descent are more hirsute. For these reasons, family history and ethnic background are important considerations when assessing the etiology and severity of hirsutism.

CLINICAL ASSESSMENT

Historic elements relevant to the assessment of hirsutism include the age of onset and rate of progression of hair growth and associated symptoms or signs (e.g., acne). Depending on the cause, excess hair growth is typically first noted during the second and third decades. The growth is usually slow but progressive. Sudden development and rapid progression of hirsutism suggests the possibility of an androgen-secreting neoplasm, in which case findings of virilization may also be present.

The age of onset of menstrual cycles (menarche) and the pattern of the menstrual cycle should be ascertained; irregular cycles from the time of menarche onward are more likely to result from ovarian rather than adrenal androgen excess. Associated symptoms such as galactorrhea should prompt evaluation for hyperprolactinemia (Chap. 328) and possibly hypothyroidism (Chap. 330). Hypertension, striae, easy bruising, centripetal weight gain, and weakness suggest hypercortisolism (Cushing's syndrome; Chap. 331). Rarely, patients with growth hormone excess (i.e., acromegaly) will present with hirsutism. Use of medications such as phenytoin, minoxidil, or cyclosporine may be associated with androgen-independent causes of excess hair growth (i.e., hypertrichosis). A family history of infertility and/or hirsutism may indicate disorders such as nonclassic congenital adrenal hyperplasia (CAH), a disorder particularly common in Ashkenazi Jews, among others (Chap. 331).

Physical examination should include measurement of height, weight, and calculation of body mass index (BMI). A BMI >25 kg/m^2 is indicative of excess weight for height, and values>30 kg/m^{2}are often seen in association with hirsutism. Notation should be made of blood pressure. Cutaneous signs sometimes associated with androgen excess and insulin resistance include acanthosis nigricans and skin tags.

An objective clinical assessment of hair distribution and quantity is central to the evaluation in any woman presenting with hirsutism. This assessment permits the distinction between hirsutism and hypertrichosis and provides a baseline reference point to gauge the response to treatment. *Hypertrichosis* refers to the excessive growth of androgen-independent hair which is vellus, prominent in nonsexual areas, and most commonly familial or caused by metabolic disorders (e.g., thyroid disturbances, anorexia nervosa) or medications (e.g., phenytoin, minoxidil or cyclosporine).

A simple and commonly used method to grade hair growth is the modified scale of Ferriman and Gallwey (Fig. 53-1), where each of nine androgen-sensitive sites is graded from 0 to 4. Approximately 95% of Caucasian women have a score below 8 on this scale; thus, it is normal for most women to have some hair growth in androgen-sensitive sites. Scores above 8 suggest an excess of androgen-mediated hair growth, a finding that should be assessed further by hormonal evaluation (see below). In racial/ethnic groups that are less likely to manifest hirsutism (e.g., Asian women), additional cutaneous evidence of androgen excess should be sought, including pustular acne or thinning hair.

HORMONAL EVALUATION

Androgens are secreted by both the ovaries and adrenal glands in response to their respective tropic hormones, luteinizing hormone (LH) and adrenocorticotropic hormone (ACTH). The principal circulating steroids involved in the etiology of hirsutism are androstenedione, dehydroepiandrosterone (DHEA) and its sulfated form (DHEAS), and testosterone. The ovaries and adrenal glands normally contribute about equally to testosterone production. Further, approximately half of the total testosterone originates from direct glandular secretion, and the remainder is derived from the peripheral conversion of androstenedione and DHEA (Chap. 335).

Although it is the most important circulating androgen, testosterone is, in effect, the penultimate androgen in mediating hirsutism; it is converted to the more potent dihydrotestosterone (DHT) by the enzyme 5a-reductase, which is located in the pilosebaceous unit. DHT has a higher affinity for, and slower dissociation from, the androgen receptor. The local production of DHT allows it to serve as the primary mediator of androgen action at the level of the pilosebaceous unit. There are two isoenzymes of 5a-reductase: type 2 is found in the prostate gland and in hair follicles, whereas type 1 is primarily found in sebaceous glands.

One approach to testing for hyperandrogenemia is depicted in Fig. 53-2. This involves measuring blood levels of testosterone andDHEAS. It is also important to measure the level of free (or unbound) testosterone, because it is the fraction of testosterone that is not bound to its carrier protein, sex-hormone binding globulin (SHBG), that is biologically available. Hyperinsulinemia and/or androgen excess decrease hepatic production of SHBG, often resulting in levels of total testosterone within the high-normal range at a time when the free hormone is substantially elevated. Because adrenal androgens are readily suppressed by low doses of glucocorticoids, the dexamethasone androgen-suppression test may broadly distinguish ovarian from adrenal androgen overproduction. A blood sample is obtained before and after administering dexamethasone (0.5 mg orally every 6 h for 4 days). An adrenal source is suggested by suppression of plasma free testosterone into the normal range; incomplete suppression suggests ovarian androgen excess.

A baseline plasma total testosterone level >12 nmol/L (>3.5 ng/mL) usually indicates a virilizing tumor, whereas a level>7 nmol/L (>2 ng/mL) is suggestive. A basalDHEASlevel >18.5 umol/L (>7000 ug/L) suggests an adrenal tumor. Although DHEAS has been proposed as a "marker" of predominant adrenal androgen excess, it is not unusual to find modest elevations in DHEAS among women withPCOS. Computed tomography (CT) or magnetic resonance imaging (MRI) should be used to localize an adrenal mass, and ultrasound will usually suffice to identify an ovarian mass, if clinical evaluation and hormonal levels suggest these possibilities.

PCOSis the most common cause of ovarian androgen excess (Chap. 336). However, the increased ratio ofLH to follicle-stimulating hormone that is often seen in carefully studied patients with PCOS may not be exhibited in up to half of these women due to the pulsatility of gonadotropins. If performed, ultrasound shows enlarged ovaries and/or increased stroma in many women with PCOS. However, polycystic ovaries may also be found in women without clinical or laboratory features of PCOS. Therefore, polycystic ovaries are a relatively insensitive and nonspecific finding for the diagnosis of ovarian hyperandrogenism. Though it is not widely used, gonadotropin-releasing hormone agonist testing can be used to make a specific diagnosis of ovarian hyperandrogenism. A peak 17-hydroxyprogesterone level ³7.8 nmol/L (³2.6 ug/L), after the administration of 100 ug nafarelin (or 10 ug/kg leuprolide) subcutaneously, is virtually diagnostic of ovarian hyperandrogenism.

NonclassicCAH is most commonly due to 21-hydroxylase deficiency but can also be caused by autosomal recessive defects in other steroidogenic enzymes necessary for adrenal corticosteroid synthesis (Chap. 331). Because of the enzyme defect, the adrenal gland cannot secrete glucocorticoids efficiently (especially cortisol). This results in diminished negative feedback inhibition ofACTH, leading to compensatory hyperplasia of the adrenal cortex and accumulation of steroid precursors proximal to the enzyme defect. These precursors are subsequently converted to androgen.

Deficiency of 21-hydroxylase can be reliably excluded by determining a morning 17-hydroxyprogesterone level <6 nmol/L (<2 ug/L) (drawn in the follicular phase). Alternatively, 21-hydroxylase deficiency can be diagnosed by measurement of 17-hydroxyprogesterone 1 h after administration of 250 ug of syntheticACTH(cosyntropin) intravenously. Measurement after ACTH is slightly more cumbersome, though the results obtained in this manner are highly reproducible and can be compared to published nomograms.

TREATMENT

Treatment of hirsutism may be accomplished pharmacologically and by mechanical means of hair removal. Nonpharmacologic treatments should be considered in all patients, either as the only treatment or as an adjunct to drug therapy.

Nonpharmacologic treatments include (1) bleaching; (2) depilatory (removal from the skin surface) such as shaving and chemical treatments; or (3) epilatory (removal of the hair including the root) such as plucking, waxing, electrolysis, and laser therapy. Despite perceptions to the contrary, shaving does not increase the rate or density of hair growth. Chemical depilatory treatments may be useful for mild hirsutism that affects only limited skin areas, though they can cause skin irritation. Wax treatment removes hair temporarily but is uncomfortable. Electrolysis is effective for more permanent hair removal, particularly in the hands of a skilled electrologist. Laser phototherapy appears to be efficacious for hair removal. It delays hair regrowth and causes permanent hair removal in some patients. The long-term effects and complications associated with laser treatment are being evaluated.

Pharmacologic therapy for androgen excess is directed at interrupting one or more of the steps in the pathway leading to its expression: (1) suppression of adrenal and/or ovarian androgen production; (2) enhancement of androgen-binding to plasma-binding proteins, particularlySHBG; (3) impairment of the peripheral conversion of androgen precursors to active androgen; and (4) inhibition of androgen action at the target tissue level. Attenuation of hair growth is typically not evident until 4 to 6 months after initiation of medical treatment and, in most cases, leads to a modest reduction in hair growth.

Combination estrogen-progestin therapy, in the form of an oral contraceptive, is usually the first-line endocrine treatment for hirsutism and acne, after cosmetic and dermatologic management. The estrogenic component of most oral contraceptives currently in use is either ethinyl estradiol or mestranol. The suppression ofLH leads to reduced production of ovarian androgens. The reduced androgen levels also result in a dose-related increase inSHBG, thereby lowering the fraction of unbound plasma testosterone. Combination therapy has also been demonstrated to decreaseDHEAS, perhaps by reducingACTHlevels. Estrogens also have a direct, dose-dependent suppressive effect on sebaceous cell function.

The choice of a specific oral contraceptive should be predicated on the progestational component, as progestins vary in their suppressive effect onSHBGlevels and in their androgenic potential. Ethynodiol diacetate has relatively low androgenic potential, whereas progestins such as norgestrel and levonorgestrel are particularly androgenic, as judged from their attenuation of the estrogen-induced increase in SHBG. Norgestimate exemplifies the newer generation of progestins that are virtually nonandrogenic. Oral contraceptives are contraindicated in women with a history of thromboembolic disease or in women with breast cancer or other estrogen-dependent cancers (Chap. 336). There is a relative contraindication to the use of oral contraceptives in smokers or in those with hypertension or a history of migraine headaches. In most trials, estrogen-progestin therapy alone improves the extent of acne by a maximum of 50 to 70%. In contrast, the effect on hair growth may not be evident for 6 months, and the maximum effect may require 9 to 12 months owing to the length of the hair growth cycle. Improvements in hirsutism are typically in the range of 20%, and often there is little more than arrest of further progression of hair growth.

Adrenal androgens are more sensitive than cortisol to the suppressive effects of glucocorticoids. Therefore, glucocorticoids are the mainstay of treatment in patients withCAH. Although glucocorticoids have been reported to restore ovulatory function in some women withPCOS, this effect is highly variable. Because of side effects from excessive glucocorticoids, low doses should be used. Dexamethasone (0.2 to 0.5 mg) or prednisone (5 to 10 mg) should be given at bedtime to achieve maximal suppression by inhibiting the nocturnal surge ofACTH.

Cyproterone acetate is the prototypic antiandrogen. It acts mainly by competitive inhibition of the binding of testosterone andDHT to the androgen receptor. In addition, it may act to enhance the metabolic clearance of testosterone by inducing hepatic enzymes. Although not available for use in the United States, cyproterone acetate is widely used in Canada, Mexico, and Europe. Cyproterone (50 to 100 mg) is given on days 1 to 15 and ethinyl estradiol (50 ug) is given on days 5 to 26 of the menstrual cycle. Side effects of cyproterone acetate include irregular uterine bleeding, nausea, headache, fatigue, weight gain, and decreased libido.

Spironolactone, usually used as a mineralocorticoid antagonist, is also a weak antiandrogen. It is almost as effective as cyproterone acetate when used at high enough doses (100 to 200 mg daily). Patients should be monitored intermittently for hyperkalemia or hypotension, though these side effects are uncommon. Pregnancy should be avoided because of the risk of feminization of a male fetus. Spironolactone can also cause menstrual irregularity. It is often used in combination with an oral contraceptive, which helps in prevention of pregnancy and suppression of ovarian androgen production.

Flutamide is a potent nonsteroidal antiandrogen that is effective in treating hirsutism, but concerns about the induction of hepatocellular dysfunction have limited its use.

Finasteride is a competitive inhibitor of 5a-reductase type 2. Beneficial effects on hirsutism have been reported, but the prominence of 5a-reductase type 1 in the pilosebaceous unit appears to account for its limited efficacy. Finasteride would also be expected to impair sexual differentiation in a male fetus, and thus it should not be used in women who may become pregnant.

A prospective, randomized trial comparing low-dose flutamide, finasteride, and combination cyproterone acetate-ethinyl estradiol demonstrated relative superiority of flutamide and cyproterone acetate-ethinyl estradiol in the treatment of hirsutism. Ultimately, the choice of any specific agent(s) must be tailored to the unique needs of the patient being treated. As noted previously, pharmacologic treatments for hirsutism should be used in conjunction with nonpharmacologic approaches. Patients should be reminded about the relatively slow and usually modest responses to pharmacologic treatment. It is also helpful to review the pattern of female hair distribution in the normal population to dispel unrealistic expectations.

(Bibliography omitted in Palm version)

Back to Table of Contents

54. INFERTILITY AND FERTILITY CONTROL - *Janet E. Hall*

The concept of reproductive choice is now firmly entrenched in developed countries and has dramatically altered reproductive behavior. The availability of effective contraceptive methods prevents unintended pregnancies and gives women the option of pursuing educational and career opportunities without interruption. Population control also has important economic and social implications. Infertility, on the other hand, can be accompanied by substantial stress and disappointment. Fortunately, the ability to diagnose and to treat various causes of infertility now provides an array of effective new approaches to this condition.

INFERTILITY

DEFINITION AND PREVALENCE

Infertility is defined as the inability to conceive after 12 months of unprotected sexual intercourse. In a study of 5574 English and American women who ultimately conceived, pregnancy occurred in 50% within 3 months, 72% within 6 months, and 85% within 12 months. These findings are consistent with predictions based on *fecundability*, the probability of achieving pregnancy in one menstrual cycle (approximately 20 to 25% in healthy young couples). Assuming a fecundability of 0.25, 98% of couples should conceive within 13 months. Based on this definition, the National Survey of Family Growth reports a 14% rate of infertility in the United States in married women aged 15 to 44. The infertility rate has remained relatively stable over the past 30 years, although the proportion of couples without children has risen, reflecting a trend to delay childbearing. This trend has important implications because of an age-related decrease in fecundability, which begins at age 35, and decreases markedly after age 40.

CAUSES OF INFERTILITY

There is a spectrum of infertility, ranging from reduced conception rates or the need for medical intervention to irreversible causes of infertility (*sterility*). Infertility can be attributed primarily to male factors in 25%, female factors in 58%, and is unexplained in about 17% of couples (Fig. 54-1). Not uncommonly, both male and female factors contribute to infertility.

Approach to the Patient

Initial Evaluation In all couples presenting with infertility, the initial evaluation includes discussion of the appropriate timing of intercourse and a description of the range of investigations that may be required. A brief description of infertility treatment options, including adoption, should be reviewed. Initial investigations are focused on determining whether the primary cause of the infertility is male, female, or both. These investigations include a semen analysis in the male, confirmation of ovulation in the female, and, in the majority of situations, documentation of tubal patency in the female. Although frequently used in the past, recent studies have not supported the efficacy of postcoital testing of sperm interaction with cervical mucus as a routine component of initial testing. Strategies for further evaluation are described below and in Chaps. 335 and336. In some cases, after an extensive workup excluding all male and female factors, a specific cause cannot be identified and infertility may ultimately be classified as unexplained.

Psychological Aspects of Infertility Infertility is invariably associated with psychological stress related not only to the diagnostic and therapeutic procedures themselves but also to repeated cycles of hope and loss associated with each new procedure or cycle of treatment that does not result in the birth of a child. These feelings are often combined with a sense of isolation from friends and family. Counseling and stress-management techniques should be introduced early in the evaluation of infertility. In addition to the psychological benefits of stress management, it is possible that stress contributes to infertility in some couples (e.g., impaired ovulation). Importantly, infertility and its treatment do not appear to be associated with long-term psychological sequelae.

Female Causes Abnormalities in menstrual function constitute the most common cause of female infertility. These disorders, which include ovulatory dysfunction and abnormalities of the uterus or outflow tract, may present as amenorrhea (absence of menses) or as irregular or short menstrual cycles. A careful history and physical examination and a limited number of laboratory tests will help to determine whether the abnormality is: (1) hypothalamic or pituitary [low follicle-stimulating hormone (FSH), luteinizing hormone (LH), and estradiol with or without an increase in prolactin]; (2) polycystic ovarian syndrome (PCOS; irregular cycles and hyperandrogenism in the absence of other causes of androgen excess); (3) ovarian (low estradiol with increased FSH); or (4) uterine or outflow tract abnormality. The frequency of these diagnoses depends on whether the amenorrhea is primary or occurs after normal puberty and menarche (Fig. 54-1).**The approach to further evaluation of these disorders is described in detail in Chap. 52.*

OVULATORY DYSFUNCTION In women with a history of regular menstrual cycles,*evidence of ovulation* should be sought by using urinary ovulation predictor kits (they reflect the preovulatory gonadotropin surge but do not confirm ovulation), basal body temperature charts, or a mid-luteal phase progesterone level. The mid-luteal phase progesterone increase (usually>3 ng/mL) confirms ovulation and corpus luteum function and is responsible for the rise in basal body temperature [>0.3°C (>0.6°F) for 10 days]. An endometrial biopsy to exclude luteal phase insufficiency is no longer considered an essential part of the infertility workup for most patients. Even in the presence of ovulatory cycles, evaluation of *ovarian reserve* is recommended for women over 35 by measurement ofFSH on day 3 of the cycle or in response to clomiphene, an estrogen antagonist (see below). An FSH level <10 IU/mL on cycle day 3 predicts adequate ovarian oocyte reserve. Inhibin B, an ovarian hormone that selectively suppresses FSH, is being investigated as an additional marker of ovarian reserve.

TUBAL DISEASE This may result from pelvic inflammatory disease (PID), appendicitis, endometriosis, pelvic adhesions, tubal surgery, and previous use of an intrauterine device (IUD). However, a cause is not identified in up to 50% of patients with documented tubal factor infertility. Because of the high prevalence of tubal disease, testing should occur early in the majority of couples with infertility. Subclinical infections with *Chlamydia trachomatis* may be an underdiagnosed cause of tubal infertility and requires the treatment of both partners. A hysterosalpingogram (HSG) is the most common screening test and will determine the presence of tubal patency and identify potential abnormalities of the uterine cavity.

ENDOMETRIOSIS Endometriosis is defined as the presence of endometrial glands or stroma outside the endometrial cavity and uterine musculature. Its presence is suggested by a history of dyspareunia (painful intercourse), worsening dysmenorrhea that often begins before menses, or by a thickened rectovaginal septum or deviation of the cervix on pelvic examination. The pathogenesis of the infertility associated with endometriosis is unclear but may involve indirect effects on the normal endometrium as well as the direct effects of adhesions in advanced disease. Endometriosis is often clinically silent, however, and can only be excluded definitively by laparoscopy.

Male Causes Known causes of male infertility include primary testicular disease, disorders of sperm transport, and hypothalamic-pituitary disease resulting in secondary hypogonadism. However, the etiology is not ascertained in up to half of men with suspected male factor infertility (Fig. 54-1). The key initial diagnostic test is a *semen analysis*. Although 95% confidence limits can be used to define normal semen parameters, data relating sperm counts to fecundability are more useful. Such studies suggest that sperm counts of<20 million/mL, with a motility of less than 40%, are associated with an increased risk of infertility. Analysis of sperm morphology is less well validated, but >40% normal forms are usually present in fertile men. Successful in vitro fertilization (IVF) can usually be accomplished with >14% normal forms (using strict Kruger criteria), whereas low fertilization is seen with<4% normal forms. Other tests such as the hamster egg penetration test and the zona-binding assay are not of proven value.

Testosterone levels should be measured if the sperm count is low on repeated examination or if there is clinical evidence of hypogonadism. A low testosterone level may result from *primary gonadal deficiency*; in this condition, levels ofLH andFSH will be elevated. Less commonly, low testosterone and decreased spermatogenesis result from hypothalamic or pituitary disease, in which case the LH and FSH levels will be low (Chap. 335).

Abnormalities of spermatogenesis may have a genetic component. Y chromosome microdeletions and substitutions are increasingly recognized as a cause of *azoospermia* (absence of sperm) or *oligospermia* (low sperm count). Microdeletions (Yq6 region) have also been identified in a subset of men with elevatedFSHlevels or otherwise idiopathic infertility. Several candidate genes have been identified including *DAZ* (deleted in azoospermia) and *YRRM* (Y chromosome RNA recognition motif).

Acquired disorders of the testes are often associated with impaired spermatogenesis with relatively preserved Leydig cell function; thus, testosterone levels may be normal. Such abnormalities include viral orchitis (especially mumps) and other infectious causes such as tuberculosis or sexually transmitted diseases (STDs), chemotherapy (especially the alkylating agents cyclophosphamide and chlorambucil), ionizing radiation, and drugs that may impair fertility directly or through inhibition of testicular androgen production or action. Anabolic androgen abuse should be considered in a well-androgenized man with low gonadotropins and testosterone but a suppressed sperm count. Prolonged elevation of testicular temperature may impair spermatogenesis, e.g., after an acute febrile illness or in association with varicocele. A potential role for environmental toxins as a cause of impaired spermatogenesis has been suggested based on an apparent decrease in sperm counts over the past several decades, but a direct cause-and-effect relationship has not been established.

SECONDARY HYPOGONADISM Low gonadotropin levels, associated with low testosterone, may signal the presence of a pituitary macroadenoma or hypothalamic tumor (in both cases prolactin levels may be elevated; Chap. 328) or may be the first presentation of hemochromatosis (Chap. 345) or other systemic illness. Recent studies have identified several genetic causes of gonadotropin-releasing hormone (GnRH) deficiency (*KAL* and *DAX-1*), as well as mutations that lead to isolated gonadotropin deficiency (GnRH receptor,LHb,FSHbmutations) (Chap. 328).

DISORDERED SPERM TRANSPORT Patients with low sperm counts and normal hormonal levels may be found to have obstructive abnormalities of the vas deferens or epididymus. The most common causes of vas deferens obstruction are previous vasectomy or accidental ligation during inguinal surgery. Patency rates with microsurgical reversal techniques are high in the first 3 years after vasectomy but decrease markedly thereafter. Congenital absence of the vas deferens can be diagnosed by a deficiency of fructose in the ejaculate and is often associated with an abnormality of the cystic fibrosis transmembrane regulator (*CFTR*) gene. Young's syndrome, characterized by inspissated secretions, can also preclude normal sperm transport.

TREATMENT

The treatment of infertility should be tailored to the problems unique to each couple (Table 54-1). In many situations, including unexplained infertility, mild to moderate endometriosis, and/or borderline semen parameters, a stepwise approach to infertility is optimal, beginning with low-risk interventions and moving to more invasive, higher risk interventions only if necessary. After determination of all infertility factors and their correction, if possible, this approach might include, in increasing order of complexity: (1) expectant management, (2) clomiphene citrate (see below) with or without intrauterine insemination (IUI), (3) gonadotropins with or without IUI, and (4)IVF. The time used to complete the evaluation, correction, and expectant management can be longer in women <30, but this process should be advanced rapidly in women >35. In some situations expectant management will not be appropriate.

Ovulatory Dysfunction Treatment of ovulatory dysfunction should first be directed at identification of the etiology of the disorder to allow specific management when possible. Dopamine agonists, for example, may be indicated in patients with hyperprolactinemia (Chap. 328); lifestyle modification may be successful in women with low body weight or a history of intensive exercise (Chap. 78).

*PulsatileGnRH*is highly effective for restoring ovulation in patients with hypothalamic amenorrhea. When administered subcutaneously by an automated pump at a physiologic dose and frequency, pulsatile GnRH induces normalLH andFSHdynamics. Direct comparisons between pulsatile GnRH and gonadotropin treatment for ovulation induction indicate similar pregnancy rates; pulsatile GnRH is associated with lower rates of multiple gestation and virtually no risk of ovarian hyperstimulation.

Clomiphene citrate is a nonsteroidal estrogen antagonist that increasesFSH andLHlevels by blocking estrogen negative feedback at the hypothalamus. The efficacy of clomiphene for ovulation induction is highly dependent on patient selection. It induces ovulation in ~60% of women withPCOS and is the initial treatment of choice in these patients. The starting dose is 50 mg daily for 5 days beginning on day 5 of a spontaneous cycle or after a progestin-induced withdrawal bleed. The dose can be increased to 150 mg, if necessary, in subsequent cycles, and human chorionic gonadotropin (hCG) can be added as the ovulatory stimulus. In women with PCOS, the use of insulin-sensitizing agents, such as metformin appears to be particularly effective in combination with clomiphene.

Gonadotropins are highly effective for ovulation induction in women with hypogonadotropic hypogonadism andPCOS. Gonadotropins are also used to induce multiple follicular recruitment in unexplained infertility and in older reproductive-aged women, particularly in conjunction withIUI. Disadvantages include a significant risk of multiple gestation and the risk of ovarian hyperstimulation, a side effect that is more common in women with PCOS. However, careful monitoring and a conservative approach to ovarian stimulation reduce these risks; gonadotropin stimulation is an effective and safe treatment when applied by experienced practitioners. Currently available gonadotropins include urinary preparations ofLH andFSH, highly purified FSH, and recombinant FSH. Though FSH is the key component, there is growing data that the addition of some LH (orhCG) may improve results, particularly in hypogonadotropic patients.

None of these methods are effective in women with premature ovarian failure in whom donor oocyte or adoption are the methods of choice.

Tubal Disease If hysterosalpingography suggests a tubal or uterine cavity abnormality, or if a patient is ³35 at the time of initial evaluation, laparoscopy with tubal lavage is recommended, often with a hysteroscopy. Although tubal reconstruction may be attempted if tubal disease is identified, it is generally being replaced by the use ofIVF, as these patients are at increased risk of developing an ectopic pregnancy.

Endometriosis Though 60% of women with minimal or mild endometriosis may conceive within 1 year without treatment, laparoscopic resection or ablation appear to improve conception rates. Medical management of advanced stages of endometriosis is widely used for symptom control but has not been shown to enhance fertility (Chap. 336). In moderate to severe endometriosis, conservative surgery is associated with pregnancy rates of 50 and 39% respectively, compared with rates of 25 and 5% with expectant management alone. In some patients,IVF may be the treatment of choice.

Male Factor Infertility The treatment options for male factor infertility have expanded greatly in recent years. Secondary hypogonadism is highly amenable to treatment with pulsatileGnRH or gonadotropins (Chap. 335). In vitro techniques have provided new opportunities for patients with primary testicular failure and disorders of sperm transport. Choice of initial treatment options depends on sperm concentration and motility. Expectant management should be attempted initially in men with mild male factor infertility (sperm count of 15 to 20´10$_6$/mL and normal motility). Moderate male factor infertility (10 to 15´10$_6$/mL and 20 to 40% motility) should begin withIUIalone or in combination with treatment of the female partner with clomiphene or gonadotropins, but it may requireIVF with or without intracytoplasmic sperm injection (ICSI). For men with a severe defect (sperm count of <10´10$_6$/mL, 10% motility), IVF with ICSI or donor sperm should be used.

Assisted Reproductive Technologies The development of assisted reproductive technologies (ART) has dramatically altered the treatment of male and female infertility.IVF is indicated for patients with many causes of infertility that have not been successfully managed with more conservative approaches. IVF orICSI is often the treatment of choice in couples with a significant male factor or tubal disease, whereas IVF using donor oocytes is used in patients with premature ovarian failure and in women of advanced reproductive age. Success rates depend on the age of the woman and the cause of the infertility and are generally 18 to 24% per cycle when initiated in women <40. In women>40, there is a marked decrease in both the number of oocytes retrieved and their ability to be fertilized. Though often effective, IVF is expensive and requires careful monitoring of ovulation induction and invasive techniques including the aspiration of multiple follicles. IVF is associated with a significant risk of multiple gestation (29% twins, 7% triplets, and 0.6% higher order multiples). More recently developed blastocyst transfer protocols decrease the number of transfers but increase pregnancy rates.

CONTRACEPTION

Though various forms of contraception are widely available, approximately 30% of births in the United States are the result of unintended pregnancy. Teenage pregnancies continue to represent a serious public health problem in the United States, with>1 million unintended pregnancies each year -- a significantly greater incidence than in other industrialized nations (Chap. 8).

Contraceptive methods are widely used (Table 54-2). Only 15% of couples report having unprotected sexual intercourse in the past 3 months. A reversible form of contraception is used by >50% of couples. Sterilization (in either the male or female) has been employed as a permanent form of contraception by about 25% of couples. Pregnancy termination is relatively safe when directed by health care professionals but is rarely the option of choice.

No single contraceptive method is ideal, although all are safer than carrying a pregnancy to term. The effectiveness of a given method of contraception is dependent on the efficacy of the method itself, compliance, and appropriate use. Knowledge of the advantages and disadvantages of each contraceptive is essential for counseling an individual about the methods that are safest and most consistent with his or her lifestyle. Discrepancies between theoretical and actual effectiveness emphasize the importance of patient education and compliance when considering various forms of contraception (Table 54-2).

BARRIER METHODS

Barrier contraceptives, such as condoms, diaphragms, cervical caps, and spermicides, are easily available, reversible, and have fewer side effects than hormonal methods.

However, their effectiveness is highly dependent on compliance and proper use (Table 54-2). A major advantage of barrier contraceptives is the protection provided againstSTDs(Chap. 132). Consistent use is associated with a decreased risk of gonorrhea, nongonococcal urethritis, and genital herpes, probably due in part to the concomitant use of spermicides. Condom use also reduces the transmission of HIV infection. Natural membrane condoms may be less effective than latex condoms, and petroleum-based lubricants can degrade condoms and decrease their efficacy for preventing HIV infection. A highly effective female condom, which also provides protection against STDs, was approved in 1994 but has not achieved widespread use.

STERILIZATION

Sterilization is the method of birth control most frequently chosen by fertile men and multiparous women >30 (Table 54-2). Sterilization refers to a procedure that prevents fertilization by surgical interruption of the fallopian tubes in women or the vas deferens in men. Although tubal ligation and vasectomy are potentially reversible, these procedures should be considered permanent and should not be undertaken without careful patient counseling.

Several methods of *tubal ligation* have been developed, all of which are highly effective with a 10-year cumulative pregnancy rate of 1.85 per 100 women. However, when pregnancy does occur, the risk of ectopic pregnancy may be as high as 30%. The success rate of tubal reanastomosis depends on the method used -- the clip, silastic band, and modified Pomeroy procedures are easier to reverse than the Irving, Uchida, and electrocoagulation methods. Even after successful reversal, the risk of ectopic pregnancy remains great. In addition to prevention of pregnancy, tubal ligation reduces the risk of ovarian cancer, possibly by limiting the upward migration of potential carcinogens.

Vasectomy is an outpatient surgical procedure that has little risk and is highly effective. The development of azoospermia may be delayed for 2 to 6 months, and other forms of contraception must be used until two sperm-free ejaculations provide proof of sterility. Reanastomosis may restore fertility in 30 to 50% of men, but the success rate appears to decline with time after vasectomy and may be influenced by nonmechanical factors such as the development of anti-sperm antibodies.

INTRAUTERINE DEVICES

IUDsinhibit pregnancy primarily through a spermicidal effect caused by a sterile inflammatory reaction produced by the presence of a foreign body in the uterine cavity. There may also be effects on cervical mucus sperm transport through the oviduct. IUDs provide a high level of efficacy in the absence of systemic metabolic effects. An additional advantage is that ongoing motivation is not required to ensure efficacy once the device has been placed. However, only 1% of women in the United States use this method compared to a utilization rate of 15 to 30% in much of Europe and Canada. This relatively low utilization rate continues despite evidence that the newer devices are not associated with increased rates of pelvic infection and infertility, as occurred with earlier devices. Screening forSTDshould be performed prior to insertion, and an IUD should not be used in women at high risk for development of STD or in women at high risk for bacterial endocarditis. In addition, the IUD may not be effective in women with uterine leiomyomas because they alter the size or shape of the uterine cavity. IUD use is associated with increased menstrual blood flow, although this is less pronounced with the progesterone-releasing IUD than the copper-containing device.

HORMONAL METHODS

No male hormonal contraceptive methods are currently approved in the United States. However, hormonal methods of male contraception, includingGnRH-mediated suppression of the hypothalamic-pituitary-gonadal axis in combination with testosterone replacement, are under investigation.

Oral Contraceptive Pills Because of their ease of use and efficacy, oral contraceptive pills are the most widely used form of hormonal contraception. They act by suppressing ovulation, changing cervical mucus, and altering the endometrium. The current formulations are made from synthetic estrogens and progestins. The estrogen component of the pill consists of ethinyl estradiol or mestranol, which is metabolized to ethinyl estradiol. Multiple synthetic progestins are used. Norethindrone and its derivatives are used in many formulations. Low-dose norgestimate and third-generation progestins (desogestrel, gestodene) have a less androgenic profile; levonorgestrel appears to be the most androgenic of the progestins and should be avoided in patients with hyperandrogenic symptoms. The three major formulations of oral contraceptives include: (1) fixed-dose estrogen-progestin combination, (2) phasic estrogen-progestin combination, and (3) progestin only. Each of these formulations is administered daily for 3 weeks followed by a week of no medication during which menstrual bleeding generally occurs.

Current doses of ethinyl estradiol range from 20 to 50 ug. However, indications for the 50-ug dose are rare, and the majority of formulations contain 35 ug of ethinyl estradiol. The reduced estrogen and progesterone content in the second- and third-generation pills has decreased both side effects and risks associated with oral contraceptive use. At the currently used doses, patients must be cautioned not to miss pills due to the potential for ovulation. Side effects, including break-through bleeding, amenorrhea, and weight gain, are often responsive to a change in formulation. There is no evidence that low-dose oral contraceptives increase the risk of cardiovascular disease in women <30 or in nonsmoking women without additional risk factors. However, the risk of myocardial infarction and stroke in women who smoke is increased by the use of oral contraceptives. The risk of developing hypertension is increased somewhat, even with the low-dose preparations. An increased risk of venous thromboembolism occurs with all oral contraceptives and may be even greater with the third-generation preparations. The factor V Leiden mutation and other thrombophilic disorders (Chap. 117) are important risk factors for venous thrombosis during oral contraceptive therapy. However, biochemical or genetic screening for these disorders before starting oral contraceptives is not cost-effective at present. In most studies, oral contraceptive use has not been shown to increase the risk of breast cancer, but there is a slight increase in the risk of cervical cancer. Risks for endometrial and ovarian cancer are decreased in oral contraceptive users.

Previous thromboembolic events or stroke are absolute contraindications for the use of oral contraceptive pills. A history of hormone-dependent tumors and liver disease are also contraindications. Oral contraceptive pills should not be given in pregnancy or in women with undiagnosed uterine bleeding or amenorrhea.

The microdose progestin-only minipill is less effective as a contraceptive, having a pregnancy rate of 2 to 7 per 100 women-years. However, it may be appropriate for women with cardiovascular disease or for women who cannot tolerate synthetic estrogens.

Injectable Contraceptives Depot medroxyprogesterone acetate (Depo-Provera) and Norplant (Table 54-2) act primarily by inhibiting ovulation and causing changes in the endometrium and cervical mucus that result in decreased implantation and sperm transport. Depo-Provera is effective for 3 months, but return of fertility after discontinuation may be delayed for up to 12 to 18 months. Norplant requires surgical insertion but is effective for up to 5 years afer insertion; fertility is possible shortly after its removal. The U.S. Food and Drug Administration (FDA) has recently approved the use of covered rods in addition to the capsules. Amenorrhea, irregular bleeding, and weight gain are the most common adverse effects associated with both injectable forms of contraception. An injectable progestin/estrogen combination contraceptive will be available soon. It requires monthly injection, but irregular bleeding and weight gain are less common. A major advantage of the injectable progestin-based contraceptives is the apparent lack of increased arterial and venous thromboembolic events.

POSTCOITAL CONTRACEPTION

Postcoital contraceptive methods prevent implantation or cause regression of the corpus luteum and are highly efficacious if used appropriately. Although postcoital contraception is not specifically licensed for use in the United States, an FDA notice published in 1997 indicated that certain oral contraceptive pills could be used within 72 h of unprotected intercourse [Ovral (2 tablets 12 h apart) and Lo/Ovral (4 tablets 12 h apart)]. The Preven Emergency Contraceptive Kit contains four combination tablets (50 mg ethinyl estradiol and 0.25 mg levonorgestrel) and a pregnancy kit to rule out pregnancy before taking the pills. Side effects are common with these high doses of hormones and include nausea, vomiting, and breast soreness. Recent studies suggest that 600 mg mifepristone (RU486), a progesterone receptor antagonist, may be equally as effective or more effective than hormonal regimens, with fewer side effects. Mifepristone is not currently available in the United States.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 9 -ALTERATIONS IN THE SKIN

55. APPROACH TO THE PATIENT WITH A SKIN DISORDER - *Thomas J. Lawley*, *Kim B. Yancey*

The challenge of examining the skin lies in distinguishing normal from abnormal, significant findings from trivial ones, and in integrating pertinent signs and symptoms into an appropriate differential diagnosis. The fact that the largest organ in the body is visible is both an advantage and a disadvantage to those who examine it. It is advantageous because no special instrumentation, other than a magnifying glass, is necessary and because the skin can be biopsied with little morbidity. However, the casual observer can be overwhelmed by a variety of stimuli and overlook important, subtle signs of skin or systemic disease. For instance, the sometimes minor differences in color and shape that distinguish a malignant melanoma (see Plate IIC-30) from a benign pigmented nevus (seePlate IIC-28) can be difficult to recognize. To aid in the interpretation of skin lesions, a variety of descriptive terms have been developed to characterize cutaneous lesions (Tables 55-1 and55-2 andFig. 55-1) and to formulate a differential diagnosis (Table 55-3). For instance, the finding of large numbers of scaling papules, usually indicative of a primary skin disease, places the patient in a different diagnostic category than would hemorrhagic papules, which may indicate vasculitis or sepsis (seePlates IIE-71 andIID-44, respectively). It is important to differentiate primary skin lesions from secondary skin changes. If the examiner focuses on linear erosions overlying an area of erythema and scaling, he or she may incorrectly assume that the erosion is the primary lesion and the redness and scale are secondary, while the correct interpretation would be that the patient has a pruritic eczematous dermatitis and the erosions have been caused by scratching.

Approach to the Patient

In examining the skin it is usually advisable to assess the patient before taking a history. This way, the entire cutaneous surface is sure to be evaluated, and objective findings can be integrated with relevant historic data. Four basic features of any cutaneous lesion must be noted and considered in the examination of skin: the distribution of the eruption, the type(s) of primary lesion, the shape of individual lesions, and the arrangement of the lesions. In the initial examination it is important that the patient be disrobed as completely as possible. This will minimize chances of missing important individual skin lesions and make it possible to assess the distribution of the eruption accurately. The patient should first be viewed from a distance of about 1.5 to 2 m (4 to 6 ft) so that the general character of the skin and the distribution of lesions can be evaluated. Indeed, distribution of lesions often correlates highly with diagnosis (Fig. 55-2). For example, a hospitalized patient with a generalized erythematous exanthem is more likely to have a drug eruption than is a patient with a similar rash limited to the sun-exposed portions of the face. The presence or absence of lesions on mucosal surfaces should also be determined. Once the distribution of the lesions has been established, the nature of the primary lesion must be determined. Thus, when lesions are distributed on elbows, knees, and scalp, the most likely possibility based solely on distribution is psoriasis or dermatitis herpetiformis (seePlates IIA-3 andIIE-68, respectively). The primary lesion in psoriasis is a scaly papule that soon forms erythematous plaques covered with a white scale, whereas that of dermatitis herpetiformis is an urticarial papule that quickly becomes a small vesicle. In this manner, identification of the primary lesion directs the examiner toward the proper diagnosis. Secondary changes in skin can also be quite helpful. For example, scale represents excessive epidermis, while crust is the result of an inadequate or discontinuous epithelial cell layer. Palpation of skin lesions can also yield insight into the character of an eruption. Thus red papules on the lower extremities that blanch with pressure can be a manifestation of many different diseases, but hemorrhagic red papules that do not blanch with pressure indicate palpable purpura characteristic of necrotizing vasculitis (seePlate IIE-71).

The shape of lesions is also an important feature. Flat, round, erythematous papules and plaques are common in many cutaneous diseases. However, target-shaped lesions that consist in part of erythematous plaques are specific for erythema multiforme (see Plate IIE-67). In the same way, the arrangement of individual lesions is important. Erythematous papules and vesicles can occur in many conditions, but their arrangement in a specific linear array suggests an external etiology such as allergic contact (see Plate IIA-8) or primary irritant dermatitis. In contrast, lesions with a generalized arrangement are common and suggest a systemic etiology.

As in other branches of medicine, a complete history should be obtained to emphasize the following features:

1. Evolution of lesions

a. Site of onset

b. Manner in which eruption progressed or spread

c. Duration

d. Periods of resolution or improvement in chronic eruptions

2. Symptoms associated with the eruption

a. Itching, burning, pain, numbness

b. What, if anything, has relieved symptoms

c. Time of day when symptoms are most severe

3. Current or recent medications (prescribed as well as over-the-counter)

4. Associated systemic symptoms (e.g., malaise, fever, arthralgias)

5. Ongoing or previous illnesses

6. History of allergies

7. Presence of photosensitivity

8. Review of systems

DIAGNOSTIC TECHNIQUES

Many skin diseases can be diagnosed on gross clinical appearance, but sometimes relatively simple diagnostic procedures can yield valuable information. In most instances, they can be performed at the bedside with a minimum of equipment.

Skin Biopsy A skin biopsy is a straightforward minor surgical procedure; however, it is important to biopsy the anatomic site most likely to yield diagnostic findings. This decision may require expertise in skin diseases and knowledge of superficial anatomic structures in selected areas of the body. In this procedure, a small area of skin is anesthetized with 1% lidocaine with or without epinephrine. The skin lesion in question can be excised with a scalpel or removed by punch biopsy. In the latter technique, a punch is pressed against the surface of the skin and rotated with downward pressure until it penetrates to the subcutaneous tissue. The circular biopsy is then lifted with forceps, and the bottom is cut with iris scissors. Biopsy sites may or may not need suture closure, depending on size and location.

KOH Preparation A potassium hydroxide (KOH) preparation is performed on scaling skin lesions when a fungal etiology is suspected. The edge of such a lesion is scraped gently with a scalpel blade, and the removed scale is collected on a glass microscope slide and treated with 1 to 2 drops of a solution of 10 to 20% KOH. KOH dissolves keratin and allows easier visualization of fungal elements. Brief heating of the slide accelerates dissolution of keratin. When the preparation is viewed under the microscope, the refractile hyphae will be seen more easily when the light intensity is reduced. This technique can be utilized to identify hyphae in dermatophyte infections (see Plate IID-51), pseudohyphae and budding yeast in *Candida* infections (see Plate IID-43), and fragmented hyphae and spores in tinea versicolor. The same sampling technique can be used to obtain scale for culture of selected pathogenic organisms.

Tzanck Smear A Tzanck smear is a cytologic technique most often used in the diagnosis of herpesvirus infections [simplex or varicella-zoster (seePlates IID-36 andIID-37)]. An early vesicle, not a pustule or crusted lesion, is unroofed, and the base of the lesion is scraped gently with a scalpel blade. The material is placed on a glass slide, air-dried, and stained with Giemsa or Wright's stain. Multinucleated giant cells suggest the presence of herpes, but culture or immunofluorescence testing must be performed to identify the specific virus.

Diascopy Diascopy is designed to assess whether a skin lesion will blanch with pressure as, for example, in determining whether a red lesion is hemorrhagic or simply blood-filled. For instance, a hemangioma (see Plate IIA-17) will blanch with pressure, whereas a purpuric lesion caused by necrotizing vasculitis (seePlate IIE-71) will not. Diascopy is performed by pressing a microscope slide or magnifying lens against a specified lesion and noting the amount of blanching that occurs. Granulomas often have an "apple jelly" appearance on diascopy.

Wood's Light A Wood's lamp generates 360-nm ultraviolet (or "black") light that can be used to aid the evaluation of certain skin disorders. For example, a Wood's lamp will cause erythrasma (a superficial, intertriginous infection caused by *Corynebacterium minutissimum*) to show a characteristic coral red color, and wounds colonized by *Pseudomonas* to appear pale blue. Tinea capitis caused by certain dermatophytes such as *Microsporum canis* or *M. audouini* exhibits a yellow fluorescence. Pigmented lesions of the epidermis such as freckles are accentuated, while dermal pigment such as postinflammatory hyperpigmentation fades under a Wood's light. Vitiligo (seePlate IIA-11) appears totally white under a Wood's lamp, and previously unsuspected areas of involvement often become apparent. A Wood's lamp may also aid in the demonstration of tinea versicolor and in recognition of ash leaf spots in patients with tuberous sclerosis.

Patch Tests Patch testing is designed to document sensitivity to a specific antigen. In this procedure, a battery of suspected allergens is applied to the patient's back under occlusive dressings and allowed to remain in contact with the skin for 48 h. The dressings are removed, and the area is examined for evidence of delayed hypersensitivity reactions (e.g., erythema, edema, or papulovesicles). This test is best performed by physicians with special expertise in patch testing and is often helpful in the evaluation of patients with chronic dermatitis.

(Bibliography omitted in Palm version)

Back to Table of Contents

56. ECZEMA, PSORIASIS, CUTANEOUS INFECTIONS, ACNE, AND OTHER COMMON SKIN DISORDERS - *Robert A. Swerlick*, *Thomas J. Lawley*

ECZEMA AND DERMATITIS

Eczema, or dermatitis, is a reaction pattern that presents with variable clinical and histologic findings and is the final common expression for a number of disorders, including atopic dermatitis, allergic contact and irritant contact dermatitis, dyshidrotic eczema, nummular eczema, lichen simplex chronicus, asteatotic eczema, and seborrheic dermatitis. Primary lesions may include papules, erythematous macules, and vesicles, which can coalesce to form patches and plaques. In severe eczema, secondary lesions from infection or excoriation, marked by weeping and crusting, may predominate. Long-standing dermatitis is often dry and is characterized by thickened, scaling skin (*lichenification*).

ATOPIC DERMATITIS

Atopic dermatitis (AD) is the cutaneous expression of the atopic state, characterized by a family history of asthma, hay fever, or dermatitis in up to 70% of patients. The criteria for the diagnosis of atopic eczema are shown inTable 56-1. The prevalence of atopic dermatitis is increasing worldwide, with a point prevalence in Norwegian school children as high as 23%.

The etiology ofAD is only partially defined. There is a clear genetic predisposition. When both parents are affected by AD, over 80% of their children manifest the disease. When only one parent is affected, the prevalence drops to slightly over 50%. A number of genes have been tentatively linked to AD including genes coding for IgE, the high-affinity IgE receptor, mast cell tryptase, and interleukin (IL) 4. Patients with AD may display a variety of immunoregulatory abnormalities including increased IgE synthesis; increased specific IgE to foods, aeroallergens, bacteria, and bacterial products; increased expression of CD23 (low-affinity IgE receptor) on monocytes and B cells; impaired delayed type hypersensitivity reactions; and increased type II and decreased type I cytokine responses.

The clinical presentation often varies with age. Half of patients withADpresent within the first year of life, and 80% present by 5 years of age. Some 80% ultimately coexpress allergic rhinitis or asthma later in life. The infantile pattern is characterized by weeping inflammatory patches and crusted plaques that occur on the face, neck, extensor surfaces, and groin. The childhood and adolescent pattern is marked by dermatitis of flexural skin, particularly in the antecubital and popliteal fossae (seePlate IIA-4). AD may resolve spontaneously in adults, but the dermatitis will persist into adult life in over half of individuals affected as children. The distribution of lesions may be similar to those seen in childhood. However, adults affected with AD frequently have localized disease, manifesting as hand eczema or lichen simplex chronicus (see below).

Pruritus is a prominent characteristic of AD, and many of the cutaneous findings in affected patients are secondary to rubbing and scratching. Other cutaneous stigmata of AD are perioral pallor, an extra fold of skin beneath the lower eyelid (Dennie's line), increased palmar markings, and increased incidence of cutaneous infections, particularly with *Staphylococcus aureus*. Atopic individuals often have dry itchy skin, abnormalities in cutaneous vascular responses, and, in some instances, elevations in serum IgE.

Histologic examinaton of the skin affected byAD may demonstrate features of acute or chronic dermatitis. Immunopathology shows activated, memory T helper cells, which express the cutaneous lymphocyte antigen, the ligand for the inducible endothelial cell adhesion molecule E-selectin. AD skin lesions may also demonstrate IgE-bearing CD1a+ positive Langerhans cells, and these cells have been implicated in AD disease pathophysiology through mediation of hypersensitivity responses to environmental antigens.

TREATMENT

Therapy of ADshould be based on avoidance of cutaneous irritants, adequate cutaneous hydration, judicious use of low- or midpotency topical glucocorticoids, and prompt treatment of secondarily infected skin lesions. Patients should be instructed to bathe using warm, but not hot, water and to limit their use of soap. Immediately after bathing while the skin is still moist, the skin should be lubricated with a low- or midpotency topical glucocorticoid in a cream or ointment base. Potent fluorinated topical glucocorticoids should not be used on the face or intertriginous areas. It takes a minimum of 30 g of glucocorticoid ointment to cover the entire body surface of an average adult.

Crusted and weeping skin lesions should be treated with systemic antibiotics with activity against *S. aureus* since secondary infection often exacerbates eczema. The frequency of macrolide-resistant organisms makes the use of penicillinase-resistant penicillins or cephalosporins preferable. Dicloxacillin or cephalexin (250 mg four times daily for 7 to 10 days) is generally adequate to decrease heavy colonization. As an adjunct, the use of triclosan-containing antibacterial washes and intermittent nasal mupirocin may be useful as prophylactic measures. The role of dietary allergens in atopic dermatitis is controversial, and there is little evidence that they play any role outside of infancy.

Control of pruritus is essential for treatment, sinceAD often represents "an itch that rashes." Antihistamines are useful to control the pruritus, but sedation may limit their usefulness. Unlike their effects in urticaria, nonsedating antihistamines are of little use since the effectiveness of antihistamines in the treatment of pruritus associated with AD is primarily related to their sedative effects as opposed to any specific action on histamine-mediated pathways.

Treatment with systemic glucocorticoids should be limited to severe exacerbations unresponsive to conservative topical therapy. In the patient with chronicAD, therapy with systemic glucocorticoids will generally clear the skin only briefly, but cessation of the systemic therapy will invariably be accompanied by return, if not worsening, of the dermatitis. Patients who do not respond to conventional therapies should be considered for patch testing to rule out allergic contact dermatitis. Immunotherapy with aeroallergens has not proven useful in AD, unlike its effect in allergic rhinitis and extrinsic asthma.

CONTACT DERMATITIS

Contact dermatitis is an inflammatory process in skin caused by an exogenous agent or agents that directly or indirectly injure the skin. This injury may be caused by an inherent characteristic of a compound -- irritant contact dermatitis (ICD). An example of ICD would be dermatitis induced by a concentrated acid or base. Agents that cause allergic contact dermatitis (ACD) induce an antigen-specific immune response. The clinical lesions of contact dermatitis may be acute (wet and edematous) or chronic (dry, thickened, and scaly), depending on the persistence of the insult (see Plate IIA-8). The most common presentation of contact dermatitis is hand eczema, and it is frequently related to occupational exposures. Occupation-related contact dermatitis represents a significant proportion of occupation-induced injury, affecting over 60,000 persons annually.

ICDis generally strictly demarcated and often localized to areas of thin skin (eyelids, intertriginous areas) or to areas where the irritant was occluded. Lesions may range from minimal skin erythema to areas of marked edema, vesicles, and ulcers. Chronic low-grade irritant dermatitis is the most common type of ICD and the most common area of involvement is the hands (see below). The most common irritants encountered are chronic wet work, soaps, and detergents. Treatment should be directed to avoidance of irritants and use of protective gloves or clothing.

ACDis a manifestation of delayed type hypersensitivity mediated by memory T lymphocytes in the skin. The most common cause of ACD is exposure to plants, specifically to members of the family Anacardiaceae; including the genera *Toxicodendrun*, *Anacardium*, *Gluta*, *Mangifera*, and *Semecarpus*. Poison ivy, poison oak, and poison sumac are members of the genus *Toxicodendron* and cause an allergic reaction marked by erythema, vesiculation, and severe pruritus. The eruption is often linear, corresponding to areas where plants have touched the skin. However, other allergens may be more difficult to identify, especially if the exposure is chronic and the skin becomes thickened and scaly. The sensitizing antigen common to these plants is urushiol, an oleoresin containing the active ingredient pentadecylcatechol. The oleoresin may adhere to skin, clothing, tools, and pets, and contaminated articles may cause dermatitis even after prolonged storage. Blister fluid does not contain urushiol and is not capable of inducing skin eruption in exposed subjects.

TREATMENT

IfACD is suspected and an offending agent is identified and removed, the eruption will resolve. Usually, treatment with high-potency fluorinated topical glucocorticoids is enough to relieve symptoms while the ACD runs its course. For those patients who require systemic therapy, a tapering course over 2 to 3 weeks given as single morning doses is the preferred method.

Identification of a contact allergen can be a difficult and time-consuming task. Patients with dermatitis unresponsive to conventional therapy or with an unusual and patterned distribution should be suspected of havingACD. They should be questioned carefully regarding occupational exposures, topical medicaments, and oral medications.

Common sensitizers include preservatives in topical preparations, nickel sulfate, potassium dichromate, thimerosal in ocular preparations, neomycin sulfate, fragrances, formaldehyde, and rubber-curing agents. Patch testing is helpful in identifying these agents, but should not be attempted on patients with widespread active dermatitis or on those taking systemic glucocorticoids.

HAND ECZEMA

Hand eczema is a very common, chronic skin disorder. It represents a large proportion of occupation-associated skin disease. It may be associated with other cutaneous disorders such as atopic dermatitis or may occur by itself. Similar to other forms of dermatitis, both exogenous and endogenous factors play important roles in the expression of hand dermatitis. Chronic, excessive exposure to water and detergents may initiate or aggravate this disorder. It may present with dryness and cracking of the skin of the hands as well as with variable amounts of erythema and edema. Often, the dermatitis will begin under rings where water and irritants are trapped. A variant of hand dermatitis, dyshidrotic eczema, presents with multiple, intensely pruritic, small papules and vesicles occurring on the thenar and hypothenar eminences and the sides of the fingers (see Plate IA-5). Lesions tend to occur in crops that slowly form crusts and heal.

The evaluation of a patient with hand eczema should include an assessment of potential occupation-associated exposures. Predominant involvement of the dorsal surface of the hands with sparing of the palmar surface suggests a possible contact dermatitis. The history should be directed to identifying possible irritant or allergen exposures. The use of rubber gloves to protect dermatitic skin is sometimes associated with the development of delayed type hypersensitivity reactions to agents used for cross-linking rubber. Such reactions can be detected by patch testing. Less commonly, patients may manifest hand dermatitis as a consequence of developing immediate type hypersensitivity reactions to latex. These are of particular concern since these patients are at risk for anaphylactic reactions. The most sensitive method of detection is the use of scratch testing with latex extract. However, this should be done with extreme caution only in a setting where an anaphylactic reaction can be treated. A latex radioallergosorbent test is available but is only about 60%sensitive.

TREATMENT

Therapy of hand dermatitis is directed toward avoidance of irritants, identification of possible contact allergens, treatment of coexistent infection, and application of topical glucocorticoids. Whenever possible, the hands should be protected by gloves, preferably vinyl. Most patients can be treated with cool moist compresses (dressings) to dry and debride acute inflammatory lesions and to decrease swelling, followed by application of a mid- to high-potency topical glucocorticoid in a cream or ointment base. As with atopic dermatitis, treatment of secondary infection by staphylococci or streptococci is essential for good control. Additionally, patients with hand dermatitis should be examined for dermatophyte infection by KOH preparation and culture (see below).

NUMMULAR ECZEMA

Nummular eczema is characterized by circular or oval "coinlike" lesions. Initially, this eruption consists of small edematous papules that become crusted and scaly. The most common locations are on the trunk or the extensor surfaces of the extremities, particularly on the pretibial areas or dorsum of the hands. It occurs more frequently in men and is most commonly seen in middle age. The etiology of nummular eczema is unknown. Whether nummular eczema represents a variant of atopic eczema is controversial. The treatment of nummular eczema is similar to that for other forms of dermatitis.

LICHEN SIMPLEX CHRONICUS

Lichen simplex chronicus may represent the end stage of a variety of pruritic and eczematous disorders. It consists of a well-circumscribed plaque or plaques with lichenified or thickened skin due to chronic scratching or rubbing. Common areas involved include the posterior nuchal region, dorsum of the feet, or ankles. Treatment of lichen simplex chronicus centers around breaking the cycle of chronic itching and scratching, which often occur during sleep. High-potency topical glucocorticoids are helpful in alleviating pruritus in most cases, but in recalcitrant cases, application of topical glucocorticoids under occlusion or intralesional injection of glucocorticoids may be required. Oral antihistamines such as hydroxyzine (10 to 50 mg every 6 h) or tricyclic antidepressants with antihistaminic activity such as doxepin (10 to 25 mg at bedtime) are useful as antipruritics primarily due to their sedating action, and are particularly useful at bedtime (see above). Patients need to be counseled regarding driving or operating heavy equipment after taking these medications due to their potentially potent sedative activity.

ASTEATOTIC ECZEMA

Asteatotic eczema, also known as xerotic eczema or "winter itch," is a mildly inflammatory variant of dermatitis that develops most commonly on the lower legs of elderly individuals during dry times of year. Fine cracks, with or without erythema, characteristically develop on the anterior surface of the lower extremities. Pruritus is variable. Asteatotic eczema responds well to avoidance of irritants, rehydration of the skin, and application of topical emollients.

STASIS DERMATITIS AND STASIS ULCERATION

Stasis dermatitis develops on the lower extremities secondary to venous incompetence and chronic edema. Early findings in stasis dermatitis consist of mild erythema and scaling associated with pruritus. The typical initial site of involvement is the medial aspect of the ankle, often over a distended vein (seePlate IIA-7). As the disorder progresses, the dermatitis becomes progressively pigmented, due to chronic erythrocyte extravasation leading to cutaneous hemosiderin deposition. As with other forms of dermatitis, stasis dermatitis may become acutely inflamed, with crusting and exudate. Chronic stasis dermatitis is often associated with dermal fibrosis that is recognized clinically as brawny edema of the skin. Stasis dermatitis is often complicated by secondary infection and contact dermatitis. Severe stasis dermatitis may precede the development of stasis ulcers.

TREATMENT

Avoidance of irritants and use of emollients and/or midpotency topical glucocorticoids are the cornerstones of therapy for stasis dermatitis. Control of chronic edema is important to prevent leg ulcers. Patients should be encouraged to elevate the affected extremity when sitting. A compression stocking with a gradient of at least 30 to 40 mmHg is most effective for edema control and is much more effective for preventing chronic edema than is antiembolism hose.

Stasis ulcers are difficult to treat, and resolution of these lesions is slow even under the best of circumstances. It is extremely important to elevate the affected limb as much as possible. The ulcer should be kept clear of necrotic material by gentle debridement and covered with a semipermeable dressing under pressure. Glucocorticoids should not be applied to ulcers, since they may retard healing. Secondarily infected lesions should be treated with appropriate oral antibiotics, but it should be noted that all ulcers will become colonized with bacteria, and the purpose of antibiotic therapy should not be to clear all bacterial growth. Some ulcers may take months to heal or require skin grafting.

SEBORRHEIC DERMATITIS

Seborrheic dermatitis is a common, chronic disorder, characterized by greasy scales overlying erythematous patches or plaques. The most common location is in the scalp where it may be recognized as severe dandruff. On the face, seborrheic dermatitis affects the eyebrows, eyelids, glabella, nasolabial fold, or ears (seePlate IIA-6). Scaling within the external ear is often mistaken for a chronic fungal infection (otomycosis), and postauricular dermatitis often becomes macerated and tender. Additionally, seborrheic dermatitis may develop in the central chest, axilla, groin, submammary folds, and gluteal cleft. Rarely, it may cause a widespread generalized dermatitis. Seborrheic dermatitis is usually symptomatic, with patients complaining of itching or burning.

Seborrheic dermatitis may be evident within the first few weeks of life, and within this context it occurs in the scalp ("cradle cap"), face, or groin. It is rarely seen in children beyond infancy but becomes evident again during adult life. Although it is frequently seen in patients with Parkinson's disease, in those who have had cerebrovascular accidents, and in those with human immunodeficiency virus (HIV) infection, the overwhelming majority of individuals with seborrheic dermatitis have no underlying disorder.

TREATMENT

Treatment with low-potency topical glucocorticoids in conjunction with shampoos containing coal tar and/or salicylic acid is generally sufficient to control activity of this disorder. High-potency topical glucocorticoid solutions (betamethasone or fluocinonide) are effective for control of scalp involvement. Fluorinated topical glucocorticoids should not be used on the face since this is often associated with the development of rebound worsening and steroid-induced rosacea or atrophy.

PAPULOSQUAMOUS DISORDERS (Table 56-2)

PSORIASIS

Psoriasis is one of the most common dermatologic diseases, affecting up to 1 to 2% of the world's population. It is a chronic inflammatory skin disorder clinically characterized by erythematous, sharply demarcated papules and rounded plaques, covered by silvery micaceous scale. The skin lesions of psoriasis are variably pruritic. Traumatized areas often develop lesions of psoriasis (Koebner or isomorphic phenomenon). Additionally, other external factors may exacerbate psoriasis including infections, stress, and medications (lithium, beta blockers, and antimalarials).

The most common variety of psoriasis is called *plaque type*. Patients with plaque-type psoriasis will have stable, slowly growing plaques, which remain basically unchanged for long periods of time. The most common areas for plaque psoriasis to occur are the elbows, knees, gluteal cleft, and the scalp. Involvement tends to be symmetric. *Inverse psoriasis* affects the intertriginous regions including the axilla, groin, submammary region, and navel, it also tends to affect the scalp, palms, and soles. The individual lesions are sharply demarcated plaques (seePlate IIA-3) but may be moist due to their location. Plaque psoriasis generally develops slowly and runs an indolent course. It rarely remits spontaneously.

Eruptive psoriasis (guttate psoriasis) is most common in children and young adults. It develops acutely in individuals without psoriasis or in those with chronic plaque psoriasis. Patients present with many small erythematous, scaling papules, frequently after upper respiratory tract infection withb-hemolytic streptococci. The differential diagnosis should include pityriasis rosea and secondary syphilis. Patients with psoriasis may also develop pustular lesions. These may be localized to the palms and soles or may be generalized and associated with fever, malaise, diarrhea, and arthralgias.

About half of all patients with psoriasis have fingernail involvement, appearing as punctate pitting, nail thickening, or subungual hyperkeratosis. About 5 to 10% of patients with psoriasis have associated joint complaints, and these are most often found in patients with fingernail involvement. Although some have the coincident occurrence of classic rheumatoid arthritis (Chap. 312), many have joint disease that falls into one of three types associated with psoriasis: (1) asymmetric inflammatory arthritis most commonly involving the distal and proximal interphalangeal joints and less commonly the knees, hips, ankles, and wrists; (2) a seronegative rheumatoid arthritis-like disease; a significant portion of these patients go on to develop a severe destructive arthritis; or (3) disease limited to the spine (psoriatic spondylitis).

The etiology of psoriasis is still poorly understood. There is clearly a genetic component to psoriasis. Over 50% of patients with psoriasis report a positive family history, and a 65 to 72% concordance among monozygotic twins has been reported in twin studies. Psoriasis has been linked to HLA-Cw6 and, to a lesser extent, to HLA-DR7. Evidence has accumulated clearly indicating a role for T cells in the pathophysiology of psoriasis. Stimulation of immune function with cytokines such asIL-2 has been associated with abrupt worsening of preexisting psoriasis, and bone marrow transplantation has resulted in clearance of disease. Psoriatic lesions are characterized by infiltration of skin with activated memory T cells, with CD8+ cells predominating in the epidermis. Agents that inhibit activated T cell function are often effective for the treatment of severe psoriasis.

Presumably, cytokines from activated T cells elaborate growth factors that stimulate keratinocyte hyperproliferation.

TREATMENT

Treatment of psoriasis depends on the type, location, and extent of disease. All patients should be instructed to avoid excess drying or irritation of their skin and to maintain adequate cutaneous hydration. Most patients with localized plaque-type psoriasis can be managed with midpotency topical glucocorticoids, although their long-term use is often accompanied by loss of effectiveness (tachyphylaxis). Crude coal tar (1 to 5% in an ointment base) is an old but useful method of treatment in conjunction with ultraviolet light therapy. A topical vitamin D analogue (calcipitriol) is also efficacious in the treatment of psoriasis.

Ultraviolet light is an effective therapy for patients with widespread psoriasis. The ultraviolet B (UV-B) spectrum is effective alone, or may be combined with coal tar (Goeckerman regimen) or anthralin (Ingram regimen). Natural sunlight or an artificial light source can be used. The combination of the ultraviolet A (UV-A) spectrum with either oral or topical psoralens (PUVA) is also extremely effective for the treatment of psoriasis, but long-term use may be associated with an increased incidence of squamous cell cancer and melanoma of the skin.

Various other agents can be used for widespread psoriatic disease. Methotrexate is an effective agent, especially in patients with associated psoriatic arthritis. Liver toxicity from long-term use limits its use to patients with widespread disease not responsive to less aggressive modalities. The synthetic retinoid, acetretin, has been shown to be effective in some patients with severe psoriasis but is a potent teratogen, thus limiting its use in women with childbearing potential. The evidence implicating psoriasis as a T cell-mediated disorder has created a new perspective relating to the treatment of psoriasis. Based on this presumed disease mechanism, immunomodulatory therapy utilizing cyclosporine has proven to be highly effective in selected patients with severe, crippling, and potentially life-threatening disease.

LICHEN PLANUS

Lichen planus (LP) is a papulosquamous disorder in which the primary lesions are pruritic, polygonal, flat-topped, violaceous papules. Close examination of the surface of these papules often reveals a network of gray lines (Wickham's striae). The skin lesions may occur anywhere but have a predilection for the wrists, shins, lower back, and genitalia (see Plate IIA-9). Involvement of the scalp may lead to hair loss. LP commonly involves mucous membranes, particularly the buccal mucosa, where it can present as a white netlike eruption. Its etiology is unknown, but cutaneous eruptions clinically resembling LP have been observed after administration of numerous drugs, including diuretics, gold, antimalarials, penicillamine, and phenothiazines, and in patients with skin lesions of chronic graft-versus-host disease. Additionally, LP associated with abnormal liver function has been correlated with viral hepatitis, particularly hepatitis C infection. The course of LP is variable, but most patients have spontaneous remissions 6 months to 2 years after the onset of disease. Topical glucocorticoids are the mainstay of therapy.

PITYRIASIS ROSEA

Pityriasis rosea (PR) is a papulosquamous eruption of unknown etiology that occurs more commonly in the spring and fall. Its first manifestation is the development of a 2- to 6-cm annular lesion (the herald patch). This is followed in a few days to a few weeks by the appearance of many smaller annular or papular lesions with a predilection to occur on the trunk (see Plate IIA-13). The lesions are generally oval, with their long axis parallel to the skin-fold lines. Individual lesions may range in color from red to brown and have a trailing scale. PR shares many clinical features with the eruption of secondary syphilis, but palm and sole lesions are extremely rare in PR and common in secondary syphilis. The eruption tends to be moderately pruritic and lasts 3 to 8 weeks. Treatment is generally directed at alleviating pruritus and consists of oral antihistamines, midpotency topical glucocorticoids, and, in some cases, the use ofUV-Bphototherapy.

CUTANEOUS INFECTIONS (Table 56-3)

IMPETIGO AND ECTHYMA

Impetigo is a common superficial bacterial infection of skin caused by group Ab-hemolytic streptococci (Chap. 140) or *S. aureus* (Chap. 139). The primary lesion is a superficial pustule that ruptures and forms a characteristic yellow-brown honey-colored crust (seePlate IID-38). Lesions caused by staphylococci may be tense, clear bullae, and this less common form of the disease is called *bullous impetigo*. Lesions may occur on normal skin or in areas already affected by another skin disease. Ecthyma is a variant of impetigo that generally occurs on the lower extremities and causes punched-out ulcerative lesions. Treatment of both ecthyma and impetigo involves gentle debridement of adherent crusts, which is facilitated by the use of soaks and topical antibiotics, in conjunction with appropriate oral antibiotics.

ERYSIPELAS AND CELLULITIS

SeeChap. 128

DERMATOPHYTOSIS

Dermatophytes are fungi that infect skin, hair, and nails and include members of the genera *Trichophyton*, *Microsporum*, and *Epidermophyton*. Infection of the foot (tinea pedis) is most common and is often chronic; it is characterized by variable erythema and edema, scaling, pruritus, and occasionally vesiculation. Involvement may be widespread or localized, but almost invariably the web space between the fourth and fifth toes is affected. Infection of the nails (tinea unguium) occurs in many patients with tinea pedis and is characterized by opacified, thickened nails and subungual debris. The groin is the next most commonly involved area (tinea cruris), with males affected much more often than females. It presents as a scaling erythematous eruption that spares the scrotum. Microscopic examination of either untreated tinea pedis or tinea cruris scale after digestion with KOH preparation will generally demonstrate hyphae.

Dermatophyte infection of the scalp (tinea capitis) has returned in epidemic proportions, particularly affecting inner city children. The predominant organism is *T. tonsurans*. This organism can produce an inflammatory or relatively noninflammatory infection that may present with either well-defined or irregular, diffuse areas of mild scaling and hair loss. Tinea corporis, or infection on non-hair-bearing skin, may have a variable appearance, depending on the extent of the associated inflammatory reaction (see Plate IID-51). It may have the typical annular appearance of "ringworm" or appear as deep inflammatory nodules (on the scalp known as a *kerion*) or granulomas. KOH examination of scale or hair from patients with tinea capitis or inflammatory tinea corporis often does not reveal hyphae, and diagnosis may require culture or biopsy.

TREATMENT

Both topical and systemic therapies may be used to treat dermatophyte infection. Treatment depends on the site involved and the type of infection. Topical therapy is generally effective for uncomplicated tinea corporis, tinea cruris, and limited tinea pedis. It is not effective as monotherapy for tinea capitis or tinea unguium. Topical imidazoles (miconazole, ketoconazole, econazole, clotrimazole, oxiconazole, and sulconazole), triazoles (terconazole), and allylamines (terbinafine and naftifine) may all be effective topical therapies for dermatophyte infections. Haloprogin, undecylic acid, ciclopirox-olamine, and tolnaftate are also effective, but nystatin is not active against dermatophytes. Treatment should continue until the patient is clear of infection by clinical examination and culture. Tinea pedis often requires longer treatment courses and is associated with a high relapse rate.

Griseofulvin is the drug of choice for dermatophyte infections requiring systemic therapy. A daily dose of 500 mg of microsized or 350 mg of ultramicrosized griseofulvin administered with a fatty meal is an adequate dose for most dermatophyte infections. The duration of therapy may be as short as 2 weeks for uncomplicated tinea corporis but may be as long as 6 to 12 months for nail infections. The most common side effects of griseofulvin are gastrointestinal distress and headache. Dermatophyte infection of hair-bearing areas (e.g., tinea capitis) requires systemic antifungal therapy. The usual adult dose of griseofulvin is 1 g of microsized or 0.5 g of ultramicrosized given daily, and treatment should be continued for 6 to 8 weeks. Children should be treated with 15 to 20 mg/kg as a single daily dose given with a fatty meal. The adjunctive use of topical antifungal agents in addition to systemic therapy may be useful, but topical therapy alone is not adequate. Markedly inflammatory tinea capitis may result in scarring and hair loss, and systemic or topical glucocorticoids may be helpful in preventing this sequela. Recent studies in children have also suggested that both itraconazole (3 to 5 mg/kg for 6 to 10 weeks) and terbinafine (125 mg/d for 6 weeks) may be effective treatments for tinea capitis.

Until recently, griseofulvin was the recommended therapy for dermatophyte infection of the nails. However, despite prolonged treatment, cure rates were poor. Itraconazole given as either continuous daily therapy (200 mg/d for 3 months) or pulses (200 mg twice daily for 1 week per month for 3 consecutive months) has been shown to be a safe and effective therapy. Itraconazole has the potential for interactions with other drugs requiring the P450 enzyme system for metabolism. Similarly, terbinafine (250 mg/d for 3 months) has shown similar cure rates. Only limited data are available on the dosing and effectiveness of the newer antifungal agents in tinea corporis, tinea cruris, and uncomplicated tinea pedis.

TINEA VERSICOLOR

Tinea versicolor is caused by a nondermatophyte dimorphic fungus that is a normal inhabitant of the skin. As the yeast form *Pityrosporum orbiculare*, it generally does not cause disease (except for folliculitis in certain individuals). However, in some individuals, it converts to the hyphal form and causes characteristic lesions. The expression of infection is promoted by heat and humidity. The typical lesions consist of oval scaly macules, papules, and patches concentrated on the chest, shoulders, and back but only rarely on the face or distal extremities. On dark skin, they often appear as hypopigmented areas, while on light skin, they are slightly hyperpigmented. In some darkly pigmented individuals, they may only appear as scaling patches. A KOH preparation from scaling lesions will demonstrate a confluence of short hyphae and round spores (so-called spaghetti and meatballs). Solutions containing sulfur, salicylic acid, or selenium sulfide will clear the infection if used daily for a week and then intermittently thereafter. Treatment with a single 400-mg dose of ketaconazole is also effective.

CANDIDIASIS

Candidiasis is a fungal infection caused by a related group of yeasts, whose manifestations may be localized to the skin, or rarely, may be systemic and life-threatening. The causative organism is usually *Candida albicans*, but may also be *C. tropicalis*, *C. parapsilosis*, or *C. krusei*. These organisms are normal saprophytic inhabitants of the gastrointestinal tract but may overgrow (usually due to broad-spectrum antibiotic therapy) and cause disease at a number of cutaneous sites. Other predisposing factors include diabetes mellitus, chronic intertrigo, oral contraceptive use, and cellular immune deficiency. Candidiasis is a very common infection in HIV-infected individuals (Chap. 309). The oral cavity is commonly involved. Lesions may occur on the tongue or buccal mucosa (thrush) and appear as white plaques (see Plate IID-43). Microscopic examination of scrapings demonstrate both pseudohyphae and yeast forms. Fissured, macerated lesions at the corners of the mouth (perleche) are often seen in individuals with poorly fitting dentures and may also be associated with candidal infection. Additionally, candidal infections have an affinity for sites that are chronically wet and macerated and may occur around nails (onycholysis and paronychia) and in intertriginous areas. Intertriginous lesions are characteristically edematous, erythematous, and scaly, with scattered "satellite pustules." In males, there is often involvement of the penis and scrotum as well as the inner aspect of the thighs. In contrast to dermatophyte infections, candidal infections are frequently accompanied by a marked inflammatory response. Diagnosis of candidal infection is based upon the clinical pattern and demonstration of yeast on KOH preparation, or culture.

TREATMENT

Treatment routinely involves removing any predisposing factors such as antibiotic therapy or chronic wetness and the use of appropriate topical or systemic antifungal therapy. Effective topical agents include nystatin or topical azoles (miconazole, clotrimazole, econazole, or ketoconazole). These agents are generally effective in clearing mucous membrane or glabrous skin involvement in nonimmunosuppressed patients. The associated inflammatory response that often accompanies candidal infection on glabrous skin should be treated with a mild glucocorticoid lotion or cream (2.5% hydrocortisone). Systemic therapy is generally reserved for immunosuppressed patients or individuals with chronic or recurrent disease who fail to respond to or tolerate appropriate topical therapy. Vulvovaginal candidiasis may respond to treatment with a single dose of fluconazole (150 mg). Chronic recurrent oral or vaginal candidiasis may be treated with weekly to monthly oral fluconazole (150 to 200 mg) in conjunction with topical therapy.

WARTS

Warts are cutaneous neoplasms that are caused by papilloma viruses. Over 50 different human papilloma viruses (HPV) have been described, and this number will almost certainly continue to grow. Typical verruca vulgaris lesions are sessile, dome-shaped, usually about a centimeter in diameter, and their surface is made up of many small filamentous projections. The HPV that cause typical verruca vulgaris also cause typical plantar warts, flat warts (or verruca plana), and filiform warts in intertriginous areas. Plantar warts are endophytic and are covered by thick keratin. Paring of the wart will generally demonstrate a central core of keratinized debris and punctate bleeding points. Filiform warts are most commonly seen on the face, neck, and skin folds and present as papillomatous lesions on a narrow base. Flat warts are only slightly elevated and have a velvety, nonverrucous surface. They have a propensity for the face, arms, and legs and are often spread by shaving.

MultipleHPVtypes have been associated with genital tract lesions. They generally begin as small papillomas that may grow to form large fungating lesions. In women, they may involve either the labia, perineum, or perianal skin. Additionally, the mucosa of the vagina, urethra, and anus can be involved, as well as the cervical epithelium. In men, the lesions often occur initially in the coronal sulcus, but may be seen on the shaft of the penis, the scrotum, perianal skin, or in the urethra.

Within the past decade, appreciable evidence has accumulated that suggestsHPVplays a role in the development of neoplasia of the uterine cervix and external genitalia (Chap. 97). HPV types 16 and 18 have been most intensely studied, while recent evidence also implicates other types. Lesions may initially appear as small, flat, velvety, hyperpigmented papules occurring on the genitalia or perianal skin. Histologic examination of biopsies from affected sites may reveal changes associated with typical warts and/or features typical of intraepidermal carcinoma (Bowen's disease). Squamous cell carcinomas associated with HPV infections have also been observed in extragenital skin (Chap. 86). This is most commonly seen in patients immunosuppressed after organ transplantation.

TREATMENT

There are many modalities available to treat warts, but no single therapy is universally effective. Factors that influence the choice of therapy include the location of the wart, extent of disease, the age and immunologic status of the patient, and the patient's desire for therapy. Perhaps the most useful and convenient method for treating warts in almost any location is cryotherapy with liquid nitrogen. Equally effective, but requiring much more patient compliance, is the use of keratolytic agents such as salicylic acid plasters or combinations of lactic acid and salicylic acid. For genital warts, application of podophyllin solution is moderately effective but may be associated with marked local reactions in certain individuals. Dilute preparations of purified podophyllin permit physician-directed by patient-applied use, facilitating treatment of mucosal warts. Topical imiquimod, a potent inducer of local cytokine release, has also been approved for use in genital warts. Other topical agents that are used include trichloracetic acid or cantharidin. Electrodessication and curettage or CO_2laser excision are also effective therapies but require local anesthesia. Recurrence of warts appears to be common to all these modalities because viral genomic material is present in normal-appearing skin adjacent to the clinical lesions.

Treatment of warts should be tempered by the observation that an overwhelming majority of warts in normal individuals resolve spontaneously within 1 to 2 years. Also, only an extremely small proportion of warts is associated with neoplasia, and those are almost exclusively located on the genitalia or perianal skin.

HERPES SIMPLEX

**See Chap. 182*

HERPES ZOSTER

**See Chap. 183*

ACNE

ACNE VULGARIS

Acne vulgaris is a self-limited disorder primarily of teenagers and young adults, although perhaps 10 to 20% of adults may continue to experience some form of the disorder. The permissive factor for the expression of the disease in adolescence is the increase in sebum release by sebaceous glands after puberty. Small cysts, called *comedones*, form in hair follicles due to blockage of the follicular orifice by retention of sebum and keratinous material. The activity of lipophilic yeast (*Pityrosporum orbiculare*) and bacteria (*Proprionobacterium acnes*) within the comedones releases free fatty acids from sebum, causes inflammation within the cyst, and results in rupture of the cyst wall. An inflammatory foreign-body reaction develops as a result of extrusion of oily and keratinous debris from the cyst.

The clinical hallmark of acne vulgaris is the comedone, which may be closed (whitehead) or open (blackhead). Closed comedones appear as 1- to 2-mm pebbly white papules, which are accentuated when the skin is stretched. They are the precursors of inflammatory lesions of acne vulgaris. The contents of closed comedones are not easily expressed. Open comedones, which rarely result in inflammatory acne lesions, have a large dilated follicular orifice and are filled with easily expressible oxidized, darkened, oily debris. Comedones are usually accompanied by inflammatory lesions: papules, pustules, or nodules.

The earliest lesions seen in early adolescence are generally mildly inflamed or noninflammatory comedones on the forehead. Subsequently, more typical inflammatory lesions develop on the cheeks, nose, and chin (seePlate IIA-1). The most common location for acne is the face, but involvement of the chest and back is not uncommon. Most disease remains mild and does not lead to scarring. However, a small number of patients develop large inflammatory cysts and nodules, which may drain and result in significant scarring.

Exogenous and endogenous factors can alter the expression of acne vulgaris. Friction and trauma may rupture preexisting microcomedones and elicit inflammatory acne lesions. This is commonly seen with headbands or chin straps of athletic helmets. Application of comedogenic topical agents in cosmetics or hair preparations or chronic topical exposure to certain industrial compounds that are comedogenic may elicit or aggravate acne. Glucocorticoids, applied topically or administered systemically in high doses, may also elicit acne. Other systemic medications such as lithium, isoniazid, halogens, phenytoin, and phenobarbital may produce acneiform eruptions, or aggravate preexisting acne.

TREATMENT

Treatment of acne vulgaris is directed toward elimination of comedones by normalization of follicular keratinization, decreasing sebaceous gland activity, decreasing the population of lipophilic bacteria and yeast, and decreasing inflammation. Acne vulgaris may be treated with either local or systemic medications. Minimal to moderate, pauci-inflammatory disease may respond adequately to local therapy alone. Although areas affected with acne should be kept clean, there is little evidence to suggest that removal of surface oils plays an important role in therapy. Overly vigorous scrubbing may aggravate acne due to mechanical rupture of comedones. Topical agents such as retinoic acid, benzoyl peroxide, or salicylic acid may alter the pattern of epidermal desquamation, preventing the formation of comedones and aiding in the resolution of preexisting cysts. Topical antibacterial agents such as benzoyl peroxide, azelaic acid, topical erythromycin (with or without zinc), clindamycin, or tetracycline are also useful adjuncts to therapy.

Patients with moderate to severe acne with a prominent inflammatory component will benefit from the addition of systemic therapy. Oral tetracyclines or erythromycin in doses of 250 to 1000 mg/d will decrease follicular colonization with some of the lipophilic organisms. They also appear to have an anti-inflammatory effect independent of their antibacterial effect. Female patients who do not respond to oral antibiotics may benefit from hormonal therapy. Women placed on oral contraceptives containing ethinyl estradiol and norgestimate have demonstrated improvement in their acne when compared to a placebo control.

Severe nodulocystic acne not responsive to oral antibiotics, hormonal therapy, or topical therapy may be treated with the synthetic retinoid isotretinoin. It is used at doses of 0.5 to 2.0 mg/kg as a single daily dose for 15 to 20 weeks. The use of this drug is limited by its teratogenicity, and female patients must be screened for pregnancy prior to initiating therapy, maintain a method of birth control during therapy, and be screened for pregnancy during treatment. Patients receiving this medication develop extremely dry skin and cheilitis and must be followed for development of hypertriglyceridemia.

ACNE ROSACEA

Acne rosacea is an inflammatory disorder predominantly affecting the central face. It is seen almost exclusively in adults, only rarely affecting patients under 30 years of age. Rosacea is seen more often in women, but those most severely affected are men. It is characterized by the presence of erythema, telangiectases, and superficial pustules (see Plate IIA-2), but is not associated with the presence of comedones. Rosacea only rarely involves the chest or back.

There is a relationship between the tendency for pronounced facial flushing and the subsequent development of acne rosacea. Often, individuals with rosacea initially demonstrate a pronounced flushing reaction. This may be in response to heat, emotional stimuli, alcohol, hot drinks, or spicy foods. As the disease progresses, the flush persists longer and longer and may eventually become permanent. Papules, pustules, and telangiectases can become superimposed on the persistent flush. Rosacea of very long standing may lead to connective tissue overgrowth, particularly of the nose (rhinophyma). Rosacea may also be complicated by various inflammatory disorders of the eye, including keratitis, blepharitis, iritis, and recurrent chalazion. These ocular problems are potentially sight-threatening and warrant ophthalmologic evaluation.

TREATMENT

Acne rosacea can generally be treated effectively with oral tetracycline in doses ranging from 250 to 1000 mg/d. Topical metronidazole or sodium sulfacetamide has also been shown to be effective. In addition, the use of low-potency, nonfluorinated topical glucocorticoids, particularly after cool soaks, is helpful in alleviating facial erythema. Fluorinated topical glucocorticoids should be avoided since chronic use of these preparations may actually elicit rosacea. Topical therapy is not effective treatment for ocular disease.

(Bibliography omitted in Palm version)

Back to Table of Contents

57. SKIN MANIFESTATIONS OF INTERNAL DISEASE - *Jean L. Bolognia*, *Irwin M. Braverman*

It is now a generally accepted concept in medicine that the skin can show signs of internal disease. Therefore, in textbooks of medicine one finds a chapter describing in detail the major systemic disorders that can be identified by cutaneous signs. The underlying assumption of such a chapter is that the clinician has been able to identify the disorder in the patient and needs only to read about it in the textbook. In reality, concise differential diagnoses and the identification of these disorders are actually difficult for the nondermatologist because he or she is not well versed in the recognition of cutaneous lesions or their spectrum of presentations. Therefore, the authors of this chapter have decided to cover this particular topic of cutaneous medicine not by discussing individual disorders but by describing and discussing the various presenting clinical signs and symptoms that indicate the presence of these disorders. Concise differential diagnoses will be generated in which the significant diseases will be briefly discussed and distinguished from the more common disorders that have no significance for internal diseases. The latter disorders are reviewed in table form and always need to be excluded when considering the former. For a detailed description of individual diseases, the reader should consult a dermatologic text.

PAPULOSQUAMOUS SKIN LESIONS (Table 57-1)

When an eruption is characterized by elevated lesions, papules (<1 cm) or plaques (>1 cm), in association with scale, it is referred to as *papulosquamous*. The most common papulosquamous diseases -- *psoriasis*, *tinea*, *pityriasis rosea*, and *lichen planus* -- are primary cutaneous disorders (Chap. 56). When psoriatic lesions are accompanied by arthritis, the possibility of psoriatic arthritis or *Reiter's disease* should be considered. A history of oral ulcers, conjunctivitis, uveitis, and/or urethritis points to the latter diagnosis. In *guttate psoriasis* there is an acute onset of small, widely scattered, uniform lesions, often in association with a streptococcal infection. Lithium, beta blockers, HIV infection, and a rapid taper of systemic glucocorticoids are also known to exacerbate psoriasis.

Whenever the diagnosis of pityriasis rosea or lichen planus is made, it is important to review the patient's medications because the eruption can be treated by simply discontinuing the offending agent. Pityriasis rosea-like drug eruptions are seen most commonly with beta blockers, angiotensin-converting enzyme (ACE) inhibitors, gold, and metronidazole, while the drugs that can produce a lichenoid eruption include gold, antimalarials, thiazides, quinidine, phenothiazines, sulfonylureas, and ACE inhibitors. Lichen planus-like lesions are also observed in chronic graft-versus-host disease.

Parapsoriasis is an intermediate disease, for it can remain solely as a primary cutaneous disease or it can progress to cutaneous T cell lymphoma (CTCL) after a latency period of as long as 40 years. There are several forms of parapsoriasis, including small plaque (0.5 to 5 cm), large plaque (>6 cm), and retiform. The lesions of both small plaque and large plaque parapsoriasis are thin and salmon-pink in color with fine white scale. In small plaque forms, the lesions are commonly on the trunk but can be widely scattered. In large plaque forms, the most common location is the "girdle" area, and fine wrinkling secondary to epidermal atrophy is often seen. Retiform parapsoriasis forms a netlike pattern, and the individual papules are red-brown and flat-topped. The latter two forms of parapsoriasis, large plaque and retiform, can progress to CTCL.

A clue to the development of*CTCL*within lesions of large plaque or retiform parapsoriasis is an increase in the palpable component of the plaque (increased infiltration). In its early stages, CTCL may be confused with ezcema or psoriasis, but it often fails to respond to the appropriate therapy for those inflammatory diseases. The diagnosis of CTCL is established by skin biopsy in which collections of atypical T lymphocytes are found in the epidermis and dermis. As the disease progresses, cutaneous tumors and lymph node involvement may appear.

In *secondary syphilis* there are scattered red-brown papules with thin scale. The eruption often involves the palms and soles and can resemble pityriasis rosea. Associated findings are helpful in making the diagnosis and include annular plaques on the face, nonscarring alopecia, condyloma lata (broad-based and moist), and mucous patches as well as lymphadenopathy, malaise, fever, headache, and myalgias. The interval between the primary chancre and the secondary stage is usually 4 to 8 weeks, and spontaneous resolution without appropriate therapy is seen.

ERYTHRODERMA (Table 57-2)

Erythroderma is the term used when the majority of the skin surface is erythematous (red in color). There may be associated scale, erosions, or pustules as well as shedding of the hair and nails. Potential systemic manifestations include fever, chills, hypothermia, reactive lymphadenopathy, peripheral edema, hypoalbuminemia, and high-output cardiac failure. The major etiologies of erythroderma are (1) *cutaneous diseases* such as psoriasis and dermatitis (Table 57-3); (2) *drugs*; (3) *systemic diseases*, most commonlyCTCL; and (4) *idiopathic*. In the first three groups, the location and description of the initial lesions, prior to the development of the erythroderma, aid in the diagnosis. For example, a history of red scaly plaques on the elbows and knees would point to psoriasis. It is also important to examine the skin carefully for a migration of the erythema and associated secondary changes such as pustules or erosions. Migratory waves of erythema studded with superficial pustules are seen in *pustular psoriasis*.

Drug-induced erythroderma (exfoliative dermatitis) may begin as a morbilliform eruption (Chap. 59) or may arise as diffuse erythema. Fever and peripheral eosinophilia often accompany the eruption, and occasionally there is an associated allergic interstitial nephritis. A number of drugs can produce an erythroderma, including penicillins, sulfonamides, carbamazepine, phenytoin, gold, allopurinol, and captopril. While reactions to anticonvulsants can lead to a pseudolymphoma syndrome, reactions to allopurinol may be accompanied by hepatitis, gastrointestinal bleeding, and nephropathy.

The most common malignancy that is associated with erythroderma isCTCL; in some series, up to 25% of the cases of erythroderma were due to CTCL. The patient may progress from isolated plaques and tumors, but more commonly the erythroderma is present throughout the course of the disease (Sezary syndrome). In the Sezary syndrome, there are circulating atypical T lymphocytes, pruritus, and lymphadenopathy. In cases of erythroderma where there is no apparent cause (idiopathic), longitudinal follow-up is mandatory to monitor for the possible development of CTCL. There have been isolated case reports of erythroderma secondary to some solid tumors -- lung, liver, prostate, thyroid, and colon -- but it is usually in a late stage of the disease.

ALOPECIA (Table 57-4)

The two major forms of alopecia are scarring and nonscarring. In *scarring alopecia* there is associated fibrosis, inflammation, and loss of hair follicles. A smooth scalp with a decreased number of follicular openings is usually observed clinically, but in some cases the changes are seen only in biopsy specimens from the affected areas. In *nonscarring alopecia* the hair shafts are gone, but the hair follicles are preserved, explaining the reversible nature of nonscarring alopecia.

Primary cutaneous disorders are the most common causes of nonscarring alopecia and they include *telogen effluvium*, *androgenetic alopecia*, *alopecia areata*, *tinea capitis*, and *traumatic alopecia* (Table 57-5). In women with androgenetic alopecia, an elevation in circulating levels of androgens may be seen as a result of ovarian or adrenal gland dysfunction. When there are signs of virilization, such as a deepened voice and enlarged clitoris, the possibility of an ovarian or adrenal gland tumor should be considered.

Exposure to various *drugs* can also cause diffuse hair loss, usually by inducing a telogen effluvium. An exception is the anagen effluvium observed with antimitotic agents such as daunorubicin. Alopecia is a side effect of the following drugs: warfarin, heparin, propylthiouracil, carbimazole, vitamin A, isotretinoin, acetretin, lithium, beta blockers, colchicine, and amphetamines. Fortunately, spontaneous regrowth usually follows discontinuation of the offending agent.

Less commonly, nonscarring alopecia is associated with *lupus erythematosus* and *secondary syphilis.* In systemic lupus there are two forms of alopecia -- one is scarring secondary to discoid lesions (see below) and the other is nonscarring. The latter form may be diffuse and involve the entire scalp, or it may localized to the frontal scalp and result in multiple short hairs ("lupus hairs"). Scattered, poorly circumscribed patches of alopecia with a "moth-eaten" appearance are a manifestation of the secondary stage of syphilis. Diffuse thinning of the hair is also associated with hypothyroidism, hyperthyroidism, and HIV infection (Table 57-4).

Scarring alopecia is more frequently the result of a primary cutaneous disorder such as *lichen planus*, *folliculitis decalvans*, *cutaneous lupus*, or *linear scleroderma* (*morphea*) than it is a sign of systemic disease. Although the scarring lesions of *discoid lupus* can be seen in patients with systemic lupus, in the majority of cases the disease process is limited to the skin. Less common causes of scarring alopecia include *sarcoidosis* (see "Papulonodular Skin Lesions") and cutaneous *metastases*.

In the early phases of discoid lupus, lichen planus, and folliculitis decalvans, there are circumscribed areas of alopecia. Fibrosis and subsequent loss of follicles are observed primarily in the center of the individual lesions, while the inflammatory process is most prominent at the periphery. The areas of active inflammation in discoid lupus are erythematous with scale, whereas the areas of previous inflammation are often hypopigmented with a rim of hyperpigmentation. In lichen planus the peripheral perifollicular macules are usually violet-colored. Complete examination of the skin and oral mucosa combined with a biopsy and direct immunofluorescence microscopy will aid in distinguishing these two entities. The peripheral active lesions in folliculitis decalvans are perifollicular pustules; these patients can develop a reactive arthritis.

FIGURATE SKIN LESIONS (Table 57-6)

In *figurate* eruptions, the lesions form rings and arcs that are usually erythematous but can be flesh-colored to brown. Most commonly, they are due to primary cutaneous diseases such as *tinea*, *urticaria*, *erythema annulare centrifugum*, and *granuloma annulare* (Chaps. 56 and58). An underlying systemic illness is found in a second, less common group of migratory annular erythemas. It includes *erythema gyratum repens*, *erythema migrans*, *erythema marginatum*, and *necrolytic migratory erythema*.

In erythema gyratum repens, one sees hundreds of mobile concentric arcs and wavefronts that resemble the grain in wood. A search for an underlying malignancy is mandatory in a patient with this eruption. Erythema migrans is the cutaneous manifestation of Lyme disease, which is caused by the spirochete *Borrelia burgdorferi*. In the initial stage (3 to 30 days after tick bite), a single annular lesion is usually seen, which can expand to ^{3}10 cm in diameter. Within several days, approximately half the patients develop multiple smaller erythematous lesions at sites distant from the bite. Associated symptoms include fever, headache, photophobia, myalgias, arthralgias, and malar rash. Erythema marginatum is seen in patients with rheumatic fever, primarily on the trunk. Lesions are pink-red in color, flat to mildly elevated, and transient.

There are additional cutaneous diseases that present as annular eruptions but lack an obvious migratory component. Examples include*CTCL*, *annular cutaneous lupus*, also referred to as *subacute lupus*, *secondary syphilis*, and *sarcoidosis* (see "Papulonodular Skin Lesions").

ACNE (Table 57-7)

Acne vulgaris and *acne rosacea* are the two major forms of acne (Chap. 56). Estrogens decrease sebaceous gland activity, whereas androgens enhance sebum production. Therefore, acne vulgaris in an adult, especially if it is of recent onset, may be a reflection of increased levels of circulating *androgens*. Dysfunction of the ovary or adrenal gland, e.g., polycystic ovary disease or Cushing's syndrome, can lead to the hormonal imbalance. Examination of the patient for signs such as hirsutism, androgenetic alopecia, hypertension, and redistribution of subcutaneous fat will aid in the diagnosis.

Exacerbations of acne vulgaris follow the ingestion of several *drugs*, such as anabolic steroids, glucocorticoids, lithium, and iodides as well as the application of oil-containing compounds. Acne-like lesions can be seen in patients with Behcet's disease (see "Ulcers"), and in immunocompromised hosts, disseminated fungal infections (e.g., cryptococcosis) may present as an acneiform eruption.

Patients with the carcinoid syndrome have episodes of flushing of the head, neck, and sometimes the trunk. Resultant skin changes of the face, in particular telangiectasias, may mimic the clinical appearance of acne rosacea.

PUSTULAR LESIONS

Acneiform eruptions (see "Acne") and *folliculitis* represent the most common pustular dermatoses. An important consideration in the evaluation of perifollicular pustules is a determination of the associated pathogen, e.g., normal flora, *Staphylococcus aureus*, *Pityrosporum*. Noninfectious forms of folliculitis include HIV-associated eosinophilic folliculitis and folliculitis secondary to drugs such as glucocorticoids and lithium. Administration of high-dose oral glucocorticoids can result in a widespread eruption of perifollicular pustules on the trunk, characterized by lesions in the same stage of development. With regard to underlying systemic diseases, pustules are a characteristic component of pustular psoriasis and can be seen in septic emboli of bacterial or fungal origin (see "Purpura").

TELANGIECTASIAS (Table 57-8)

In order to distinguish the various types of telangiectasias, it is important to examine the shape and configuration of the dilated blood vessels. *Linear telangiectasias* are seen on the face of patients with *actinically damaged skin* and *acne rosacea* and they are found on the legs of patients with *venous hypertension* and *essential telangiectasia*. Patients with an unusual form of *mastocytosis* (telangiectasia macularis eruptiva perstans), the *carcinoid* syndrome (see "Acne"), and *ataxia-telangiectasia* also have linear telangiectasias. In ataxia-telangiectasia, linear telangiectasias appear on the bulbar conjunctiva during childhood. Eventually, there is involvement of the ears, eyelids, cheeks, and/or flexural areas such as the antecubital and popliteal fossae. Lastly, linear telangiectasias are found in areas of cutaneous inflammation. For example, lesions of discoid lupus frequently have telangiectasias within them.

Poikiloderma is a term used to describe a patch of skin with (1) reticulated hypo- and hyperpigmentation, (2) wrinkling secondary to epidermal atrophy, and (3) telangiectasias. Poikiloderma does not imply a single disease entity -- it is seen in skin damaged by *ionizing radiation*, in the disorders *poikiloderma vasculare atrophicans* (PVA) and *xeroderma pigmentosum*, as well as in patients with connective-tissue diseases, primarily *dermatomyositis* (DM).PVA is a precursor lesion ofCTCL, and the areas of poikiloderma usually begin in the flexural areas of the axillae and groin.

In *scleroderma*, the dilated blood vessels have a unique configuration and are known as *mat telangiectasias*. The lesions are broad macules that usually measure 2 to 7 mm in diameter but occasionally are larger. Mats have a polygonal or oval shape, and their erythematous color may be uniform or the result of delicate telangiectasias. The most common locations for mat telangiectasias are the face, oral mucosa, and hands -- peripheral sites that are prone to intermittent ischemia. The CREST (*c*alcinosis cutis, *R*aynaud's phenomenon, *e*sophageal dysmotility, *s*clerodactyly, and *t*elangiectasia) variant of scleroderma (Chap. 313) is associated with a chronic course and anticentromere antibodies. Mat telangiectasias are an important clue to the diagnosis of the CREST syndrome as well as systemic scleroderma, for they may be the only cutaneous finding.

Periungual telangiectasias are pathognomonic signs of the three major connective tissue diseases -- *lupus erythematosus*, *scleroderma*, and*DM*. They are easily visualized by the naked eye and occur in at least two-thirds of these patients. In both DM and lupus there is associated nailfold erythema, and in DM the erythema is often accompanied by "ragged" cuticles and fingertip tenderness. Under 10´ magnification, the blood vessels in the nailfolds of lupus patients are tortuous and resemble "glomeruli," whereas in scleroderma and DM there is a loss of capillary loops and those that remain are markedly dilated.

In *hereditary hemorrhagic telangiectasia* (Osler-Rendu-Weber disease), the lesions usually appear during adulthood and are most commonly seen on the mucous membranes, face, and distal extremities, including under the nails. They represent arteriovenous (AV) malformations of the dermal microvasculature, are dark red in color, and are usually slightly elevated. When the skin is stretched over an individual lesion, an eccentric punctum with radiating legs is seen. Although the degree of systemic involvement varies in this autosomal dominant disease (due to mutations in either the endoglin or activin receptor-like kinase gene), the major symptoms are recurrent epistaxis and gastrointestinal bleeding. The fact that these mucosal telangiectasias are actually AV communications helps to explain their tendency to bleed.

HYPOPIGMENTATION (Table 57-9)

Disorders of hypopigmentation are classified as either diffuse or localized. The classic example of *diffuse hypopigmentation* is *oculocutaneous albinism* (OCA). The most common forms are due to mutations in the tyrosinase gene (type I) or the *P* gene (type II); patients with type IA OCA have a total lack of enzyme activity. At birth, different forms of OCA can appear similar -- white hair, gray-blue eyes, and pink-white skin. However, the patients with no tyrosinase activity maintain this phenotype, whereas those with decreased activity or *P* gene mutations will acquire some pigmentation of the eyes, hair, and skin as they age. The degree of pigment formation is also a function of racial background, and the pigmentary dilution is readily apparent when patients are compared to their first-degree relatives.

The ocular findings inOCAcorrelate with the degree of hypopigmentation and include decreased visual acuity, nystagmus, photophobia, and monocular vision. Generalized vitiligo, phenylketonuria, and homocystinuria are other unusual causes of diffuse pigmentary dilution. In generalized vitiligo, melanocytes are not found in affected skin, whereas in OCA they are present but have decreased activity. Appropriate laboratory tests exclude the other disorders of metabolism.

The differential diagnosis of *localized hypomelanosis* includes the following primary cutaneous disorders: *idiopathic guttate hypomelanosis*, *postinflammatory hypopigmentation*, *tinea (pityriasis) versicolor*, *vitiligo*, *chemical leukoderma*, *nevus depigmentosus* (see below), and *piebaldism* (Table 57-9). In this group of diseases, the areas of involvement are macules or patches with a decrease or absence of pigmentation. Patients with vitiligo also have an increased incidence of several autoimmune disorders, including hypothyroidism, Graves' disease, pernicious anemia, Addison's disease, uveitis, alopecia areata, chronic mucocutaneous candidiasis, and the polyglandular autoimmune syndromes (types I and II). Diseases of the thyroid gland are the most frequently associated disorders, occurring in up to 30% of patients with vitiligo. Circulating autoantibodies are often found, and the most common ones are antithyroglobulin, antimicrosomal, and antiparietal cell antibodies.

There are three systemic diseases that should be considered in a patient with skin findings suggestive of vitiligo -- *Vogt-Koyanagi-Harada syndrome*, *scleroderma*, and *melanoma-associated leukoderma*. A history of aseptic meningitis, nontraumatic uveitis, tinnitus, hearing loss, and/or dysacousis points to the diagnosis of the Vogt-Koyanagi-Harada syndrome. In these patients, the face and scalp are the most common locations of pigment loss. The vitiligo-like leukoderma seen in patients with scleroderma has a clinical resemblance to idiopathic vitiligo that has begun to repigment as a result of treatment; that is, perifollicular macules of normal pigmentation are seen within areas of depigmentation. The basis of this leukoderma is unknown; there is no evidence of inflammation in areas of involvement, but it can resolve if the underlying connective tissue disease becomes inactive. In contrast to idiopathic vitiligo, melanoma-associated leukoderma often begins on the trunk, and its appearance should prompt a search for metastatic disease. The possibility exists that the destruction of normal melanocytes is the result of an immune response against malignant melanocytes.

There are two systemic disorders that may have the cutaneous findings of piebaldism (Table 57-10). They are *Hirschsprung's disease* and *Waardenburg's syndrome*. A possible explanation for both disorders is an abnormal embryonic migration or survival of two neural crest-derived elements, one of them being melanocytes and the other myenteric ganglion cells (Hirschsprung's disease) or auditory nerve cells (Waardenburg's syndrome). The latter syndrome is characterized by congenital sensorineural hearing loss, dystopia canthorum (lateral displacement of the inner canthi but normal interpupillary distance), heterochromic irises, and a broad nasal root, in addition to the piebaldism. Patients with Waardenburg's syndrome have been shown to have mutations in two genes that encode DNA-binding proteins, *PAX-3* and *MITF*, while patients with Hirschsprung's disease and white spotting have mutations in one of three genes -- endothelin 3, endothelin B receptor, and *SOX-10*.

In *tuberous sclerosis*, the earliest cutaneous sign is an ash leaf spot. These lesions are often present at birth and are usually multiple; however, detection may require Wood's lamp examination, especially in fair-skinned individuals. The pigment within them is reduced but not absent. The average size is 1 to 3 cm, and the common shapes are polygonal and lance-ovate. Examination of the patient for additional cutaneous signs such as adenoma sebaceum (multiple angiofibromas of the face), ungual and gingival fibromas, fibrous plaques of the forehead, and connective tissue nevi (shagreen patches) is recommended. It is important to remember that an ash leaf spot on the scalp will result in *poliosis*, which is a circumscribed patch of gray-white hair. Internal manifestations include seizures, mental retardation, central nervous system (CNS) and retinal hamartomas, renal angiomyolipomas, and cardiac rhabdomyomas. The latter can be detected in up to 60% of children (<18 years) with tuberous sclerosis by echocardiography.

Nevus depigmentosus is a stable, well-circumscribed hypomelanosis that is present at birth. There is usually a single circular or rectangular lesion, but occasionally the nevus has a dermatomal or whorled pattern. It is important to distinguish this more common entity from ash leaf spots especially when there are multiple lesions. In *hypomelanosis of Ito*, swirls and streaks of hypopigmentation run parallel to one another in a pattern that resembles a marble cake. Lesions may progress or regress with time, and in up to a third of patients, associated abnormalities are found including in the musculoskeletal system (asymmetry), theCNS(seizures and mental retardation), and the eyes (strabismus and hypertelorism). Chromosomal mosaicism has been detected in these patients; this lends support to the hypothesis that the pattern is the result of the migration of two clones of primordial melanocytes, each with a different pigment potential.

Localized areas of decreased pigmentation are commonly seen as a result of cutaneous inflammation (Table 57-10) and have been observed in the skin overlying active lesions of sarcoidosis (see "Papulonodular Skin Lesions") as well as inCTCL. Cutaneous infections also present as disorders of hypopigmentation, and in *tuberculoid leprosy* there are a few asymmetric patches of hypomelanosis that have associated anesthesia, anhidrosis, and alopecia. Biopsy specimens of the palpable border show dermal granulomas that lack *Mycobacterium leprae* organisms.

HYPERPIGMENTATION (Table 57-11)

Disorders of hyperpigmentation are also divided into two groups -- localized and diffuse. The *localized* forms are due to an epidermal alteration, a proliferation of melanocytes, or an increase in pigment production. Both seborrheic keratoses and acanthosis nigricans belong to the first group. *Seborrheic keratoses* are common lesions, but in one clinical setting they are a sign of systemic disease, and that setting is the sudden appearance of multiple lesions, often with an inflammatory base and in association with acrochordons (skin tags) and acanthosis nigricans. This is termed the *sign of Leser-Trelat* and signifies an internal malignancy. *Acanthosis nigricans* can also be a reflection of an internal malignancy, most commonly of the gastrointestinal tract, and it appears as velvety hyperpigmentation, primarily in flexural areas. In the majority of patients, acanthosis nigricans is associated with obesity, but it may be a reflection of an endocrinopathy such as acromegaly, Cushing's syndrome, the Stein-Leventhal syndrome, or insulin-resistant diabetes mellitus (type A, type B, and lipoatrophic forms).

A proliferation of melanocytes results in the following pigmented lesions: *lentigo*, *melanocytic nevus*, and *melanoma* (Chap. 86). In an adult, the majority of lentigines are related to sun exposure, which explains their distribution. However, in the Peutz-Jeghers and LEOPARD [*l*entigines; *E*CG abnormalities, primarily conduction defects, *o*cular hypertelorism; *p*ulmonary stenosis and subaortic valvular stenosis; *a*bnormal genitalia (cryptorchidism, hypospadias); *r*etardation of growth; and *d*eafness (sensorineural)] syndromes, lentigines do serve as a clue to systemic disease. In the multiple lentigines or *LEOPARD syndrome*, hundreds of lentigines develop during childhood and are scattered over the entire surface of the body. The lentigines in patients with *Peutz-Jeghers syndrome* are located primarily around the nose and mouth, on the hands and feet, and within the oral cavity. While the pigmented macules on the face may fade with age, the oral lesions persist. However, similar intraoral lesions are also seen in Addison's disease and as a normal finding in darkly pigmented individuals. Patients with this autosomal dominant syndrome (due to mutations in a novel serine threonine kinase gene) have multiple benign polyps of the gastrointestinal tract, testicular tumors, and an increased risk of developing gastrointestinal (primarily colon), breast, and gynecologic cancers.

Lentigines are also seen in association with cardiac myxomas and have been described in two syndromes whose findings overlap: *LAMB* (*l*entigines, *a*trial myxomas, *m*ucocutaneous myxomas, and *b*lue nevi) *syndrome* and *NAME* [*n*evus, *a*trial myxoma, *m*yxoid neurofibroma, and *e*phelides (freckles)] *syndrome*. These patients can also have evidence of endocrine overactivity in the form of Cushing's syndrome, acromegaly, or sexual precocity.

The third type of localized hyperpigmentation is due to a local increase in pigment production, and it includes *ephelides* and cafe au lait macules (CALM). The latter are most commonly associated with two disorders -- neurofibromatosis (NF) and McCune-Albright syndrome. *CALM* are flat, uniformly light brown in color, and can vary in size from 0.5 to 12 cm. Approximately 80% of adult patients with *type I NF* will have six or more CALM measuring 1.5 cm or greater in diameter. Additional findings are discussed in the section on neurofibromas (see "Papulonodular Skin Lesions"). In comparison with NF, the CALM in patients with *McCune-Albright syndrome* [polyostotic fibrous dysplasia with precocious puberty in females due to mosaicism for an activating mutation in a G protein (G_sa) gene] are usually larger, more irregular in outline, and tend to respect the midline. CALM have also been associated with pulmonary stenosis (Watson syndrome), tuberous sclerosis, theLEOPARDsyndrome, and ataxia telangiectasia, but a few such lesions can be found in normal individuals.

In incontinentia pigmenti, dyskeratosis congenita, and bleomycin pigmentation, the areas of localized hyperpigmentation form a pattern -- swirled in the first, reticulated in the second, and flagellate in the third. Patients with the X-linked dominant disorder *incontinentia pigmenti* can have linear blisters and verrucous papules during infancy. During childhood, parallel swirls and streaks of hyperpigmentation appear on the trunk, and occasionally streaks of hypopigmentation appear on the extremities. Associated findings include seizures, mental retardation, retinal vascular abnormalities, and delayed or impaired dentition. Biopsy specimens of the streaks will show pigment within dermal macrophages ("incontinent pigment"). In *dyskeratosis congenita*, atrophic reticulated hyperpigmentation is seen on the neck, thighs, and trunk and is accompanied by nail dystrophy, pancytopenia, and leukoplakia of the oral and anal mucosa. The latter often develops into squamous cell carcinoma. In addition to the flagellate pigmentation (linear streaks) on the trunk, patients receiving bleomycin often have hyperpigmentation on the elbows, knees, and small joints of the hand.

Localized hyperpigmentation is seen as a side effect of several other *systemic medications*, including those that produce fixed drug reactions [phenolphthalein, nonsteroidal anti-inflammatory drugs (NSAIDs), sulfonamides, and barbiturates] and those that can complex with melanin (antimalarials). Fixed drug eruptions recur in the same location as circular areas of erythema that can become bullous and then resolve as brown macules. The eruption usually appears within hours of administration of the offending agent, and common locations include the genitalia, extremities, and perioral region. Chloroquine and hydroxychloroquine produce gray-brown to blue-black discoloration of the shins, hard palate, and face, while blue macules can be seen on the lower extremities and in sites of inflammation with prolonged minocycline administration. Estrogen in oral contraceptives can induce melasma -- symmetric brown patches on the face, especially the cheeks, upper lip, and forehead. Similar changes are seen in pregnancy, in patients receiving hydantoin, and in the adult form of Gaucher's disease. In the latter group there is also hyperpigmentation of the distal lower extremities.

In the *diffuse* forms of hyperpigmentation, the darkening of the skin may be of equal intensity over the entire body or may be accentuated in sun-exposed areas. The causes of diffuse hyperpigmentation can be divided into four groups -- endocrine, metabolic, autoimmune, and drugs. The endocrinopathies that frequently have associated hyperpigmentation include *Addison's disease*, *Nelson syndrome*, and *ectopic ACTH syndrome*. In these diseases, the increased pigmentation is diffuse but is accentuated in the palmar creases, sites of friction, scars, and the oral mucosa. An overproduction of the pituitary hormonesa-MSH (melanocyte-stimulating hormone) and ACTH can lead to an increase in melanocyte activity. These peptides are products of the proopiomelanocortin gene and exhibit homology; e.g.,a-MSHand ACTH share 13 amino acids. A minority of the patients with Cushing's disease or hyperthyroidism have generalized hyperpigmentation.

The metabolic causes of hyperpigmentation include *porphyria cutanea tarda* (PCT), *hemochromatosis*, *vitamin B_{12}deficiency*, *folic acid deficiency*, *pellagra*, *malabsorption*, and *Whipple's disease*. In patients with PCT (see "Vesicles/Bullae"), the skin darkening is seen in sun-exposed areas and is a reflection of the photoreactive properties of porphyrins. The increased level of iron in the skin of patients with hemochromatosis stimulates melanin pigment production and leads to the classic bronze color. Patients with pellagra have a brown discoloration of the skin, especially in sun-exposed areas, as a result of nicotinic acid (niacin) deficiency. In the areas of increased pigmentation, there is a thin varnish-like scale. These changes are also seen in patients who are vitamin B_6deficient, have functioning carcinoid tumors (increased consumption of niacin), or take isoniazid. Approximately 50% of the patients with Whipple's disease have an associated generalized hyperpigmentation in association with diarrhea, weight loss, arthritis, and lymphadenopathy. A diffuse slate-blue color is seen in patients with melanosis secondary to metastatic melanoma. Although there is a debate as to whether the color is due to single-cell metastases in the dermis or to a widespread deposition of melanin resulting from the high concentration of circulating melanin precursors, there is more evidence to support the latter.

Of the autoimmune diseases associated with diffuse hyperpigmentation, *biliary cirrhosis* and *scleroderma* are the most common, and occasionally, both disorders are seen in the same patient. The skin is dark brown in color, especially in sun-exposed areas. In biliary cirrhosis the hyperpigmentation is accompanied by pruritus, jaundice, and xanthomas, whereas in scleroderma it is accompanied by sclerosis of the extremities, face, and, less commonly, the trunk. Additional clues to the diagnosis of scleroderma are telangiectasias, calcinosis cutis, Raynaud's phenomenon, and distal ulcerations (see "Telangiectasias"). The differential diagnosis of cutaneous sclerosis with hyperpigmentation includes the POEMS [*p*olyneuropathy; *o*rganomegaly (liver, spleen, lymph nodes); *e*ndocrinopathies (impotence, gynecomastia); *M*-protein; and *s*kin changes] syndrome. The skin changes include hyperpigmentation, skin thickening, hypertrichosis, and angiomas.

In the late 1980s, an epidemic of the eosinophilia-myalgia syndrome was described that was presumably due to contaminated L-tryptophan preparations. In addition to maculopapular eruptions and alopecia, large areas of scleroderma-like induration were observed with overlying hyperpigmentation.

Diffuse hyperpigmentation that is due to *drugs* or *metals* can result from one of several mechanisms -- induction of melanin pigment formation, complexing of the drug or its metabolites to melanin, and deposits of the drug in the dermis. Busulfan, cyclophosphamide, long-term, high-dose ACTH, and inorganic arsenic induce pigment production. Complexes containing melanin or hemosiderin plus the drug or its metabolites are seen in patients receiving chlorpromazine and minocycline. The sun-exposed skin as well as the conjunctivae of patients on long-term, high-dose chlorpromazine can become blue-gray in color. Patients taking minocycline may develop a diffuse blue-gray, muddy appearance in sun-exposed areas in addition to pigmentation of the mucous membranes, teeth, nails, bones, and thyroid. Administration of amiodarone can result in both a phototoxic eruption (exaggerated sunburn) and/or a brown or blue-gray discoloration of sun-exposed skin. Biopsy specimens of the latter show yellow-brown granules in dermal macrophages, which represent intralysosomal accumulations of lipids, amiodarone, and its metabolites. Actual deposits of a particular drug or metal in the skin are seen with silver (argyria), where the skin appears blue-gray in color; gold (chrysiasis), where the skin has a brown to blue-gray color; and clofazimine, where the skin appears reddish brown. The associated hyperpigmentation is accentuated in sun-exposed areas, and discoloration of the eye is seen with gold (sclerae) and clofazimine (conjunctivae).

VESICLES/BULLAE (Table 57-12)

Depending on their size, cutaneous blisters are referred to as *vesicles* (<0.5 cm) or *bullae* (>0.5 cm). The primary blistering disorders include *pemphigus vulgaris*, *pemphigus foliaceus*, *pemphigus erythematosus*, *paraneoplastic pemphigus*, *bullous pemphigoid*, *herpes gestationis*, *cicatricial pemphigoid*, *epidermolysis bullosa acquisita*, *linear IgA disease*, and *dermatitis herpetiformis* (Chap. 58).

Vesicles and bullae are also seen in *contact dermatitis*, both allergic and irritant forms (Chap. 56). When there is a linear arrangement of vesicular lesions, an exogenous cause should be suspected. Bullous disease secondary to the ingestion of drugs can take one of several forms, including phototoxic eruptions, isolated bullae, toxic epidermal necrolysis, and erythema multiforme (Chap. 59). Clinically, phototoxic eruptions resemble an exaggerated sunburn with diffuse erythema and bullae in sun-exposed areas. The most commonly associated drugs are thiazides, doxycycline, sulfonamides,NSAIDs, and psoralens. The development of a phototoxic eruption is dependent on the doses of both the drug and UV-A irradiation.

Toxic epidermal necrolysis (TEN) is characterized by bullae that arise on widespread areas of erythema and then slough. This results in large areas of denuded skin. The associated morbidity, such as sepsis, and mortality are relatively high and are a function of the extent of epidermal necrosis. In addition, these patients may also have involvement of the mucous membranes and intestinal tract. Drugs are the primary cause of TEN, and the most common offenders are phenytoin, barbiturates, sulfonamides, penicillins, andNSAIDs. Severe acute graft-versus-host disease (grade 4) also can resemble TEN.

In *erythema multiforme* (EM), the primary lesions are pink-red macules and edematous papules, the centers of which may become vesicular. The clue to the diagnosis of EM, as opposed to a drug-induced morbilliform exanthem, is the development of a "dusky" violet color or petechiae in the center of the lesions. Target or iris lesions are also characteristic of EM and arise as a result of active centers and borders in combination with centrifugal spread. However, iris lesions need not be present to make the diagnosis of EM. Preferred sites of involvement include the distal extremities and mucous membranes (oral, nasal, ocular, and genital). Hemorrhagic crusts of the lips are characteristic of EM as well as herpes simplex, pemphigus vulgaris, and paraneoplastic pemphigus. Fever, malaise, myalgias, sore throat, and cough may precede or accompany the eruption. The lesions of EM usually resolve over 3 to 6 weeks but may be recurrent.

Drugs can induceEM, in particular sulfonamides, phenytoin, barbiturates, penicillins, and carbamazepine, but they do not cause the majority of cases, especially in young adults. Infections with herpes simplex are the most common cause of EM in this age group, and the lesions appear 7 to 12 days after the viral eruption. Other infectious agents associated with EM include *Mycoplasma pneumoniae*, dimorphic fungi, and several viruses (echovirus, coxsackievirus, Epstein-Barr, and influenza). EM can also follow vaccinations with BCG, poliomyelitis, or vaccinia viruses; radiation therapy; and exposure to environmental toxins.

In addition to primary blistering disorders and hypersensitivity reactions, bacterial and viral infections can lead to vesicles and bullae. The most common infectious agents are herpes simplex (Chap. 182), herpes varicella-zoster (Chap. 183), and staphylococci (Chap. 139).

Staphylococcal scalded-skin syndrome (SSSS) and *bullous impetigo* are two blistering disorders associated with staphylococcal (phage group II) infection. In SSSS, the initial findings are redness and tenderness of the central face, neck, trunk, and intertriginous zones. This is followed by short-lived flaccid bullae and a slough or exfoliation of the superficial epidermis. Crusted areas then develop, characteristically around the mouth. SSSS is distinguished fromTEN by the following features: younger age group, more superficial site of blister formation, no oral lesions, shorter course, less morbidity and mortality, and an association with staphylococcal exfoliative toxin ("exfoliatin"), not drugs. A rapid diagnosis of SSSS versus TEN can be made by a frozen section of the blister roof or exfoliative cytology of the blister contents. In SSSS the site of staphylococcal infection is usually extracutaneous (conjunctivitis, rhinorrhea, otitis media, pharyngitis, tonsillitis), and the cutaneous lesions are sterile, whereas in bullous impetigo the skin lesions are the site of infection. Impetigo is more localized than SSSS and usually presents with honey-colored crusts. Occasionally, superficial purulent blisters also form. *Cutaneous emboli* from gram-negative infections may present as isolated bullae, but the base of the lesion is purpuric or necrotic, and it may develop into an ulcer (see "Purpura").

Several metabolic disorders are associated with blister formation, including diabetes mellitus, renal failure, and porphyria. Local hypoxia secondary to decreased cutaneous blood flow can also produce blisters, which explains the presence of bullae over pressure points in comatose patients (coma bullae). In *diabetes mellitus*, tense bullae with clear viscous fluid arise on normal skin. The lesions can be as large as 6 cm in diameter and are located on the distal extremities. There are several types of porphyria, but the most common form with cutaneous findings is*PCT*. In sun-exposed areas (primarily the face and hands), the skin is very fragile, and trauma leads to erosions and tense vesicles. These lesions then heal with scarring and formation of milia; the latter are firm, 2- to 3-mm white or yellow papules that represent epidermoid inclusion cysts. Associated findings can include hypertrichosis of the lateral malar region (males) or face (females) and, in sun-exposed areas, hyperpigmentation and firm sclerotic plaques. An elevated level of urinary uroporphyrins confirms the diagnosis and is due to a decrease in uroporphyrinogen decarboxylase activity. Precipitating agents include alcohol, iron, chlorinated hydrocarbons, and hepatitis C infection.

The differential diagnosis ofPCTincludes (1) *porphyria variegata* -- the skin signs of PCT plus the systemic findings of acute intermittent porphyria; it has a diagnostic plasma porphyrin fluorescence emission at 626 nm; (2) *drug-induced bullous photosensitivity* (pseudoporphyria) -- the clinical and histologic findings are similar to PCT, but porphyrins are normal; etiologic agents include naproxen, furosemide, tetracycline, and nalidixic acid; (3) *bullous dermatosis of hemodialysis* -- the same appearance as PCT, but porphyrins are usually normal or occasionally borderline elevated; patients have chronic renal failure and are on hemodialysis; (4) PCT associated with hepatomas, hepatic carcinomas, and hemodialysis; and (5) *epidermolysis bullosa acquisita* (Chap. 58).

EXANTHEMS (Table 57-13)

Exanthems are characterized by an acute generalized eruption. The two most common presentations are erythematous macules and papules (morbilliform) and confluent blanching erythema (scarlatiniform). *Morbilliform* eruptions are usually due to either *drugs* or *viral infections*. For example, up to 5% of the patients receiving penicillins, sulfonamides, phenytoin, or gold will develop a maculopapular eruption. Accompanying signs may include pruritus, fever, eosinophilia, and transient lymphadenopathy. Similar maculopapular eruptions are seen in the classic childhood viral exanthems, including (1) *rubeola* (measles) -- a prodrome of coryza, cough, and conjunctivitis followed by Koplik's spots on the buccal mucosa; the eruption begins behind the ears, at the hairline, and on the forehead and then spreads down the body, often becoming confluent; (2) *rubella* -- it begins on the forehead and face and then spreads down the body; it resolves in the same order and is associated with retroauricular and suboccipital lymphadenopathy; and (3) *erythema infectiosum* (fifth disease) -- erythema of the cheeks is followed by a reticulated pattern on extremities; it is secondary to a parvovirus B19 infection, and an associated arthritis is seen in adults.

Both measles and rubella are seen in unvaccinated young adults, and an atypical form of measles is seen in adults immunized with either killed measles vaccine or killed vaccine followed in time by live vaccine. In contrast to classic measles, the eruption of atypical measles begins on the palms, soles, wrists, and knuckles, and the lesions may become purpuric. The patient with atypical measles can have pulmonary involvement and be quite ill. Rubelliform and roseoliform eruptions are also associated with *Epstein-Barr virus* (5 to 15% of patients), *echovirus*, *coxsackievirus*, and *adenovirus* infections. Detection of specific IgM antibodies or fourfold elevations in IgG antibodies allows the proper diagnosis. Occasionally, a maculopapular eruption is the result of a drug-viral interaction. For example, about 95% of the patients with infectious mononucleosis who are given ampicillin will develop a rash.

Of note, early in the course of infections with *Rickettsia* and *meningococcus*, prior to the development of purpura, the lesions may be erythematous macules and papules. This is also the case in chickenpox prior to the development of vesicles. Maculopapular eruptions are associated with early *HIV infection*, early secondary *syphilis*, *typhoid fever*, and *acute graft-versus-host disease*. In the last, lesions frequently begin on the palms and soles; the macular rose spots of typhoid fever involve primarily the anterior trunk.

The prototypic *scarlatiniform* eruption is seen in *scarlet fever* and is due to an erythrotoxin produced by group A b-hemolytic streptococcal infections, most commonly pharyngitis. This eruption is characterized by diffuse erythema, which begins on the neck and upper trunk, and red perifollicular puncta. Additional findings include a white strawberry tongue (white coating with red papillae) followed by a red strawberry tongue (red tongue with red papillae); petechiae of the palate; a facial flush with circumoral pallor; linear petechiae in the antecubital fossae; and desquamation of the involved skin, palms, and soles 5 to 20 days after onset of the eruption. A similar desquamation of the palms and soles is seen in toxic shock syndrome, Kawasaki's disease, and after severe febrile illnesses. Certain strains of staphylococci also produce an erythrotoxin that leads to the same clinical findings as in streptococcal scarlet fever, except that the antistreptolysin O titers are not elevated.

In *toxic shock syndrome* (TSS), staphylococcal (phage group I) infections produce an exotoxin (TSST-1) that causes the fever and rash, as well as enterotoxins. Initially, the majority of cases were reported in menstruating women who were using tampons. However, other sites of infection, including wounds and vaginitis, may produce TSS. The diagnosis of TSS is based on clinical criteria (Chap. 139), and three of these involve mucocutaneous sites. The clinical criteria are (1) fever; (2) diffuse erythema of the skin; (3) desquamation of the palms and soles 1 to 2 weeks after onset of illness; (4) hypotension; and (5) involvement of three or more organ systems, including the gastrointestinal tract, muscles, kidney, liver,CNS, hematologic (thrombocytopenia), and mucous membranes. The latter is characterized as hyperemia of the vagina, oropharynx, or conjunctivae. Similar systemic findings have been described in *streptococcal toxic shock-like syndrome* (Chap. 140), and although an exanthem is seen less often than in TSS due to a staphylococcal infection, the underlying infection is often in the soft tissue.

The cutaneous eruption in *Kawasaki's disease* (mucocutaneous lymph node syndrome) (Chap. 317) is polymorphous, but the two most common forms are morbilliform and scarlatiniform. The majority of cases are seen in children less than 5 years of age, but adult cases have been reported. The diagnosis is based on a fever lasting more than 5 days plus four of the five following criteria: (1) bilateral conjunctival injection; (2) exanthem; (3) cervical lymphadenopathy, usually unilateral; (4) erythema and edema of the hands and feet followed by desquamation; and (5) diffuse erythema of the oropharynx, red strawberry tongue, and erosions with crusting on the lips. This clinical picture can resembleTSS and scarlet fever, but clues to the diagnosis of Kawasaki's disease are the cervical lymphadenopathy, lip erosions, and increased platelets. The most serious associated systemic finding in this disease is coronary aneurysm secondary to arteritis. Aneurysms may lead to sudden death, primarily within the first 30 days of the illness. Scarlatiniform eruptions are also seen in the early phase of SSSS(see "Vesicles/Bullae") and as reactions to drugs.

URTICARIA (Table 57-14)

Urticaria (hives) are transient lesions that are composed of a central wheal surrounded by an erythematous halo. Individual lesions are round, oval, or figurate and are often pruritic. *Acute* and *chronic* urticaria have a wide variety of allergic etiologies. Less common systemic causes of urticaria are mastocytosis (urticaria pigmentosa), hyperthyroidism, malignancy, and juvenile rheumatoid arthritis (JRA). In JRA, the lesions coincide with the fever spike and are transient but not migratory as in erythema marginatum.

The common *physical urticarias* include dermographism, solar urticaria, cold urticaria, and cholinergic urticaria. Patients with *dermographism* exhibit linear wheals following minor pressure or scratching of the skin. It is a common disorder, affecting approximately 5% of the population. *Solar urticaria* characteristically occur within minutes of sun exposure and are a skin sign of one systemic disease -- erythropoietic protoporphyria. In addition to the urticaria, these patients have subtle pitted scarring of the nose and hands. *Cold urticaria* are precipitated by exposure to the cold, and therefore exposed areas are usually affected. In some cases, the disease is associated with abnormal circulating proteins -- more commonly cryoglobulins and less commonly cryofibrinogens and cold agglutinins. Additional systemic symptoms include wheezing and syncope, thus explaining the need for these patients to avoid swimming in cold water. *Cholinergic urticaria* are precipitated by heat, exercise, or emotion and are characterized by small wheals with relatively large flares. They are occasionally associated with wheezing.

Whereas urticaria are the result of dermal edema, subcutaneous edema leads to the clinical picture of *angioedema.* Sites of involvement include the eyelids, lips, tongue, larynx, and gastrointestinal tract as well as the subcutaneous tissue. Angioedema occurs alone or in combination with urticaria, including urticarial vasculitis and the physical urticarias. Both acquired and hereditary (autosomal dominant) forms of angioedema occur (Chap. 310), and in the latter, urticaria is rarely seen.

Urticarial vasculitis is an immune complex disease that may be confused with simple urticaria. In contrast to simple urticaria, individual lesions tend to last longer than 24 h and usually develop central petechiae that can be observed even after the urticarial phase has resolved. The patient may also complain of burning rather than pruritus. On biopsy, there is a leukocytoclastic vasculitis of the small blood vessels. Although many cases of urticarial vasculitis are idiopathic in origin, it can be a reflection of an underlying systemic illness such as lupus erythematosus, Sjogren's syndrome, or hereditary complement deficiency. There is a spectrum of urticarial vasculitis that ranges from purely cutaneous to multisystem involvement. The most common systemic signs and symptoms are arthralgias and/or arthritis, nephritis, and crampy abdominal pain, with asthma and chronic obstructive lung disease seen less often. Hypocomplementemia occurs in one- to two-thirds of patients, even in the idiopathic cases. Urticarial vasculitis can also be seen in patients with *hepatitis B* and *hepatitis C* infections, *serum sickness*, and *serum sickness-like illnesses*.

PAPULONODULAR SKIN LESIONS (Table 57-15)

In the *papulonodular diseases*, the lesions are elevated above the surface of the skin and may coalesce to form plaques. The location, consistency, and color of the lesions are the keys to their diagnosis; this section is organized on the basis of color.

White Lesions In *calcinosis cutis* there are firm white to white-yellow papules with an irregular surface. When the contents are discharged, a chalky white material is seen. *Dystrophic* calcification is seen at sites of previous inflammation or damage to the skin. It develops in acne scars as well as on the distal extremities of patients with scleroderma and in the subcutaneous tissue and intermuscular fascial planes inDM. The latter is more extensive and is more commonly seen in children. An elevated calcium phosphate product, as in secondary hyperparathyroidism, can lead to nodules of *metastatic* calcinosis cutis, which tend to be subcutaneous and periarticular. This form is often accompanied by calcification of muscular arteries and subsequent ischemic necrosis (calciphylaxis).

Skin-Colored Lesions There are several types of skin-colored lesions, including epidermoid inclusion cysts, lipomas, rheumatoid nodules, neurofibromas, angiofibromas, neuromas, and adnexal tumors such as tricholemmomas. Both *epidermoid inclusion cysts* and *lipomas* are very common mobile subcutaneous nodules -- the former are rubbery and compressible and drain cheeselike material (sebum and keratin) if incised. Lipomas are firm and somewhat lobulated on palpation. When extensive facial epidermoid inclusion cysts develop in childhood or there is a family history of such lesions, the patient should be examined for other signs of Gardner syndrome, including osteomas and desmoid tumors. *Rheumatoid nodules* are firm, 0.5- to 4-cm nodules that tend to localize around pressure points, especially the elbows. They are seen in approximately 20% of patients with rheumatoid arthritis and 6% of patients with Still's disease. Biopsies of the nodules show palisading granulomas. Similar lesions that are smaller and shorter-lived are seen in rheumatic fever.

Neurofibromas (benign Schwann cell tumors) are soft papules or nodules that exhibit the "button-hole" sign, that is, they invaginate into the skin with pressure in a manner similar to a hernia. Single lesions are seen in normal individuals, but multiple neurofibromas, usually in combination with six or moreCALMmeasuring >1.5 cm (see "Hyperpigmentation") and multiple Lisch nodules, are seen in von Recklinghausen's disease (NF type I). Lisch nodules are 1-mm yellow-brown spots within the iris that are best observed with slit-lamp examination. Additional manifestations include axillary freckling and peripheral andCNStumors (Chap. 370). In some patients the neurofibromas are localized and unilateral, whereas in others they are limited to the CNS.

Angiofibromas are firm, pink to skin-colored papules that measure from 3 mm to several centimeters in diameter. When they are located on the central cheeks (adenoma sebaceum) or multiple fibromas are seen around the nails, the patient has tuberous sclerosis. It is an autosomal disorder due to mutations in two different genes, and the associated findings are discussed in the section on ash leaf spots as well as inChap. 370. Multiple facial angiofibromas have also been observed in patients with multiple endocrine neoplasia (MEN) syndrome, type 1.

Neuromas (benign proliferations of nerve fibers) are also firm, skin-colored papules. They are more commonly found at sites of amputation and as rudimentary supernumerary digits. However, when there are multiple neuromas on the eyelids, lips, distal tongue, and/or oral mucosa, the patient should be investigated for other signs of theMENsyndrome, type 2b. Associated findings include marfanoid habitus, protuberant lips, intestinal ganglioneuromas, and medullary thyroid carcinoma (>75% of patients;Chap. 339).

Adnexal tumors are derived from pluripotential cells of the epidermis that can differentiate toward hair, sebaceous, apocrine, or eccrine glands or remain undifferentiated. *Basal cell epitheliomas* (BCEs) are examples of adnexal tumors that have little or no evidence of differentiation. Clinically, they are translucent papules with rolled borders, telangiectasias, and central erosion. BCEs commonly arise in sun-damaged skin of the head and neck. When a patient has multiple BCEs, especially prior to age 30, the possibility of the basal cell nevus syndrome should be raised. It is inherited as an autosomal dominant trait and is associated with jaw cysts, palmar and plantar pits, frontal bossing, medulloblastomas and calcification of the falx cerebri and diaphragma sellae. *Tricholemmomas* are also skin-colored adnexal tumors but differentiate toward hair follicles and can have a wartlike appearance. The presence of multiple tricholemmomas on the face and cobblestoning of the oral mucosa points to the diagnosis of Cowden's disease (multiple hamartoma syndrome) due to mutations in the *PTEN* gene. Internal organ involvement (in decreasing order of frequency) includes fibrocystic disease and carcinoma of the breast, adenomas and carcinomas of the thyroid, and gastrointestinal polyposis. Keratoses of the palms, soles, and dorsa of the hands are also seen.

Pink Lesions The cutaneous lesions associated with primary systemic *amyloidosis* are pink in color and translucent. Common locations are the face, especially the periorbital and perioral regions, and flexural areas. On biopsy, homogeneous deposits of amyloid are seen in the dermis and in the walls of blood vessels; the latter lead to an increase in vessel wall fragility. As a result, petechiae and purpura develop in clinically normal skin as well as in lesional skin following minor trauma, hence the term "pinch purpura." Amyloid deposits are also seen in the striated muscle of the tongue and result in macroglossia.

Even though specific mucocutaneous lesions are rarely seen in secondary amyloidosis and are present in only about 30% of the patients with primary amyloidosis, a rapid diagnosis of systemic amyloidosis can be made by an examination of abdominal subcutaneous fat. By special staining, deposits are seen around blood vessels or individual fat cells in 40 to 50% of patients. There are also three forms of amyloidosis that are limited to the skin and that should not be construed as cutaneous lesions of systemic amyloidosis. They are macular amyloidosis (upper back), lichenoid amyloidosis (usually lower extremities), and nodular amyloidosis. In macular and lichenoid amyloidosis, the deposits are composed of altered epidermal keratin. Recently, macular and lichenoid amyloidosis have been associated withMENsyndrome, type 2a.

Patients with *multicentric reticulohistiocytosis* also have pink-colored papules and nodules on the face and mucous membranes as well as on the extensor surface of the hands and forearms. They have a polyarthritis that can mimic rheumatoid arthritis clinically. On histologic examination, the papules have characteristic giant cells that are not seen in biopsies of rheumatoid nodules. Pink to skin-colored papules that are firm, 2 to 5 mm in diameter, and often in a linear arrangement are seen in patients with *papular mucinosis*. This disease is also referred to as *lichen myxedematosus* or *scleromyxedema*. The latter name comes from the brawny induration of the face and extremities that may accompany the papular eruption. Biopsy specimens of the papules show localized mucin deposition, and serum protein electrophoresis demonstrates a monoclonal spike of IgG, usually with a l light chain.

Yellow Lesions Several systemic disorders are characterized by yellow-colored cutaneous papules or plaques -- hyperlipidemia (xanthomas), gout (tophi), diabetes (necrobiosis lipoidica), pseudoxanthoma elasticum, and Torre syndrome (sebaceous tumors). Eruptive xanthomas are the most common form of *xanthomas*, and are associated with hypertriglyceridemia (types I, III, IV, and V). Crops of yellow papules with erythematous halos occur primarily on the extensor surfaces of the extremities and the buttocks, and they spontaneously involute with a fall in serum triglycerides. Increasedb-lipoproteins (primarily types II and III) result in one or more of the following types of xanthoma: xanthelasma, tendon xanthomas, and plane xanthomas. Xanthelasma are found on the eyelids, whereas tendon xanthomas are frequently associated with the Achilles and extensor finger tendons; plane xanthomas are flat and favor the palmar creases, face, upper trunk, and scars. Tuberous xanthomas are frequently associated with hypertriglyceridemia, but they are also seen in patients with hypercholesterolemia (type II) and are found most frequently over the large joints or hand. Biopsy specimens of xanthomas show collections of lipid-containing macrophages (foam cells).

Patients with several disorders, including biliary cirrhosis, can have a secondary form of hyperlipidemia with associated tuberous and planar xanthomas. However, patients with myeloma have *normolipemic* flat xanthomas. This latter form of xanthoma may be³12 cm in diameter and is most frequently seen on the upper trunk or side of the neck. It is also important to note that the most common setting for eruptive xanthomas is uncontrolled diabetes mellitus. The least specific sign for hyperlipidemia is xanthelasma, because at least 50% of the patients with this finding have normal lipid profiles.

In *tophaceous gout* there are deposits of monosodium urate in the skin around the joints, particularly those of the hands and feet. Additional sites of *tophi* formation include the helix of the ear and the olecranon and prepatellar bursae. The lesions are firm, yellow in color, and occasionally discharge a chalky material. Their size varies from 1 mm to 7 cm, and the diagnosis can be established by polarization of the aspirated contents of a lesion. Lesions of *necrobiosis lipoidica* are found primarily on the shins (90%), and patients can have diabetes mellitus or develop it subsequently. Characteristic findings include a central yellow color, atrophy (transparency), telangiectasias, and an erythematous border. Ulcerations can also develop within the plaques. Biopsy specimens show necrobiosis of collagen, granulomatous inflammation, and obliterative endarteritis.

In *pseudoxanthoma elasticum* (PXE) there is an abnormal deposition of calcium on the elastic fibers of the skin, eye, and blood vessels. In the skin, the flexural areas such as the neck, axillae, antecubital fossae, and inguinal area are the primary sites of involvement. Yellow papules coalesce to form reticulated plaques that have an appearance similar to that of plucked chicken skin. In severely affected skin, hanging, redundant folds develop. Some patients have a more subtle macular form of the disease, and careful inspection is required. Biopsy specimens of involved skin show swollen and irregularly clumped elastic fibers with deposits of calcium. In the eye, the calcium deposits in Bruch's membrane lead to angioid streaks and choroiditis; in the arteries of the heart, kidney, gastrointestinal tract, and extremities, the deposits lead to angina, hypertension, gastrointestinal bleeding, and claudication, respectively. Long-term administration of D-penicillamine can lead to PXE-like skin changes as well as elastic fiber alterations in internal organs.

Adnexal tumors that have differentiated toward sebaceous glands include sebaceous adenoma, sebaceous epithelioma, sebaceous carcinoma, and sebaceous hyperplasia. Except for sebaceous hyperplasia, which is commonly seen on the face, these tumors are fairly rare. Patients with Torre syndrome have *sebaceous adenomas*, and in the majority of cases there are multiple such tumors. These patients can also have sebaceous carcinomas and sebaceous hyperplasia as well as keratoacanthomas. The internal manifestations of Torre syndrome include *multiple* carcinomas of the gastrointestinal tract (primarily colon) as well as cancers of the larynx, genitourinary tract, and endometrium. Some patients also have a strong family history of cancer.

Red Lesions Cutaneous lesions that are red in color have a wide variety of etiologies; in an attempt to simplify their identification, they will be subdivided into papules, papules/plaques, and subcutaneous nodules. Common red papules include *arthropod bites* and *cherry hemangiomas*; the latter are small, bright-red, dome-shaped papules that represent benign proliferation of capillaries. In patients with AIDS, the development of multiple red hemangioma-like lesions points to bacillary angiomatosis, and biopsy specimens show clusters of bacilli that stain positive with the Warthin-Starry stain; the pathogens have been identified as *Bartonella henselae* and *B. quintana*. Disseminated visceral disease is seen primarily in immunocompromised hosts but can occur in immunocompetent individuals.

Multiple *angiokeratomas* are seen in Fabry's disease, an X-linked recessive lysosomal storage disease that is due to a deficiency of a-galactosidase A. The lesions are red to red-blue in color and can be quite small in size (1 to 3 mm), with the most common location being the lower trunk. Associated findings include chronic renal failure, peripheral neuropathy, and corneal opacities (cornea verticillata). Electron photomicrographs of angiokeratomas and clinically normal skin demonstrate lamellar lipid deposits in fibroblasts, pericytes, and endothelial cells that are diagnostic of this disease. Widespread acute eruptions of erythematous papules are discussed in the section on exanthems.

There are several infectious diseases that present as erythematous papules or nodules in a sporotrichoid pattern, that is, in a linear arrangement along the lymphatic channels. The two most common etiologies are *Sporothrix schenckii* (sporotrichosis) and *M. marinum* (atypical mycobacteria). The organisms are introduced as a result of trauma, and a primary inoculation site is often seen in addition to the lymphatic nodules. Additional causes include *Nocardia*, *Leishmania*, and other dimorphic fungi; culture of lesional tissue will aid in the diagnosis.

The diseases that are characterized by erythematous plaques with scale are reviewed in the papulosquamous section, and the various forms of dermatitis are discussed in the section on erythroderma. Additional disorders in the differential diagnosis of red papules/plaques include *erysipelas*, *polymorphous light eruption* (PMLE), *lymphocytoma cutis*, *cutaneous lupus*, *lymphoma cutis*, and *leukemia cutis*. The first three diseases represent primary cutaneous disorders. PMLE is characterized by erythematous papules and plaques in a primarily sun-exposed distribution -- dorsum of the hand, extensor forearm, and face. Lesions follow exposure to UV-B and/or UV-A, and in northern latitudes PMLE is most severe in the late spring and early summer. A process referred to as "hardening" occurs with continued UV exposure, and the eruption fades, but in temperate climates it will recur in the spring. PMLE must be differentiated from cutaneous lupus, and this is accomplished by histologic examination and direct immunofluorescence of the lesions. Lymphocytoma cutis (pseudolymphoma) is a *benign* polyclonal proliferation of lymphocytes in the skin that presents as infiltrated pink-red to red-purple papules and plaques; it must be distinguished from lymphoma cutis.

Several types of red plaques are seen in patients with systemic *lupus*, including (1) erythematous urticarial plaques across the cheeks and nose in the classic butterfly rash; (2) erythematous discoid lesions with fine or "carpet-tack" scale, telangiectasias, central hypopigmentation, peripheral hyperpigmentation, follicular plugging, and atrophy located on the face, scalp, external ears, arms, and upper trunk; and (3) psoriasiform or annular lesions of subacute lupus with hypopigmented centers located on the face, extensor arms, and upper trunk. Additional cutaneous findings include (1) a violaceous flush on the face and V of the neck; (2) urticarial vasculitis (see "Urticaria"); (3) lupus panniculitis (see below); (4) diffuse alopecia; (5) alopecia secondary to discoid lesions; (6) periungual telangiectasias and erythema; (7) erythema multiforme-like lesions that may become bullous; and (8) distal ulcerations secondary to Raynaud's phenomenon, vasculitis, or livedoid vasculitis. Patients with only discoid lesions usually have the form of lupus that is limited to the skin. However, 2 to 10% of these patients eventually develop systemic lupus. Direct immunofluorescence of involved skin shows deposits of IgG or IgM and C3 in a granular distribution along the dermal-epidermal junction.

In *lymphoma cutis* there is a proliferation of malignant lymphocytes or histiocytes in the skin, and the clinical appearance resembles that of lymphocytoma cutis -- infiltrated pink-red to red-purple papules and plaques. Lymphoma cutis can occur anywhere on the surface of the skin, whereas the sites of predilection for lymphocytomas include the malar ridge, tip of the nose, and earlobes. Patients with non-Hodgkin's lymphomas have specific cutaneous lesions more often than those with Hodgkin's disease, and occasionally, the skin nodules precede the development of extracutaneous non-Hodgkin's lymphoma or represent the only site of involvement. Arcuate lesions are sometimes seen in lymphoma and lymphocytoma cutis as well as in CTCL. *Leukemia cutis* has the same appearance as lymphoma cutis, and specific lesions are seen more commonly in monocytic leukemias than in lymphocytic or granulocytic leukemias. Cutaneous chloromas (granulocytic sarcomas) may precede the appearance of circulating blasts in acute nonlymphocytic leukemia and, as such, represent a form of aleukemic leukemia cutis.

Common causes of erythematous subcutaneous nodules include inflamed epidermoid inclusion cysts, acne cysts, and furuncles. *Panniculitis*, an inflammation of the fat, also presents as subcutaneous nodules and is frequently a sign of systemic disease. There are several forms of panniculitis, including erythema nodosum, erythema induratum, lupus profundus, lipomembranous lipodermatosclerosis,a$_1$-antitrypsin deficiency, facticial, and fat necrosis secondary to pancreatic disease. Except for erythema nodosum, these lesions may break down and ulcerate or heal with a scar. The shin is the most common location for the nodules of erythema nodosum, whereas the calf is the most common location for lesions of erythema induratum. In erythema nodosum the nodules are initially red but then develop a blue color as they resolve. Patients with erythema nodosum but no underlying systemic illness can still have fever, malaise, leukocytosis, arthralgias, and/or arthritis. However, the possibility of an underlying illness should be excluded, and the most common associations are streptococcal infections, upper respiratory infections, sarcoidosis, and inflammatory bowel disease. The less common associations include tuberculosis, histoplasmosis, coccidioidomycosis, psittacosis, drugs (oral contraceptives, sulfonamides, aspartame, bromides, iodides), cat-scratch fever, and infections with *Yersinia*, *Salmonella*, and *Chlamydia*.

In some patients, erythema induratum/nodular vasculitis is an idiopathic disease; however, in approximately 25 to 70% of patients, polymerase chain reaction (PCR) analysis will demonstrate *M. tuberculosis* complex DNA. The lesions of lupus profundus are found primarily on the upper arms and buttocks (sites of abundant fat) and are seen in both the cutaneous and systemic forms of lupus. The overlying skin may be normal, erythematous, or have the changes of discoid lupus. The subcutaneous fat necrosis that is associated with pancreatic disease is presumably secondary to circulating lipases and is seen in patients with pancreatic carcinoma as well as in patients with acute and chronic pancreatitis. In this disorder there may be an associated arthritis, fever, and inflammation of visceral fat. Histologic examination of deep incisional biopsy specimens will aid in the diagnosis of the particular type of panniculitis.

Subcutaneous erythematous nodules are also seen in *cutaneous polyarteritis nodosa* (PAN) and as a manifestation of *systemic vasculitis*, e.g., systemic PAN, allergic granulomatosis, or Wegener's granulomatosis (Chap. 317). Cutaneous PAN presents with painful subcutaneous nodules and ulcers within a red-purple, netlike pattern of livedo reticularis. The latter is due to slowed blood flow through the superficial horizontal venous plexus. The majority of lesions are found on the lower extremity, and while arthralgias and myalgias may accompany cutaneous PAN, there is no evidence of systemic involvement. In both the cutaneous and systemic forms of vasculitis, skin biopsy specimens of the associated nodules will show the changes characteristic of a vasculitis; the size of the vessel involved will depend on the particular disease.

Red-Brown Lesions The cutaneous lesions in *sarcoidosis* (Chap. 318) are classically red to red-brown in color, and with diascopy (pressure with a glass slide) a yellow-brown residual color is observed that is secondary to the granulomatous infiltrate. The waxy papules and plaques may be found anywhere on the skin, but the face is the most common location. Usually there are no surface changes, but occasionally the lesions will have scale. Biopsy specimens of the papules show "naked" granulomas in the dermis, i.e., granulomas surrounded by a minimal number of lymphocytes. Other cutaneous findings in sarcoidosis include annular lesions with an atrophic or scaly center, papules within scars, hypopigmented macules and papules, alopecia, acquired ichthyosis, erythema nodosum, and lupus pernio (see below). Additional physical findings are peripheral lymphadenopathy and parotid and lacrimal gland enlargement. When there is cutaneous involvement of the hands, radiographs will often show lytic lesions in the underlying bone.

The differential diagnosis of sarcoidosis includes foreign-body granulomas produced by chemicals such as beryllium and zirconium, late secondary syphilis, and *lupus vulgaris*. Lupus vulgaris is a form of cutaneous tuberculosis that is seen in previously infected and sensitized individuals. There is often underlying active tuberculosis elsewhere, usually in the lungs or lymph nodes. At least 90% of the lesions occur in the head and neck area and are red-brown plaques with a yellow-brown color on diascopy. Secondary scarring and squamous cell carcinomas can develop within the plaques. Cultures orPCRanalysis of the lesions should be done because it is rare for the acid-fast stain to show bacilli within the dermal granulomas.

Sweet's syndrome is characterized by red to red-brown plaques and nodules that are frequently painful and occur primarily on the head, neck, and upper extremities. The patients also have fever, neutrophilia, and a dense dermal infiltrate of neutrophils in the lesions. In approximately 10% of the patients there is an associated malignancy, most commonly acute nonlymphocytic leukemia. Sweet's syndrome has also been reported with lymphoma, chronic leukemia, myeloma, myelodysplastic syndromes, and solid tumors (primarily of the genitourinary tract). The differential diagnosis includes neutrophilic eccrine hidradenitis and atypical forms of pyoderma gangrenosum. Extracutaneous sites of involvement include joints, muscles, eye, kidney (proteinuria, occasionally glomerulonephritis), and lung (neutrophilic infiltrates). The idiopathic form of Sweet's syndrome is seen more often in women, following a respiratory tract infection.

A generalized distribution of red-brown macules and papules is seen in the form of mastocytosis known as *urticaria pigmentosa* (Chap. 310). Each lesion represents a collection of mast cells in the dermis, with hyperpigmentation of the overlying epidermis. Stimuli such as rubbing cause these mast cells to degranulate, and this leads to the formation of localized urticaria (Darier's sign). Additional symptoms can result from mast cell degranulation and include headache, flushing, diarrhea, and pruritus. Mast cells also infiltrate various organs such as the liver, spleen, and gastrointestinal tract in up to

30 to 50% of patients with urticaria pigmentosa, and accumulations of mast cells in the bones may produce either osteosclerotic or osteolytic shadows on radiographs. In the majority of these patients, however, the internal involvement remains fairly static. A subtype of chronic leukocytoclastic vasculitis, *erythema elevatum diutinum* (EED), also presents with papules that are red-brown in color. The papules coalesce into plaques on the extensor surfaces of knees, elbows, and the small joints of the hand. Flares of EED have been associated with streptococcal infections.

Blue Lesions Lesions that are blue in color are the result of either vascular ectasias and tumors or melanin pigment in the dermis. *Venous lakes* (ectasias) are compressible dark-blue lesions that are found commonly in the head and neck region. *Venous malformations* are also compressible blue papules and nodules that can occur anywhere on the body, including the oral mucosa. When they are multiple rather than single congenital lesions, the patient may have the blue rubber bleb syndrome or Mafucci's syndrome. Patients with the blue rubber bleb syndrome also have vascular anomalies of the gastrointestinal tract that may bleed, whereas patients with Mafucci's syndrome have associated dyschondroplasia and osteochondromas. *Blue nevi* (moles) are seen when there are collections of pigment-producing nevus cells in the dermis. These benign papular lesions are dome-shaped and occur most commonly on the dorsum of the hand or foot.

Violaceous Lesions Violaceous papules and plaques are seen in *lupus pernio*, *lymphoma cutis*, and *cutaneous lupus*. Lupus pernio is a particular type of sarcoidosis that involves the tip of the nose and the earlobes, with lesions that are violaceous in color rather than red-brown. This form of sarcoidosis is associated with involvement of the upper respiratory tract. The plaques of lymphoma cutis and cutaneous lupus may be red or violaceous in color and were discussed above.

Purple Lesions Purple-colored papules and plaques are seen in vascular tumors, such as *Kaposi's sarcoma* (Chap. 309) and *angiosarcoma*, and when there is extravasation of red blood cells into the skin in association with inflammation, as in *palpable purpura* (see "Purpura"). Patients with congenital or acquiredAVfistulas and venous hypertension can develop purple papules on the lower extremities that can resemble Kaposi's sarcoma clinically and histologically; this condition is referred to as pseudo-Kaposi sarcoma (acral angiodermatitis). Angiosarcoma is found most commonly on the scalp and face of elderly patients or within areas of chronic lymphedema and presents as purple papules and plaques. In the head and neck region the tumor often extends beyond the clinically defined borders and may be accompanied by facial edema.

Brown and Black Lesions Brown- and black-colored papules are reviewed in "Hyperpigmentation."

Cutaneous Metastases These are discussed last because they can have a wide range of colors. Most commonly they present as either firm, skin-colored subcutaneous nodules or firm, red to red-brown papulonodules. The lesions of lymphoma cutis range from pink-red to plum in color, whereas metastatic melanoma can be pink, blue, or black in color. Cutaneous metastases develop from hematogenous or lymphatic spread and are most often due to the following primary carcinomas: in men, lung, colon, melanoma, and oral cavity; and in women, breast, colon, and lung. These metastatic lesions may be the initial presentation of the carcinoma, especially when the primary site is the lung, kidney, or ovary.

PURPURA (Table 57-16)

Purpura are seen when there is an extravasation of red blood cells into the dermis, and as a result, the lesions do not blanch with pressure. This is in contrast to those erythematous or violet-colored lesions that are due to localized vasodilatation -- they do blanch with pressure. Purpura (³3 mm) and petechiae (£2 mm) are divided into two major groups, palpable and nonpalpable. The most frequent causes of *nonpalpable* petechiae and purpura are primary cutaneous disorders such as *trauma*, *solar purpura*, and *capillaritis*. Less common causes are *steroid purpura* and *livedoid vasculitis* (see "Ulcers"). Solar purpura are seen primarily on the extensor forearms, while glucocorticoid purpura secondary to potent topical steroids or endogenous or exogenous Cushing's syndrome can be more widespread. In both cases there is alteration of the supporting connective tissue that surrounds the dermal blood vessels. In contrast, the petechiae that result from capillaritis are found primarily on the lower extremities. In capillaritis there is an extravasation of erythrocytes as a result of perivascular lymphocytic inflammation. The petechiae are bright red, 1 to 2 mm in size, and scattered within annular or coin-shaped yellow-brown macules. The yellow-brown color is caused by hemosiderin deposits within the dermis.

Systemic causes of nonpalpable purpura fall into several categories, and those secondary to clotting disturbances and vascular fragility will be discussed first. The former group includes *thrombocytopenia* (Chap. 116), *abnormal platelet function* as is seen in uremia, and *clotting factor defects*. The initial site of presentation for thrombocytopenia-induced petechiae is the distal lower extremity. Capillary fragility leads to nonpalpable purpura in patients with systemic *amyloidosis* (see "Papulonodular Skin Lesions"), disorders of collagen production such as *Ehlers-Danlos syndrome*, and *scurvy*. In scurvy there are flattened corkscrew hairs with surrounding hemorrhage on the lower extremities, in addition to gingivitis. Vitamin C is a cofactor for lysyl hydroxylase, an enzyme involved in the posttranslational modification of procollagen that is necessary for cross-link formation.

In contrast to the previous group of disorders, the purpura seen in the following group of diseases are associated with thrombi formation within vessels. It is important to note that these thrombi are demonstrable in skin biopsy specimens. This group of disorders includes disseminated intravascular coagulation (DIC), monoclonal cryoglobulinemia, thrombotic thrombocytopenic purpura, and reactions to warfarin. DIC is triggered by several types of infection (gram-negative, gram-positive, viral, and rickettsial) as well as by tissue injury and neoplasms. Widespread purpura and hemorrhagic infarcts of the distal extremities are seen. Similar lesions are found in purpura fulminans, which is a form of DIC associated with fever and hypotension that occurs more commonly in children following an infectious illness such as varicella, scarlet fever, or an upper respiratory tract infection. In both disorders, hemorrhagic bullae can develop in involved skin.

Monoclonal cryoglobulinemia is associated with multiple myeloma, Waldenstrom's macroglobulinemia, lymphocytic leukemia, and lymphoma. Purpura, primarily of the lower extremities, and hemorrhagic infarcts of the fingers and toes are seen in these patients. Exacerbations of disease activity can follow cold exposure or an increase in serum viscosity. Biopsy specimens show precipitates of the cryoglobulin within dermal vessels. Similar deposits have been found in the lung, brain, and renal glomeruli. Patients with *thrombotic thrombocytopenic purpura* can also have hemorrhagic infarcts as a result of intravascular thromboses. Additional signs include thrombocytopenic purpura, fever, and microangiopathic hemolytic anemia (Chap. 108).

Administration of *warfarin* can result in painful areas of erythema that become purpuric and then necrotic with an adherent black eschar. This reaction is seen more often in women and in areas with abundant subcutaneous fat -- breasts, abdomen, buttocks, thighs, and calves. The erythema and purpura develop between the third and tenth day of therapy, most likely as a result of a transient imbalance in the levels of anticoagulant and procoagulant vitamin K-dependent factors. Continued therapy does not exacerbate preexisting lesions, and patients with an inherited or acquired deficiency of protein C are at increased risk for this particular reaction as well as for purpura fulminans.

Purpura secondary to *cholesterol emboli* are usually seen on the lower extremities of patients with atherosclerotic vascular disease. They often follow anticoagulant therapy or an invasive vascular procedure such as an arteriogram but also occur spontaneously from disintegration of atheromatous plaques. Associated findings include livedo reticularis, gangrene, cyanosis, subcutaneous nodules, and ischemic ulcerations. Multiple step sections of the biopsy specimen may be necessary to demonstrate the cholesterol clefts with the vessels. Petechiae are also an important sign of *fat embolism* and occur primarily on the upper body 2 to 3 days after a major injury. By using special fixatives, the emboli can be demonstrated in biopsy specimens of the petechiae. Emboli of tumor or thrombus are seen in patients with atrial myxomas and marantic endocarditis.

In the *Gardner-Diamond syndrome* (autoerythrocyte sensitivity), female patients develop large ecchymoses within areas of painful, warm erythema. An episode of significant trauma frequently precedes the onset of this syndrome. Intradermal injections of autologous erythrocytes or phosphatidyl serine derived from the red cell membrane can reproduce the lesions in some patients; however, there are instances where a reaction is seen at an injection site of the forearm but not in the midback region. The latter has led some observers to view Gardner-Diamond syndrome as a cutaneous manifestation of severe emotional stress. *Waldenstrom's hypergammaglobulinemic purpura* is a chronic disorder characterized by petechiae on the lower extremities. There are circulating complexes of IgG-anti-IgG molecules, and exacerbations are associated with prolonged standing or walking.

Palpable purpura are further subdivided into vasculitic and embolic. In the group of vasculitic disorders, *leukocytoclastic vasculitis* (LCV), also known as *allergic vasculitis*, is the one most commonly associated with palpable purpura (Chap. 317). *Henoch-Schonlein purpura* is a subtype of acute LCV that is seen primarily in children and adolescents following an upper respiratory infection. The majority of lesions are found on the lower extremities and buttocks. Systemic manifestations include fever, arthralgias (primarily of the knees and ankles), abdominal pain, gastrointestinal bleeding, and nephritis. Direct immunofluorescence examination shows deposits of IgA within dermal blood vessel walls. In *polyarteritis nodosa*, specific cutaneous lesions result from a vasculitis of arterial vessels rather than postcapillary venules as in LCV. The arteritis leads to ischemia of the skin, and this explains the irregular outline of the purpura (see below).

Several types of infectious emboli can give rise to palpable purpura. These embolic lesions are usually *irregular* in outline as opposed to the lesions of leukocytoclastic vasculitis, which are *circular* in outline. The irregular outline is indicative of a cutaneous infarct, and the size corresponds to the area of skin that received its blood supply from that particular arteriole or artery. The palpable purpura inLCV are circular because the erythrocytes simply diffuse out evenly from the postcapillary venules as a result of inflammation. Infectious emboli are most commonly due to gram-negative cocci (meningococcus, gonococcus), gram-negative rods (Enterobacteriaceae), and gram-positive cocci (staphylococcus). Additional causes include *Rickettsia* and, in immunocompromised patients, *Candida* and opportunistic fungi.

The embolic lesions in *acute meningococcemia* are found primarily on the trunk, lower extremities, and sites of pressure, and a gunmetal-gray color often develops within them. Their size varies from 1 mm to several centimeters, and the organisms can be cultured from the lesions. Associated findings include a preceding upper respiratory tract infection, fever, meningitis,DIC, and, in some patients, a deficiency of the terminal components of complement. In *disseminated gonococcal infection* (arthritis-dermatitis syndrome), a small number of papules and vesicopustules with central purpura or hemorrhagic necrosis are found over the joints of the distal extremities. Additional symptoms include arthralgias, tenosynovitis, and fever. To establish the diagnosis, a Gram stain of these lesions should be performed. *Rocky mountain spotted fever* is a tick-borne disease that is caused by *R. rickettsii*. A several-day history of fever, chills, severe headache, and photophobia precedes the onset of the cutaneous eruption. The initial lesions are erythematous macules and papules on the wrists, ankles, palms, and soles. With time, the lesions spread centripetally and become purpuric.

Lesions of *ecthyma gangrenosum* begin as edematous, erythematous papules or plaques and then develop central purpura and necrosis. Bullae formation also occurs in these lesions, and they are frequently found in the girdle region. The organism that is classically associated with ecthyma gangrenosum is *Pseudomonas aeruginosa*, but other gram-negative rods such as *Klebsiella*, *Escherichia coli*, and *Serratia* can produce similar lesions. In immunocompromised hosts, the list of potential pathogens is expanded to include *Candida* and opportunistic fungi.

ULCERS

The approach to the patient with a cutaneous ulcer, is outlined inTable 57-17.**Peripheral vascular diseases of the extremities are reviewed in Chap. 248, as is Raynaud's phenomenon.*

Livedoid vasculitis (atrophie blanche) represents a combination of a vasculopathy with intravascular thrombosis. Purpuric lesions and livedo reticularis are found in association with painful ulcerations of the lower extremities. These ulcers are often slow to heal, but when they do, irregularly shaped white scars are formed. The majority of cases are secondary to venous hypertension, but possible underlying illnesses include cryofibrinogenemia and disorders of hypercoagulability, e.g., the antiphospholipid syndrome (Chap. 117).

In *pyoderma gangrenosum*, the border of the ulcers has a characteristic appearance of an undermined necrotic bluish edge and a peripheral erythematous halo. The ulcers often begin as pustules that then expand rather rapidly to a size as large as 20 cm. Although these lesions are most commonly found on the lower extremities, they can arise anywhere on the surface of the body, including sites of trauma (pathergy). An estimated 30 to 50% of cases are idiopathic, and the most common associated disorders are ulcerative colitis and Crohn's disease. Less commonly, it is associated with chronic active hepatitis, seropositive rheumatoid arthritis, acute and chronic granulocytic leukemia, polycythemia vera, and myeloma. Additional findings in these patients, even those with idiopathic disease, are cutaneous anergy and a benign monoclonal gammopathy. Because the histology of pyoderma gangrenosum is nonspecific, the diagnosis is made clinically by excluding less common causes of similar-appearing ulcers such as necrotizing vasculitis, Meleney's ulcer (synergistic infection at a site of trauma or surgery), dimorphic fungi, cutaneous amebiasis, spider bites, and facticial. In the myeloproliferative disorders, the ulcers may be more superficial with a pustulobullous border, and these lesions provide a connection between classic pyoderma gangrenosum and acute febrile neutrophilic dermatosis (Sweet's syndrome).

FEVER AND RASH

The major considerations in a patient with a fever and a rash are inflammatory diseases versus infectious diseases. In the hospital setting, the most common scenario is a patient who has a drug rash plus a fever secondary to an underlying infection. However, it should be emphasized that a drug reaction can lead to both a cutaneous eruption and a fever ("drug fever"). Additional inflammatory diseases that are often associated with a fever include pustular psoriasis, erythroderma, and Sweet's syndrome. Lyme disease, secondary syphilis, and viral and bacterial exanthems (see "Exanthems") are examples of infectious diseases that produce a rash and a fever. Lastly, it is important to determine whether or not the cutaneous lesions represent septic emboli (see "Purpura"). Such lesions usually have evidence of ischemia in the form of purpura, necrosis, or impending necrosis (gunmetal-gray color). In the patient with thrombocytopenia, however, purpura can be seen in inflammatory reactions such as morbilliform drug eruptions and infectious lesions.

(Bibliography omitted in Palm version)

Back to Table of Contents

58. IMMUNOLOGICALLY MEDIATED SKIN DISEASES - *Kim B. Yancey*, *Thomas J. Lawley*

A number of immunologically mediated skin diseases and cutaneous manifestations of immunologically mediated systemic disorders are now recognized as distinct entities with relatively consistent clinical, histologic, and immunopathologic findings. Many of these disorders are due to autoimmune mechanisms. Clinically, they are characterized by morbidity (pain, pruritus, disfigurement) and in some instances by mortality (largely due to loss of epidermal barrier function and/or secondary infection). The major features of the more common immunologically mediated skin diseases are summarized in this chapter (Table 58-1).

PEMPHIGUS VULGARIS

Pemphigus vulgaris (PV) is a blistering skin disease seen predominantly in elderly patients. Patients with PV have an increased incidence of the HLA-DR4 and -DRw6 serologically defined haplotypes. This disorder is characterized by the loss of cohesion between epidermal cells (a process termed *acantholysis*) with the resultant formation of intraepidermal blisters. Clinical lesions of PV typically consist of flaccid blisters on either normal-appearing or erythematous skin. These blisters rupture easily, leaving denuded areas that may crust and enlarge peripherally (Plate IIE-69). Substantial portions of the body surface may be denuded in severe cases. Manual pressure to the skin of these patients may elicit the separation of the epidermis (Nikolsky's sign). This finding, while characteristic of PV, is not specific to this disorder and is also seen in toxic epidermal necrolysis, Stevens-Johnson syndrome, and a few other skin diseases. Lesions in PV typically present on the oral mucosa, scalp, face, neck, axilla, and trunk. In half or more of patients, lesions begin in the mouth; approximately 90% of patients have oromucosal involvement at some time during the course of their disease. Involvement of other mucosal surfaces (e.g., pharyngeal, laryngeal, esophageal, conjunctival, vulval, or rectal) can occur in severe disease. Pruritus may be a feature of early pemphigus lesions; extensive denudation may be associated with severe pain. Lesions usually heal without scarring, except at sites complicated by secondary infection or mechanically induced dermal wounds. Nonetheless, postinflammatory hyperpigmentation is usually present at sites of healed lesions for some time.

Biopsies of early lesions demonstrate intraepidermal vesicle formation secondary to loss of cohesion between epidermal cells (i.e., acantholytic blisters). Blister cavities contain acantholytic epidermal cells, which appear as round homogeneous cells containing hyperchromatic nuclei. Basal keratinocytes remain attached to the epidermal basement membrane, hence blister formation is within the suprabasal portion of the epidermis. Lesional skin may contain focal collections of intraepidermal eosinophils within blister cavities; dermal alterations are slight, often limited to an eosinophil-predominant leukocytic infiltrate. Direct immunofluorescence microscopy of lesional or intact patient skin shows deposits of IgG on the surface of keratinocytes; in contrast, deposits of complement components are typically found in lesional but not uninvolved skin. Deposits of IgG on keratinocytes are derived from circulating autoantibodies directed against cell-surface antigens. Circulating autoantibodies can be demonstrated in 80 to 90% ofPVpatients by indirect immunofluorescence microscopy; monkey esophagus is the optimal substrate for demonstration of these autoantibodies. Patients with PV have

IgG autoantibodies directed against desmogleins (Dsgs), transmembrane desmosomal glycoproteins that belong to the cadherin supergene family of calcium-dependent adhesion molecules. While Dsg3 is specifically recognized by PV autoantibodies, approximately 50% of PV sera also contain IgG against Dsg1. Most patients with early PV and only mucosal involvement have only anti-Dsg3 autoantibodies, whereas most patients with advanced disease (i.e., involvement of skin and mucosa) have both anti-Dsg3 and anti-Dsg1 autoantibodies. Recent studies have shown that the anti-Dsg autoantibody profile in these patients' sera as well as the tissue distribution of Dsg3 and Dsg1 determine the site of blister formation in patients with pemphigus. Experimental studies have also shown that these autoantibodies are pathogenic (i.e., responsible for blister formation) and that their titer correlates with disease activity.

PVcan be life-threatening. Prior to the availability of glucocorticoids, the mortality ranged from 60 to 90%; the current mortality is approximately 5%. Common causes of morbidity and mortality are infection and complications of treatment with glucocorticoids. Bad prognostic factors include advanced age, widespread involvement, and the requirement for high doses of glucocorticoids (with or without other immunosuppressive agents) for control of disease. The course of PV in individual patients is variable and difficult to predict. Some patients achieve remission (40% of patients in some series), but others may require long-term treatment or succumb to complications of their disease or its treatment. The mainstay of treatment is systemic glucocorticoids. Patients with moderate to severe disease are usually started on prednisone, 60 to 80 mg/d. If new lesions continue to appear after 1 to 2 weeks of treatment, the dose should be increased. Many regimens combine an immunosuppressive agent with systemic glucocorticoids for control of PV. The two most frequently used are either azathioprine (1 mg/kg per day) or cyclophosphamide (1 mg/kg per day). It is important to bring severe or progressive disease under control quickly to lessen the severity and/or duration of this disorder.

PEMPHIGUS FOLIACEUS

Pemphigus foliaceus (PF) is distinguished fromPV by several features. In PF, acantholytic blisters are located high within the epidermis, usually just beneath the stratum corneum. Hence PF is a more superficial blistering disease than PV. The distribution of lesions in the two disorders is much the same, except that in PF mucous membrane lesions are very rare. Patients with PF rarely demonstrate intact blisters but rather exhibit shallow erosions associated with erythema, scale, and crust formation. Mild cases of PF resemble severe seborrheic dermatitis; severe PF may cause extensive exfoliation. Sun exposure (ultraviolet irradiation) may be an aggravating factor. A blistering skin disease endemic to south central Brazil known as *fogo selvagem*, or *Brazilian pemphigus*, is clinically, histologically, and immunopathologically indistinguishable from PF.

Patients withPF have immunopathologic features in common withPV. Specifically, direct immunofluorescence microscopy of perilesional skin demonstrates IgG on the surface of keratinocytes. As in PV, patients with PF frequently have circulating IgG autoantibodies against keratinocyte cell surface antigens. Guinea pig esophagus is the optimal substrate for indirect immunofluorescence microscopy studies of sera from patients with PF. In PF, autoantibodies are directed against Dsg1, a 160-kDa desmosomal cadherin.

As noted for PV, the autoantibody profile in patients with PF (i.e., anti-Dsg1) and the normal tissue distribution of this autoantigen (i.e., low expression in oral mucosa) is thought to account for the distribution of lesions in this disease.

Although pemphigus has been associated with several autoimmune diseases, its association with thymoma and/or myasthenia gravis is particularly notable. To date, more than 30 cases of thymoma and/or myasthenia gravis have been reported in association with pemphigus, usually withPF. Patients may also develop pemphigus as a consequence of drug exposure. The most frequently implicated agent is penicillamine; other offenders include captopril, rifampin, piroxicam, penicillin, and phenobarbital. Drug-induced pemphigus usually resembles PF rather thanPV; autoantibodies in these patients have the same antigenic specificity as they do in other pemphigus patients. In most patients, lesions resolve following discontinuation of the drug; however, some patients require treatment with systemic glucocorticoids and/or immunosuppressive agents.

PFis generally a far less severe disease thanPV and carries a better prognosis. Localized disease can be treated conservatively with topical or intralesional glucocorticoids; more active cases can usually be controlled with systemic glucocorticoids.

PARANEOPLASTIC PEMPHIGUS

Paraneoplastic pemphigus (PNP) is an autoimmune acantholytic mucocutaneous disease associated with an occult or confirmed neoplasm. Patients with PNP typically show painful mucosal erosive lesions in association with pruritic papulosquamous eruptions that often progress to blisters. Palm and sole involvement is common in these patients and raises the possibility that prior reports of neoplasia-associated erythema multiforme actually may have represented unrecognized cases of PNP. Biopsies of lesional skin from these patients show varying combinations of acantholysis, keratinocyte necrosis, and vacuolar-interface dermatitis. Direct immunofluorescence microscopy of patient skin shows deposits of IgG and complement on the surface of keratinocytes and (variably) similar immunoreactants in the epidermal basement membrane zone. Patients with PNP have IgG autoantibodies against cytoplasmic proteins that are members of the plakin family (e.g., desmoplakins I and II, bullous pemphigoid antigen 1, envoplakin, periplakin, and plectin) and cell-surface proteins that are members of the cadherin family (e.g., Dsg3 and Dsg1). Because immunoadsorption of anti-Dsg3 IgG is sufficient to eliminate the ability of PNP sera to induce blisters in an experimental passive transfer animal model, these particular autoantibodies are thought to play a key pathogenic role in blister formation in these patients.

Although PNP is generally resistant to conventional therapies (i.e., those used to treat PV), patients may improve (or even remit) following resection of underlying neoplasms. The predominant neoplasms associated with this disorder are non-Hodgkin's lymphoma, chronic lymphocytic leukemia, Castleman's disease, thymoma, and spindle cell tumors.

BULLOUS PEMPHIGOID

Bullous pemphigoid (BP) is an autoimmune subepidermal blistering disease usually seen in the elderly. Lesions typically consist of tense blisters on either normal-appearing or erythematous skin (Plate IIE-72). The lesions are usually distributed over the lower abdomen, groin, and flexor surface of the extremities; oral mucosal lesions are found in 10 to 40% of patients. Pruritus may be nonexistent or severe. As lesions evolve, tense blisters tend to rupture and be replaced by flaccid lesions or erosions with or without surmounting crust. Nontraumatized blisters heal without scarring. The major histocompatibility complex class II allele HLA-DQb1*0301 is prevalent in patients with BP. Despite isolated reports, several studies have shown that patients with BP do not have an increased incidence of malignancy in comparison with appropriately age- and sex-matched controls.

While biopsies of early lesional skin demonstrate subepidermal blisters, the histologic features depend on the character of the particular lesion. Lesions on normal-appearing skin generally show a sparse perivascular leukocytic infiltrate with some eosinophils; conversely, biopsies of inflammatory lesions typically show an eosinophil-rich infiltrate within the papillary dermis at sites of vesicle formation and in perivascular areas. In addition to eosinophils, cell-rich lesions also contain mononuclear cells and neutrophils. It is not always possible to distinguishBP from other subepidermal blistering diseases by routine histologic techniques.

Immunopathologic studies have broadened our understanding of this disease and aided its diagnosis. Direct immunofluorescence microscopy of normal-appearing perilesional skin shows linear deposits of IgG and/or C3 in the epidermal basement membrane. The sera of approximately 70% of these patients contain circulating IgG autoantibodies that bind the epidermal basement membrane of normal human skin in indirect immunofluorescence microscopy. An even higher percentage of patients shows reactivity to the epidermal side of 1 *M* NaCl split skin [an alternative immunofluorescence microscopy test substrate that is commonly used to distinguish circulating IgG anti-basement membrane autoantibodies in patients with BPfrom those in patients with similar, yet different, subepidermal blistering diseases (e.g., epidermolysis bullosa acquisita, see below)]. No correlation exists between the titer of these autoantibodies and disease activity. In BP, circulating autoantibodies recognize 230- and (in approximately 70% of BP patients) 180-kDa hemidesmosome-associated proteins in basal keratinocytes [i.e., bullous pemphigoid antigen (BPAG)1 and BPAG2, respectively]. Autoantibodies are thought to develop against these antigens (more specifically, initially against BPAG2), deposit in situ, and activate complement that subsequently produces dermal mast cell degranulation and granulocyte-rich infiltrates that cause tissue damage and blister formation.

BPmay persist for months to years, with exacerbations or remissions. Although extensive involvement may result in widespread erosions and compromise cutaneous integrity, the mortality rate is low even in the absence of treatment. Nonetheless, deaths may occur in elderly and/or debilitated patients. The mainstay of treatment is systemic glucocorticoids. Patients with local or minimal disease can sometimes be controlled with topical glucocorticoids alone; patients with more extensive lesions generally respond to systemic glucocorticoids either alone or in combination with immunosuppressive agents. Patients will usually respond to prednisone, 40 to 60 mg/d. In some instances, azathioprine (1 mg/kg per day) or cyclophosphamide (1 mg/kg per day) are necessary adjuncts.

PEMPHIGOID GESTATIONIS

Pemphigoid gestationis (PG), also known as herpes gestationis, is a rare, nonviral, subepidermal blistering disease of pregnancy and the puerperium. PG may begin during any trimester of pregnancy or present shortly after delivery. Lesions are usually distributed over the abdomen, trunk, and extremities; mucous membrane lesions are rare. Skin lesions in these patients may be quite polymorphic and consist of erythematous urticarial papules and plaques, vesiculopapules, and/or frank bullae. Lesions are almost always very pruritic. Severe exacerbations of PG frequently occur after delivery, typically within 24 to 48 h. PG tends to recur in subsequent pregnancies, often beginning earlier during such gestations. Brief flare-ups of disease may occur with resumption of menses and may develop in patients later exposed to oral contraceptives. Occasionally, infants of affected mothers demonstrate transient skin lesions.

Biopsies of early lesional skin show teardrop-shaped subepidermal vesicles forming in dermal papillae in association with an eosinophil-rich leukocytic infiltrate. Differentiation ofPG from other subepidermal bullous diseases by light microscopy is often difficult. However, direct immunofluorescence microscopy of perilesional skin from PG patients reveals the immunopathologic hallmark of this disorder -- linear deposits of C3 in the epidermal basement membrane zone. These deposits develop as a consequence of complement activation produced by low titer IgG anti-basement membrane zone autoantibodies. Recent studies have shown that the majority of PG sera contain autoantibodies that recognizeBPAG2, the same 180-kDa hemidesmosome-associated protein that is targeted by autoantibodies in roughly 70% of patients withBP -- a subepidermal bullous disease that resembles PG morphologically, histologically, and immunopathologically.

The goals of therapy in patients withPG are to prevent the development of new lesions, relieve intense pruritus, and care for erosions at sites of blister formation. Most patients require treatment with moderate doses of daily glucocorticoids (i.e., 20 to 40 mg of prednisone) at some point in their course. Mild cases (or brief flare-ups) may be controlled by vigorous use of potent topical glucocorticoids. Although PG was once thought to be associated with an increased risk of fetal morbidity and mortality, the best evidence now suggests that these infants may only be at increased risk of being slightly premature or "small for dates." Current evidence suggests that there is no difference in the incidence of uncomplicated live births in PG patients treated with systemic glucocorticoids and in those managed more conservatively. If systemic glucocorticoids are administered, newborns are at risk for development of reversible adrenal insufficiency.

DERMATITIS HERPETIFORMIS

Dermatitis herpetiformis (DH) is an intensely pruritic, papulovesicular skin disease characterized by lesions symmetrically distributed over extensor surfaces (i.e., elbows, knees, buttocks, back, scalp, and posterior neck) (Plate IIE-68). The primary lesion in this disorder is a papule, papulovesicle, or urticarial plaque. Because pruritus is prominent, patients may present with excoriations and crusted papules but no observable primary lesions. Patients sometimes report that their pruritus has a distinctive burning or stinging component; the onset of such local symptoms reliably heralds the development of distinct clinical lesions 12 to 24 h later. Almost all DH patients have an associated, usually subclinical, gluten-sensitive enteropathy (Chap. 286), and more than 90% express the HLA-B8/DRw3 and HLA-DQw2 haplotypes. DH may present at any age, including childhood; onset in the second to fourth decades is most common. The disease is typically chronic.

Biopsy of early lesional skin reveals neutrophil-rich infiltrates within dermal papillae. Neutrophils, fibrin, edema, and microvesicle formation at these sites are characteristic of early disease. Older lesions may demonstrate nonspecific features of a subepidermal bulla or an excoriated papule. Because the clinical and histologic features of this disease can be variable and resemble other subepidermal blistering disorders, the diagnosis is confirmed by direct immunofluorescence microscopy of normal-appearing perilesional skin. Such studies demonstrate granular deposits of IgA (with or without complement components) in the papillary dermis and along the epidermal basement membrane zone. IgA deposits in the skin are unaffected by control of disease with medication; however, these immunoreactants may diminish in intensity or disappear in patients maintained for long periods on a strict gluten-free diet (see below). Patients with granular deposits of IgA in their epidermal basement membrane zone typically do not have circulating IgA anti-basement membrane autoantibodies and should be distinguished from individuals with linear IgA deposits at this site (see below).

Although mostDHpatients do not report overt gastrointestinal symptoms or laboratory evidence of malabsorption, biopsies of small bowel usually reveal blunting of intestinal villi and a lymphocytic infiltrate in the lamina propria. As is true for patients with celiac disease, this gastrointestinal abnormality can be reversed by a gluten-free diet. Moreover, if maintained, this diet alone may control the skin disease and eventuate in clearance of IgA deposits from these patients' epidermal basement membrane zone. Subsequent gluten exposure in such patients alters the morphology of their small bowel, elicits a flare-up of their skin disease, and is associated with the reappearance of IgA in their epidermal basement membrane zone. Additional evidence that DH develops as a consequence of dietary gluten exposure is the demonstration of IgA anti-endomysial antibodies in these patients' sera (as found in the sera of patients with ordinary gluten-sensitive enteropathy). Recent studies have shown that such autoantibodies are directed against tissue transglutaminase. Patients with DH also have an increased incidence of thyroid abnormalities, achlorhydria, atrophic gastritis, and antigastric parietal cell antibodies. These associations likely relate to the high frequency of the HLA-B8/DRw3 haplotype in these patients, since this marker is commonly linked to autoimmune disorders. The mainstay of treatment of DH is dapsone, a sulfone. Patients respond rapidly (24 to 48 h) to dapsone but require careful pretreatment evaluation and close follow-up to ensure that complications are avoided or controlled. All patients on more than 100 mg/d dapsone will have some hemolysis and methemoglobinemia. These are expected pharmacologic side effects of this agent. Gluten restriction can control DH and lessen dapsone requirements; this diet must rigidly exclude gluten to be of maximal benefit. Many months of dietary restriction may be necessary before a beneficial result is achieved. Good dietary counselling by a trained dietitian is essential.

LINEAR IGA DISEASE

Linear IgA disease, once considered a variant form of dermatitis herpetiformis, is actually a separate and distinct entity. Clinically, these patients may resemble patients with typical cases ofDH,BP, or other subepidermal blistering diseases. Lesions typically consist of papulovesicles, bullae, and/or urticarial plaques, predominantly on extensor (as seen in "classic" DH), central, or flexural sites. Oral mucosal involvement occurs in some patients. Severe pruritus resembles that in patients with DH. Patients with linear IgA disease do not have an increased frequency of the HLA-B8/DRw3 haplotype or an associated enteropathy and hence are not candidates for a gluten-free diet.

The histologic alterations in early lesions may be virtually indistinguishable from those inDH. However, direct immunofluorescence microscopy of normal-appearing perilesional skin reveals linear deposits of IgA (and often C3) in the epidermal basement membrane zone. Most patients with linear IgA disease demonstrate circulating IgA anti-basement membrane autoantibodies against epitopes in the extracellular domain ofBPAG2, a transmembrane protein found in hemidesmosomes of basal keratinocytes. These patients generally respond to treatment with dapsone, 50 to 150 mg/d.

EPIDERMOLYSIS BULLOSA ACQUISITA

EBAis a rare, noninherited, polymorphic, subepidermal blistering disease. (The inherited form is discussed in Chap. 351.) Patients with classic or noninflammatory EBA have blisters on noninflamed skin, atrophic scars, milia, nail dystrophy, and oral lesions. Because lesions generally occur at sites exposed to minor trauma, classic EBA is considered to be a mechanobullous disease. Other patients with EBA have widespread inflammatory, scarring, bullous lesions and oromucosal involvement that resembles severeBP. Some patients present with an inflammatory bullous disease that evolves into the classic noninflammatory form of this disorder. In general, EBA is chronic; associations with multiple myeloma, amyloidosis, inflammatory bowel disease, and diabetes mellitus have been reported. The HLA-DR2 haplotype is found with increased frequency in these patients.

The histology of lesional skin varies depending on the character of the lesion being studied. Noninflammatory bullae show subepidermal blisters with a sparse leukocytic infiltrate and resemble those in patients with porphyria cutanea tarda. Inflammatory lesions consist of a subepidermal blister and neutrophil-rich leukocytic infiltrates in the superficial dermis.EBApatients have continuous deposits of IgG (and frequently C3 as well as other complement components) in a linear pattern within the epidermal basement membrane zone. Ultrastructurally, these immunoreactants are found in the sublamina densa region in association with anchoring fibrils, wheat stack-like structures that extend from the lamina densa into the underlying papillary dermis. Approximately 25 to 50% of EBA patients have circulating IgG anti-basement membrane autoantibodies directed against type VII collagen -- the collagen species that comprises anchoring fibrils. Such IgG autoantibodies bind the dermal side of 1 *M* NaCl split skin (in contrast to IgG autoantibodies in patients withBP that bind either epidermal or both sides of this indirect immunofluorescence microscopy test substrate).

Treatment ofEBA is generally unsatisfactory. Some patients with inflammatory EBA may respond to systemic glucocorticoids, either alone or in combination with immunosuppressive agents. Other patients (especially those with neutrophil-rich inflammatory lesions) may respond to dapsone. The chronic, noninflammatory form of this disease is largely resistant to treatment, although some patients may respond to cyclosporine.

CICATRICIAL PEMPHIGOID

Cicatricial pemphigoid (CP) is a rare, acquired, subepithelial blistering disease characterized by erosive lesions of mucous membranes and skin that result in scarring of at least some sites of involvement. Immunopathologically, perilesional mucosa and skin of patients with CP demonstrate in situ deposits of immunoreactants in epithelial basement membranes. Common sites of involvement include the oral mucosa (especially the gingiva) and conjunctiva; other sites that may be affected include the nasopharyngeal, laryngeal, esophageal, urogenital, and rectal mucosa. Skin lesions (present in about one-third of patients) tend to predominate on the scalp, face, and upper trunk and generally consist of a few scattered erosions or tense blisters on an erythematous or urticarial base. CP is typically a chronic and progressive disorder. Serious complications may arise as a consequence of ocular, laryngeal, esophageal, or urogenital lesions. Erosive conjunctivitis may result in shortened fornices, symblephara, ankyloblepharon, entropion, corneal opacities, and (in severe cases) blindness. Similarly, erosive lesions of the larynx may cause hoarseness, pain, and tissue loss that if unrecognized and untreated may eventuate in complete destruction of the airway. Esophageal lesions may result in stenosis and/or strictures that may place patients at risk for aspiration. Strictures may also complicate urogenital involvement.

Biopsies of lesional tissue generally demonstrate subepithelial vesiculobullae and a mononuclear leukocytic infiltrate. Neutrophils and eosinophils may be seen in biopsies of early lesions; older lesions may demonstrate a scant leukocytic infiltrate and fibrosis. Direct immunofluorescence microscopy of perilesional tissue typically demonstrates deposits of IgG, IgA, and/or C3 in these patients' epithelial basement membranes. Because many of these patients show no evidence of circulating anti-basement membrane autoantibodies, testing of perilesional skin is important diagnostically. AlthoughCP was once thought to be a single nosologic entity, it is now largely regarded as a disease phenotype that may develop as a consequence of an autoimmune reaction against a variety of different molecules in epithelial basement membranes (e.g.,BPAG2, laminin 5, type VII collagen, and other antigens yet to be completely defined). Treatment of CP is largely dependent upon sites of involvement. Due to potentially severe complications, ocular, laryngeal, esophageal, and/or urogenital involvement require aggressive systemic treatment with dapsone, prednisone, or the latter in combination with another immunosuppressive agent (e.g., azathioprine or cyclophosphamide). Less threatening forms of the disease may be managed with topical or intralesional glucocorticoids.

AUTOIMMUNE SYSTEMIC DISEASES WITH PROMINENT CUTANEOUS FEATURES

DERMATOMYOSITIS

The cutaneous manifestations of dermatomyositis (Chap. 382) are often distinctive but at times may resemble those of systemic lupus erythematosus (SLE) (Chap. 311), scleroderma (Chap. 313), or other overlapping connective tissue diseases (Chap. 313).

The extent and severity of cutaneous disease may or may not correlate with the extent and severity of the myositis. Patients with severe muscle involvement may have relatively minor skin changes, whereas patients with marked skin involvement may have mild muscle disease. The cutaneous manifestations of dermatomyositis are similar whether the disease appears in childhood or old age, except that calcification of subcutaneous tissue is a common late sequela in childhood dermatomyositis.

The cutaneous signs of dermatomyositis may precede or follow the development of myositis by weeks to years. Cases lacking muscle involvement (i.e., dermatomyositis sine myositis) have also been reported. The most common manifestation is a purple-red discoloration of the upper eyelids, sometimes associated with scaling ("heliotrope" erythema; Plate IIE-63) and periorbital edema. Erythema on the cheeks and nose in a "butterfly" distribution may resemble the eruption inSLE. Erythematous or violaceous scaling patches are common on the upper anterior chest, posterior neck, scalp, and the extensor surfaces of the arms, legs, and hands. Erythema and scaling may be particularly prominent over the elbows, knees, and the dorsal interphalangeal joints. Approximately one-third of patients have violaceous, flat-topped papules over the dorsal interphalangeal joints that are pathognomonic of dermatomyositis (Gottron's sign or Gottron's papules; Plate IIE-65). These lesions can be contrasted with the erythema and scaling on the dorsum of the fingers in some patients with SLE, which spares the skin over the interphalangeal joints. Periungual telangiectasia may be prominent, and a lacy or reticulated erythema may be associated with fine scaling on the extensor surfaces of the thighs and upper arms. Other patients, particularly those with long-standing disease, develop areas of hypopigmentation, hyperpigmentation, mild atrophy, and telangiectasia known as *poikiloderma vasculare atrophicans*. Poikiloderma is rare in both SLE and scleroderma and thus can serve as a clinical sign that distinguishes dermatomyositis from these two diseases. Cutaneous changes may be similar in scleroderma and dermatomyositis and may include thickening and binding down of the skin of the hands (sclerodactyly) as well as Raynaud's phenomenon. However, the presence of severe muscle disease, Gottron's papules, heliotrope erythema, and poikiloderma serve to distinguish patients with dermatomyositis. Skin biopsy of erythematous, scaling lesions of dermatomyositis may reveal only mild nonspecific inflammation but sometimes may show changes indistinguishable from those found in SLE, including epidermal atrophy, hydropic degeneration of basal keratinocytes, edema of the upper dermis, and a mild mononuclear cell infiltrate. Direct immunofluorescence microscopy of lesional skin is usually negative, although granular deposits of immunoglobulin(s) and complement in the epidermal basement membrane zone have been described in some patients. Treatment should be directed at the systemic disease. In the few instances where adjunctive cutaneous therapy is desirable, topical glucocorticoids are sometimes useful. These patients should avoid exposure to ultraviolet irradiation and use photoprotective measures such as sunscreens.

LUPUS ERYTHEMATOSUS

The cutaneous manifestations of lupus erythematosus (LE) (Chap. 311) can be divided into acute, subacute, and chronic (i.e., discoid LE) types. *Acute cutaneous LE* is characterized by erythema of the nose and malar eminences in a "butterfly" distribution (Plate IIE-61). The erythema is often sudden in onset, accompanied by edema and fine scale, and correlated with systemic involvement. Patients may have widespread involvement of the face as well as erythema and scaling of the extensor surfaces of the extremities and upper chest. These acute lesions, while sometimes evanescent, usually last for days and are often associated with exacerbations of systemic disease. Skin biopsy of acute lesions may show only a sparse dermal infiltrate of mononuclear cells and dermal edema. In some instances, cellular infiltrates around blood vessels and hair follicles are notable, as is hydropic degeneration of basal cells of the epidermis. Direct immunofluorescence microscopy of lesional skin frequently reveals deposits of immunoglobulin(s) and complement in the epidermal basement membrane zone. Treatment is aimed at control of systemic disease; photoprotection in this, as well as in other forms of LE, is very important.

Subacute cutaneous lupus erythematosus (SCLE) is characterized by a widespread photosensitive, nonscarring eruption. About half of these patients haveSLE in which severe renal and central nervous system involvement is uncommon. SCLE may present as a papulosquamous eruption that resembles psoriasis or annular lesions that resemble those seen in erythema multiforme. In the papulosquamous form, discrete erythematous papules arise on the back, chest, shoulders, extensor surfaces of the arms, and the dorsum of the hands; lesions are uncommon on the face, flexor surfaces of the arms, and below the waist. The slightly scaling papules tend to merge into large plaques, some with a reticulate appearance. The annular form involves the same areas and presents with erythematous papules that evolve into oval, circular, or polycyclic lesions. The lesions of SCLE are more widespread but have less tendency for scarring than do lesions of discoidLE. Skin biopsy reveals a dense mononuclear cell infiltrate around hair follicles and blood vessels in the superficial dermis, combined with hydropic degeneration of basal cells in the epidermis. Direct immunofluorescence microscopy of lesional skin reveals deposits of immunoglobulin(s) in the epidermal basement membrane zone in about half these cases. A particulate pattern of IgG deposition around basal keratinocytes has recently been associated with SCLE. Most SCLE patients have anti-Ro antibodies. Local therapy is usually unsuccessful, and most patients require treatment with aminoquinoline antimalarials. Low-dose therapy with oral glucocorticoids is sometimes necessary; photoprotective measures against both ultraviolet B and A wavelengths are very important.

Discoid lupus erythematosus (DLE) is characterized by discrete lesions, most often on the face, scalp, or external ears. The lesions are erythematous papules or plaques with a thick, adherent scale that occludes hair follicles (follicular plugging). When the scale is removed, its underside will show small excrescences that correlate with the openings of hair follicles and is termed a "carpet tack" appearance. This finding is relatively specific for DLE. Long-standing lesions develop central atrophy, scarring, and hypopigmentation but frequently have erythematous, sometimes raised borders at the periphery (Plate IIE-62). These lesions persist for years and tend to expand slowly. Only 5 to 10% of patients with DLE meet the American Rheumatism Association criteria forSLE. However, typical discoid lesions are frequently seen in patients with SLE. Biopsy of DLE lesions shows hyperkeratosis, follicular plugging, and atrophy of the epidermis. The dermal-epidermal junction reveals hydropic degeneration of basal keratinocytes, and a mononuclear cell infiltrate surrounding hair follicles and blood vessels. Direct immunofluorescence microscopy demonstrates immunoglobulin(s) and complement deposits at the basement membrane zone in about 90% of cases. Treatment is focused on control of local cutaneous disease and consists mainly of photoprotection and topical or intralesional glucocorticoids. If local therapy is ineffective, use of aminoquinoline antimalarials may be indicated.

SCLERODERMA AND MORPHEA

The skin changes of scleroderma (Chap. 313) usually begin on the hands, feet, and face, with episodes of recurrent nonpitting edema. Sclerosis of the skin begins distally on the fingers (sclerodactyly) and spreads proximally, usually accompanied by resorption of bone of the fingertips, which may have punched out ulcers, stellate scars, or areas of hemorrhage (Plate IIE-66). The fingers may actually shrink in size and become sausage-shaped, and since the fingernails are usually unaffected, the nails may curve over the end of the fingertips. Periungual telangiectasias are usually present, but periungual erythema is rare. In advanced cases, the extremities show contractures and calcinosis cutis. Face involvement includes a smooth, unwrinkled brow, taut skin over the nose, shrinkage of tissue around the mouth, and perioral radial furrowing (Plate IIE-64). Matlike telangiectasias are often present, particularly on the face and hands. Involved skin feels indurated, smooth, and bound to underlying structures; hyperpigmentation and hypopigmentation are also often present. Raynaud's phenomenon, i.e., cold-induced blanching, cyanosis, and reactive hyperemia, is present in almost all patients and can precede development of scleroderma by many years. The combination of calcinosis cutis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia has been termed the *CREST syndrome*. Anticentromere antibodies have been reported in a very high percentage of patients with the CREST syndrome but in only a small minority of patients with scleroderma. Skin biopsy reveals thickening of the dermis and homogenization of collagen bundles. Direct immunofluorescence microscopy of lesional skin is usually negative.

Morphea, which has been called *localized scleroderma*, is characterized by localized thickening and sclerosis of skin, usually affecting young adults or children. Morphea begins as erythematous or flesh-colored plaques that become sclerotic, develop central hypopigmentation, and demonstrate an erythematous border. In most cases, patients have one or a few lesions, and the disease is termed *localized morphea*. In some patients, widespread cutaneous lesions may occur, without systemic involvement. This form is called *generalized morphea*. Most patients with morphea do not have autoantibodies. Skin biopsy of morphea is indistinguishable from that of scleroderma. Linear scleroderma is a limited form of disease that presents in a linear, bandlike distribution and tends to involve deep as well as superficial layers of skin. Scleroderma and morphea are usually quite resistant to therapy. For this reason, physical therapy to prevent joint contractures and to maintain function is employed and is often helpful.

Diffuse fasciitis with eosinophilia is a clinical entity that can sometimes be confused with scleroderma. There is usually the sudden onset of swelling, induration, and erythema of the extremities frequently following significant physical exertion. The proximal portions of extremities (arms, forearms, thighs, legs) are more often involved than are the hands and feet. While the skin is indurated, it is usually not bound down as in scleroderma; contractures may occur early secondary to fascial involvement. The latter may also cause muscle groups to be separated (i.e., the "groove sign") and veins to appear depressed (i.e., sunken veins). These skin findings are accompanied by peripheral blood eosinophilia, increased erythrocyte sedimentation rate, and sometimes hypergammaglobulinemia. Deep biopsy of affected areas of skin reveals inflammation and thickening of the deep fascia overlying muscle. An inflammatory infiltrate composed of eosinophils and mononuclear cells is usually found. Patients with eosinophilic fasciitis appear to be at increased risk to develop bone marrow failure or other hematologic abnormalities. While the ultimate course of eosinophilic fasciitis is uncertain, many patients respond favorably to treatment with prednisone in doses ranging from 40 to 60 mg/d.

The *eosinophilia-myalgia syndrome*, a disorder reported in epidemic numbers in 1989 and linked to ingestion of L-tryptophan manufactured by a single company in Japan, is a multisystem disorder characterized by debilitating myalgias and absolute eosinophilia in association with varying combinations of arthralgias, pulmonary symptoms, and peripheral edema. In a later phase (i.e., 3 to 6 months after initial symptoms), these patients often develop localized sclerodermatous skin changes, weight loss, and/or neuropathy (Chap. 313). The precise cause of this syndrome, which may resemble other sclerotic skin conditions, is unknown. However, the implicated lots of L-tryptophan contained the contaminant 1,1-ethylidene bis[tryptophan]. This contaminant may be pathogenic or a marker for another substance that provokes the disorder.

(Bibliography omitted in Palm version)

Back to Table of Contents

59. CUTANEOUS DRUG REACTIONS - *Robert S. Stern*, *Olivier M. Chosidow*, *Bruce U. Wintroub*

Cutaneous reactions are among the most frequent adverse reactions to drugs. Prompt recognition of these reactions, drug withdrawal, and appropriate therapeutic interventions can minimize toxicity. This chapter focuses on adverse cutaneous reactions to drugs other than topical agents and reviews the incidence, patterns, and pathogenesis of cutaneous reactions to drugs and other therapeutic agents.

USE OF PRESCRIPTION DRUGS IN THE UNITED STATES

More than 1.5 billion prescriptions for 60,000 drug products, which include over 2000 different active agents, are dispensed each year in the United States. Hospital inpatients alone annually receive about 120 million courses of drug therapy, and half of adult Americans receive prescription drugs on a regular outpatient basis. Many additional patients use over-the-counter medicines that may cause adverse cutaneous reactions.

INCIDENCE OF CUTANEOUS REACTIONS

Although adverse drug reactions are common, it is difficult to ascertain their incidence, seriousness, and ultimate health effects. Available information comes from evaluations of hospitalized patients, epidemiologic surveys, premarketing studies, and voluntary reporting, most notably to the U.S. Food and Drug Administration's Medwatch System. None of these efforts provides comprehensive comparable data on the risk of cutaneous reactions associated with most medicines.

In one study about 2% of medical inpatients had skin reactions consisting of rash, urticaria, or pruritus during hospitalization. The overall reaction rate per course of drug therapy was about 3:1000. Among inpatients, penicillins, sulfonamides, and blood products accounted for two-thirds of cutaneous reactions. Among outpatients, reaction rates for many antibiotics were comparable to those observed in inpatients. Fluoroquinolones are notable causes of cutaneous reactions not observed in earlier studies. Reaction rates for selected commonly used antibiotics are summarized in Table 59-1. Most cutaneous reactions occur within 2 weeks of exposure to a drug. The risk of allergic reactions does not vary greatly with age or sex. Among outpatients, the risk of a reaction to an antibiotic was comparable for first and subsequent courses of a given drug.

The distribution of morphologic patterns of drug eruptions cared for within a Finnish hospital dermatology department with a special interest in fixed drug eruptions included exanthematous reactions (32%), urticaria and/or angioedema (20%), fixed drug eruptions (34%), erythema multiforme (2%), Stevens-Johnson syndrome (SJS; 1%), exfoliative dermatitis (1%), and photosensitivity reactions (3%). Other studies suggest that about 80% of all cutaneous reactions are morbilliform or erythematous, 10 to 15% are urticaria or angioedema, and all other types of reactions are relatively rare.

The relative risk ofSJS and toxic epidermal necrolysis (TEN), perhaps the most important severe cutaneous reactions, has been quantified in an international case control study and case series. Sulfonamide antibiotics, allopurinol, amine antiepileptic drugs (phenytoin and carbamazepine), and lamotrigine (a new antiepileptic) are associated with the highest risk of these reactions.

PATHOGENESIS OF DRUG REACTIONS

Untoward cutaneous responses to drugs can arise as a result of immunologic or nonimmunologic mechanisms. Immunologic reactions require activation of host immunologic pathways and are designated *drug allergy*. Drug reactions occurring through nonimmunologic mechanisms may be due to activation of effector pathways, overdosage, cumulative toxicity, side effects, ecologic disturbance, interactions between drugs, metabolic alterations, exacerbation of preexisting dermatologic conditions, or inherited protein or enzyme deficiencies. It is often not possible to specify the responsible drug or pathogenic mechanism because the skin responds to a variety of stimuli through a limited number of reaction patterns. The mechanism of many drug reactions is unknown.

IMMUNOLOGIC DRUG REACTIONS

Drugs frequently elicit an immune response, but only a small number of individuals experience clinical hypersensitivity reactions. For example, most patients exposed to penicillin develop demonstrable antibodies to penicillin but do not manifest drug reactions when exposed to penicillin. Multiple factors determine the capacity of a drug to elicit an immune response, including the molecular characteristics of the drug and host effects.

Increases in *molecular* size and complexity are associated with increased immunogenicity, and macromolecular drugs such as protein or peptide hormones are highly antigenic. Most drugs are small organic molecules <1000 Da in size, and the capacity of such small molecules to elicit an immune response depends on their ability to act as haptens, i.e., to form stable, usually covalent, bonds with tissue macromolecules, an extremely rare event.

Route of administration of a drug or simple chemical can influence the nature of the *host* immune response. For example, topical application of antigens tends to induce delayed hypersensitivity, and exposure to antigens via oral or nasal cavities stimulates production of secretory immunoglobins, IgA and IgE, and occasionally IgM. Frequency of sensitization through intravenous administration of drugs varies, but anaphylaxis is a more likely consequence with this route of exposure than following oral administration.

The degree of drug exposure and individual variability in absorption and metabolism of a given agent may alter immunogenic load. The variable degree of in vivo acetylation of hydralazine provides a clinical example of this phenomenon. Hydralazine produces a lupus-like syndrome associated with antinuclear antibody formation more frequently in patients who acetylate the drug slowly. Frequent high-dose and interrupted courses of therapy are also important risk factors for development of drug allergy.

Pathogenesis of Allergic Drug Reactions> >

> >*IgE-Dependent Reactions* IgE-dependent drug reactions are usually manifest in the

skin and gastrointestinal, respiratory, and cardiovascular systems (Chap. 310). Primary symptoms and signs include pruritus, urticaria, nausea, vomiting, cramps, bronchospasm, and laryngeal edema and, on occasion, anaphylactic shock with hypotension and death. Immediate reactions may occur within minutes of drug exposure, and accelerated reactions occur hours or days after drug administration. Accelerated reactions are usually urticarial and may include laryngeal edema. Penicillin and related drugs are the most frequent causes of IgE-dependent reactions. Release of chemical mediators such as histamine, adenosine, leukotrienes, prostaglandins, platelet-activating factor, enzymes, and proteoglycans from sensitized tissue, mast cells, or circulating basophilic leukocytes results in vasodilation and edema. Release is triggered when polyvalent drug protein conjugates cross-link IgE molecules fixed to sensitized cells. The clinical manifestations are determined by interaction of the released chemical mediator with its target organ, i.e., skin, respiratory, gastrointestinal, and/or cardiovascular systems. Certain routes of administration favor different clinical patterns (i.e., oral route: gastrointestinal effects; intravenous route: circulatory effects).

Immune-Complex-Dependent Reactions Serum sickness is produced by circulating immune complexes and is characterized by fever, arthritis, nephritis, neuritis, edema, and an urticarial, papular, or purpuric rash (Chap. 317). The syndrome requires an antigen that remains in the circulation for prolonged periods so that when antibody is synthesized, circulating antigen-antibody complexes are formed. Serum sickness was first described following administration of foreign sera, but drugs are now the usual cause. Drugs that produce serum sickness include the penicillins, sulfonamides, thiouracils, cholecystographic dyes, phenytoin, aminosalicylic acid, heparin, and antilymphocyte globulin. Cephalosporin administration in febrile children is associated with a high risk of a clinically similar reaction, but the mechanism of this reaction is unknown. In classic serum sickness, symptoms develop 6 days or more after exposure to a drug, the latent period representing the time needed to synthesize antibody. The antibodies responsible for immune-complex-dependent drug reactions are largely of the IgG or IgM class. Vasculitis, a relatively rare cutaneous complication of drugs, may also be a result of immune complex deposition (Chap. 317).

Cytotoxicity and Delayed Hypersensitivity Cytotoxicity and delayed hypersensitivity mechanisms may be important in the etiology of morbilliform exanthema, hypersensitivity syndrome,SJS, orTEN, but this is not proven. Systemic manifestations occur frequently. The nature of the antigen leading to cytotoxic reactions is unknown, but it is likely that different T lymphocyte populations are activated. T$_H$1 type cells will lead to the production of interleukin (IL)-2 and interferon (IFN)-g and subsequent activation of cytotoxic T cells. In early lesions of morbilliform exanthema or TEN, histopathologic studies have shown expression of HLA-DR and intercellular adhesion molecule (ICAM-1) by keratinocytes, CD4 cells (in the dermis), and CD8 T cells (in the epidermis) and apoptosis of keratinocytes (facilitated by tumor necrosis factora secretion and *fas*-ligand expression). T$_H$2 type cells produce cytokines such as IL-5, which may be involved in hypersensitivity syndrome (see below).

NONIMMUNOLOGIC DRUG REACTIONS

Nonimmunologic mechanisms are responsible for the majority of drug reactions; however, only the most important mechanisms will be discussed.

Nonimmunologic Activation of Effector Pathways Drug reactions may result from nonimmunologic activation of effector pathways by three mechanisms: First, drugs may release mediators directly from mast cells and basophils and present as anaphylaxis, urticaria, and/or angioedema. Urticarial anaphylactic reactions induced by opiates, polymyxin B, tubocurarine, radiocontrast media, and dextrans may occur by this mechanism. Second, drugs may activate complement in the absence of antibody. This is an additional mechanism through which radiocontrast media may act. Third, drugs such as aspirin and other nonsteroidal anti-inflammatory agents (NSAIDs) may alter pathways of arachidonic acid metabolism and induce urticaria.

Phototoxicity Phototoxic reactions may be drug-induced or may occur in metabolic disorders in which a photosensitizing chemical is overproduced. A phototoxic reaction occurs when enough chromophore (drug or metabolic product) absorbs sufficient radiation to cause a reaction or interaction with target tissue. Drug-induced phototoxic reactions can occur on first exposure. The incidence of phototoxicity is a direct function of the concentration of sensitizer and the amount of light of the appropriate wavelengths. At least three distinct photochemical mechanisms have been described: (1) the reaction between the excited state of a phototoxic molecule and a biologic target may cause formation of a covalent photoaddition product, (2) the phototoxic molecule may form stable photoproducts that are toxic to biologic substrates, and (3) radiation of a phototoxic molecule may result in transfer of energy to oxygen molecules and cause formation of toxic oxygen species, such as singlet oxygen superoxide anion, or hydroxyl radicals. Interaction of these reactive species with biologic targets produces photooxidized molecules. Phototoxic injury is usually manifest as a sunburn-like reaction.

Exacerbation of Preexisting Diseases A variety of agents can exacerbate preexisting diseases. For example, lithium can exacerbate acne and psoriasis in a dose-dependent manner. Beta-blocking agents andIFN-amay induce psoriasis. Withdrawal of glucocorticoids can exacerbate psoriasis or atopic dermatitis.

Inherited Enzyme or Protein Deficiencies Specific genetically determined defects in the ability of an individual to detoxify toxic reactive drug metabolites may predispose such individuals to the development of severe drug reactions, especially hypersensitivity syndrome, and perhapsTENassociated with use of sulfonamides and anticonvulsants.

Alterations of Immunologic Status Alterations in patients' immunologic status may also modify the risk of cutaneous reactions. Bone marrow transplant patients, HIV-infected persons, and persons with Epstein-Barr virus infection are at higher risk of developing cutaneous reactions to drugs. Skin reactions to trimethoprim-sulfamethoxazole are seen in about a third of HIV-infected users of this drug, but desensitization can be accomplished. Dapsone, trimethoprim alone, and amoxicillin-clavulanate are also frequent causes of drug eruptions in HIV-infected patients. The advent of highly active antiretroviral therapy (HAART) may have decreased the risk of cutaneous reactions in HIV patients (Chap. 309).

A CLINICAL CLASSIFICATION OF CUTANEOUS DRUG REACTIONS

URTICARIA/ANGIOEDEMA

Urticaria is a skin reaction characterized by pruritic, red wheals. Lesions may vary from a small point to a large area. Individual lesions rarely last more than 24 h. When deep dermal and subcutaneous tissues are also swollen, this reaction is known as *angioedema*. Angioedema may involve mucous membranes and may be part of a life-threatening anaphylactic reaction. Urticarial lesions, along with pruritus and morbilliform (or maculopapular) eruptions, are among the most frequent types of cutaneous reactions to drugs.

Drug-induced urticaria may be caused by three mechanisms: an IgE-dependent mechanism, circulating immune complexes (serum sickness), and nonimmunologic activation of effector pathways. IgE-dependent urticarial reactions usually occur within 36 h but can occur within minutes. Reactions occurring within minutes to hours of drug exposure are termed *immediate reactions*, whereas those that occur 12 to 36 h after drug exposure are designated *accelerated reactions*. Immune-complex-induced urticaria associated with serum sickness usually occurs from 6 to 12 days after first exposure. In this syndrome, the urticarial eruption may be accompanied by fever, hematuria, arthralgias, hepatic dysfunction, and neurologic symptoms.

Certain drugs, such asNSAIDs, angiotensin-converting enzyme (ACE) inhibitors, and radiographic dyes, may induce urticarial reactions, angioedema, and anaphylaxis in the absence of drug-specific antibody. Although ACE inhibitors, aspirin, penicillin, and blood products are the most frequent causes of urticarial eruptions, urticaria has been observed in association with nearly all drugs. Drugs also may cause chronic urticaria, which lasts more than 6 weeks. Aspirin frequently exacerbates this problem.

The treatment of urticaria or angioedema depends on the severity of the reaction and the rate at which it is evolving. In severe cases, especially with respiratory or cardiovascular compromise, epinephrine is the mainstay of therapy, but its effect is reduced in patients using beta blockers. For more seriously affected patients, treatment with systemic glucocorticoids, sometimes intravenously administered, are helpful. In addition to drug withdrawal, for patients with only cutaneous symptoms and without symptoms of angioedema or anaphylaxis, oral antihistamines are usually sufficient.

PHOTOSENSITIVITY ERUPTIONS

Photosensitivity eruptions are usually most marked in sun-exposed areas but may extend to sun-protected areas. Phototoxic reactions are more common with some drugs. Photoallergic reactions to systemically administered drugs are very rare. Phototoxic reactions usually resemble sunburn and can occur with the first exposure to a drug. Their severity depends on the tissue level of the drug, the extent of exposure to light, and the efficiency of the photosensitizer (Chap. 60).

Orally administered phototoxic drugs include many fluoroquinolones, chlorpromazine, tetracycline, thiazides, and at least twoNSAIDs(benoxaprofen and piroxicam). The majority of the common phototoxic drugs have action spectrums in the long-wave ultraviolet A (UV-A) range. Phototoxic reactions abate with removal of either the drug or ultraviolet radiation. Because UV-A and visible light, which trigger these reactions, are not easily absorbed by nonopaque sunscreens and are transmitted through window glass, these reactions may be difficult to block.

Photosensitivity reactions are treated by avoiding exposure to ultraviolet light (sunlight) and treating the reaction as one would a sunburn. Rarely, individuals develop persistent reactivity to light, necessitating long-term avoidance of sun exposure.

PIGMENTATION CHANGES

Drugs may cause a variety of pigmentary changes in the skin. Some drugs stimulate melanocytic activity and increase pigmentation. Drug deposition can also lead to pigmentation; this phenomenon occurs with heavy metals. Phenothiazines may be deposited in the skin and cause a slate-gray color. Antimalarial drugs may cause a slate-gray or yellow pigmentation. Long term minocycline use may cause slate-gray hyperpigmentation, especially in areas of chronic inflammation. Inorganic arsenic, once used to treat psoriasis, is associated with diffuse macular pigmentation. Other heavy metals that cause pigmentary changes include silver, gold, bismuth, and mercury. Long-term use of phenytoin can produce a chloasma-like pigmentation in women. Certain cytostatic agents can also cause pigmentary changes. Histologic examination is often diagnostic for drug deposition diseases.

Zidovudine (AZT) is a frequent cause of pigmentation, especially of the nails (Chap. 309). Nicotinic acid in large doses may cause brown pigmentation, and oral contraceptives may produce chloasma. In addition, amiodarone may cause violaceous hyperpigmentation that is increased in sun-exposed skin. Drugs such as heavy metals, copper antimalarial and arsenical agents, and ACTH also may discolor oral mucosa.

VASCULITIS

Cutaneous necrotizing vasculitis often presents as palpable purpuric lesions that may be generalized or limited to the lower extremities or other dependent areas (Chap. 317). Urticarial lesions, ulcers, and hemorrhagic blisters also occur. Vasculitis may involve other organs, including the liver, kidney, brain, and joints. Drugs are only one cause of vasculitis, with infection and collagen vascular disease responsible for the majority of cases.

Propylthiouracil induces a cutaneous vasculitis that is accompanied by leukopenia and splenomegaly. Direct immunofluorescent changes in these lesions suggest immune-complex deposition. Drugs implicated in vasculitic eruptions include allopurinol, thiazides, sulfonamides, penicillin, and someNSAIDs.

HYPERSENSITIVITY SYNDROME

Initially described with phenytoin, hypersensitivity syndrome presents as an erythematous eruption that may become purpuric and is accompanied by many of the following features: fever, facial and periorbital edema, tender generalized lymphadenopathy, leukocytosis (often with atypical lymphocytes and eosinophils), hepatitis, and sometimes nephritis or pneumonitis. The cutaneous reaction usually begins 1 to 6 weeks after phenytoin is begun and usually resolves with drug cessation, but symptoms, especially hepatitis, may persist. The eruption recurs with rechallenge, and cross-reactions among aromatic anticonvulsants, including phenytoin, carbamazepine, and barbiturates, are frequent. With phenytoin, an increased risk of this syndrome is associated with an inherited deficiency of epoxide hydrolase, an enzyme required for metabolism of a toxic intermediate arene oxide that is formed during metabolism of phenytoin by the cytochrome P450 system. Other drugs causing this syndrome include lamotrigine, dapsone, allopurinol, sulfonamides, minocycline, and sulfones. Systemic glucocorticoids (prednisone, 0.5 to 1.0 mg/kg) seem to reduce symptoms. Mortality as high as 10% has been reported.

WARFARIN NECROSIS OF THE SKIN

This rare reaction occurs usually between the third and tenth days of therapy with warfarin derivatives, usually in women. Lesions are sharply demarcated, erythematous, indurated, and purpuric and may resolve or progress to form large, irregular, hemorrhagic bullae with eventual necrosis and slow-healing eschar formation.

Development of the syndrome is unrelated to drug dose or underlying condition. Favored sites are breasts, thighs, and buttocks. The course is not altered by discontinuation of the drug after onset of the eruption. Similar reactions have been associated with heparin. Warfarin reactions are associated with protein C deficiency. Protein C is a vitamin K-dependent protein with a shorter half-life than other clotting proteins and is in part responsible for control of fibrinolysis. Since warfarin inhibits synthesis of vitamin K-dependent coagulation factors, warfarin anticoagulation in heterozygotes for protein C deficiency causes a precipitous fall in circulating levels of protein C, permitting hypercoagulability and thrombosis in the cutaneous microvasculature, with consequent areas of necrosis. Heparin-induced necrosis may have clinically similar features but is probably due to heparin-induced platelet aggregation with subsequent occlusion of blood vessels.

Warfarin-induced cutaneous necrosis is treated with vitamin K and heparin. Vitamin K reverses the effects of warfarin, and heparin acts as an anticoagulant. Treatment with protein C concentrates may also be helpful in individuals with deficiencies of protein C, the predisposing factor for development of these reactions.

MORBILLIFORM REACTIONS

Morbilliform or maculopapular eruptions are the most common of all drug-induced reactions, often start on the trunk or areas of pressure or trauma, and consist of erythematous macules and papules that are frequently symmetric and may become confluent. Involvement of mucous membranes, palms, and soles is variable; the eruption may be associated with moderate to severe pruritus and fever.

The pathogenesis is unclear. A hypersensitivity mechanism has been suggested, although these reactions do not always recur following drug rechallenge. Diagnosis is rarely assisted by laboratory or patch testing; differentiation from viral exanthem is the principal differential diagnostic consideration. Unless the suspect drug is essential it should be discontinued. Occasionally these eruptions may decrease or fade with continued use of the responsible drug.

Morbilliform reactions usually develop within 1 week of initiation of therapy and last 1 to 2 weeks; however, reactions to some drugs, especially penicillin and drugs with long half-lives, may begin more than 2 weeks after therapy has begun and last as long as 2 weeks after therapy has ceased.

Morbilliform eruptions are usually treated by discontinuing the suspect medications symptomatically. Oral antihistamines, emollients, and soothing baths are useful for treatment of pruritus. Short courses of potent topical glucocorticoids can reduce inflammation and symptoms and are probably helpful. The beneficial effect of systemic glucocorticoids relative to risk is less clear.

FIXED DRUG REACTIONS

These reactions are characterized by one or more sharply demarcated, erythematous lesions in which hyperpigmentation results after resolution of the acute inflammation; with rechallenge, the lesion recurs in the same (i.e., "fixed") location. Lesions often involve the lips, hands, legs, face, genitalia, and oral mucosa and cause burning. Most patients have multiple lesions. Patch testing is useful to establish the etiology. Fixed drug eruptions have been associated with phenolphthalein, sulfonamides, tetracyclines, phenylbutazone,NSAIDs, and barbiturates. Although cross-sensitivity appears to occur between different tetracycline compounds, cross-sensitivity was not elicited when different sulfonamide compounds were administered to patients as part of provocation testing.

LICHENOID DRUG ERUPTIONS

A lichenoid cutaneous reaction, clinically and morphologically indistinguishable from lichen planus, is associated with a variety of drugs and chemicals. Eosinophils are more common when the reaction is drug-induced. Gold and antimalarials are most often associated with this eruption. Antihypertensive agents, including beta blockers and captopril, have also been reported to cause lichenoid reactions.

BULLOUS ERUPTIONS

Blisters accompany a wide variety of cutaneous reactions, including fixed drug eruptions, severe morbilliform eruptions in dependent areas of the body, and phototoxic reactions.SJS andTEN are the most serious and important bullous reactions to drugs. Nalidixic acid and furosemide cause blistering eruptions indistinguishable from the primary bullous diseases. A pemphigus foliaceus-like eruption is seen with penicillamine.

PUSTULAR ERUPTIONS

Acute generalized exanthematous pustulosis is often associated with exposure to drugs, most notably antibiotics. Usually beginning on the face or intertriginous areas, small nonfollicular pustules overlying erythematous and edematous skin may coalesce and lead to superficial ulceration. Fever is present and differentiating this eruption fromTEN in its initial stages may be difficult. Acute generalized exanthematous pustulosis often begins within a few days of initiating drug treatment.

ERYTHEMA MULTIFORME

Erythema multiforme is an acute, self-limited inflammatory disorder of skin and mucous membranes characterized by distinctive iris or target lesions, usually acrally distributed and often associated with sore throat, mucosal lesions, and malaise. Classic erythema multiforme usually has nondrug causes, most commonly herpes simplex infection, and must be differentiated from trueSJS, which is usually drug related.

STEVENS-JOHNSON SYNDROME

SJSis a blistering disorder that is usually more severe than erythema multiforme. Initial presentation is often a sore throat, malaise, and fever. Within a few days, in addition to erosions of multiple mucous membranes, small blisters developing on dusky or purpuric macules or atypical target lesions characterize this eruption. Total percent of body surface area blistering and eventual detachment is less than 10%. Overlap SJS/TENshares characteristics of both SJS and TEN, with 10 to 30% of body surface area exhibiting epidermal detachment.

TOXIC EPIDERMAL NECROLYSIS

TENis the most serious cutaneous drug reaction and may be fatal. Drugs are usually the cause of TEN. Onset is generally acute and is characterized by fever>39°C (102.2°F), blisters or ulcers of multiple mucous membranes, malaise, and epidermal necrosis involving>30% of body surface area. Intestinal and pulmonary involvement is associated with a poor prognosis, as is a greater extent of epidermal detachment and older age. About 30% of affected persons die. Many treatments affecting immune response or cytokines (thalidomide) or apoptosis (intravenous immunoglobulin) have been advocated, but none have been shown to be efficacious in well-controlled trials. In spite of its theoretical potential benefits, thalidomide therapy increases TEN-associated mortality. Supportive treatment in burn units is helpful in reducing morbidity and mortality.

DRUGS OF SPECIAL INTEREST

PENICILLIN

The incidence of cutaneous reactions to penicillin is about 1%. About 85% of cutaneous reactions to penicillin are morbilliform, and about 10% are urticaria or angioedema.

IgG, IgM, and IgE antibodies can be produced; IgG and IgM anti-penicillin antibodies play a role in the development of hemolytic anemia, whereas anaphylaxis and serum sickness appear to be due to IgE antibodies in serum.

In patients with suspected IgE-mediated reactions to penicillin for whom future treatment is anticipated, accurate tests for sensitization are available. Current practice is to perform skin testing with a commercially available penicilloyl determinant preparation (Pre-pen, Kremers-Urban) and with fresh penicillin and, if possible, with another source of minor (nonpenicilloyl) determinants such as aged or base-treated penicillin. Antibodies to minor determinants are common in patients experiencing anaphylaxis, but testing with major determinants alone detects most patients at risk for anaphylaxis.

About one-fourth of patients with positive history of penicillin allergy have a positive skin test, while 6% (3 to 10%) with no history of penicillin sensitivity demonstrate a positive skin response to penicillin. Administering penicillin to those patients with a positive skin test produces reactions in a high proportion (50 to 100%); conversely, only a few patients (0.5%) with a negative skin test react to the drug, and reactions tend to be mild and to occur late. Since a false-negative skin test may occur during or just after an acute reaction, testing should be performed either prospectively or several months after a suspected reaction. As many as 80% of patients lose anaphylactic sensitivity and IgE antibody after several years. Radioallergosorbent tests and other in vitro tests offer no advantage over properly performed skin testing. Some cross-reactivity between penicillin and nonpenicillin b-lactam antibiotics (e.g., cephalosporins) occurs, but the majority of penicillin-allergic patients will tolerate cephalosporins. Persons who have negative skin tests to penicillin rarely develop reactions to cephalosporins.

In the face of a positive clinical history of penicillin reaction, another drug should be chosen. If this is not feasible or prudent (e.g., in a pregnant patient with syphilis or with enterococcal endocarditis), skin testing with penicillin is warranted. If skin tests are negative, cautious administration of penicillin is acceptable, although some recommend desensitization of such patients if the reaction was likely to be IgE-mediated. In those with positive skin tests, desensitization is mandatory if therapeutic use ofb-lactam antibiotics is to be undertaken. Various protocols are available, including oral and parenteral approaches. Oral desensitization appears to have lower risk of serious anaphylactic reactions during desensitization. However, desensitization carries the risk of anaphylaxis regardless of how it is performed. After desensitization, many patients experience non-life-threatening IgE-mediated untoward reactions to penicillin during their course of therapy. Desensitization is not effective in those with exfoliative dermatitis or morbilliform reactions due to penicillin.

NONSTEROIDAL ANTI-INFLAMMATORY DRUGS

NSAIDs, including aspirin and indomethacin (indometacin), cause two broad categories of allergic-like symptoms in susceptible individuals: (1) approximately 1% of persons experience urticaria or angioedema, and (2) about half as many (0.5%) experience rhinosinusitis and asthma; however, about 10% of adults with asthma and one-third of individuals with nasal polyposis and sinusitis may respond adversely to aspirin.

Urticaria/angioedema may be delayed up to 24 h and may occur at any age. The rhinosinusitis-asthma syndrome generally develops within 1 h of drug administration. In young patients, the reaction pattern often begins as watery rhinorrhea, which can be complicated by nasal and sinus infection, and polyposis, bloody discharge, and nasal eosinophilia. In many individuals with this syndrome, asthma that can be life-threatening eventually ensues wheneverNSAIDsare subsequently ingested, and symptoms may persist despite avoidance of these drugs. Proof of the association of symptoms and NSAID use requires either clear-cut history of symptoms following drug ingestion or an oral challenge. For the latter to be performed with relative safety, (1) asthma must be under good control, (2) the procedure must be conducted in a hospital setting by experienced personnel capable of recognizing and treating acute respiratory responses, and (3) the challenge should begin with very low doses (i.e., not>30 mg) of aspirin and increase every 1 to 2 h in doubling doses as tolerated to 650 mg.

While cross-reactivity betweenNSAIDsis common, it is not immunologic, and patients who are sensitive to NSAIDs cannot be identified by assessment of IgE antibody to aspirin, lymphocyte sensitization, or in vitro immunologic testing.

RADIOCONTRAST MEDIA

Large numbers of patients are exposed to radiocontrast agents. High-osmolality radiocontrast media are about five times more likely to induce urticaria (1%) or anaphylaxis than newer low-osmolality media. Severe reactions are rare with either type of contrast media. About one-third of those with mild reactions to previous exposure rereact on reexposure. In most cases, these reactions are probably not immunologic. Pretreatment with prednisone and diphenhydramine reduces reaction rates. Persons with a reaction to a high-osmolality contrast media should be given low-osmolality media if later contrast studies are required.

ANTICONVULSANTS

Of the anticonvulsants, the single orally administered agent with the highest risk of severe adverse cutaneous reactions is the antiseizure medicine lamotrigine. Older anticonvulsants, including phenytoin and carbamazepine, are also associated with many types of severe reactions and a high incidence of less severe reactions, particularly in children. In addition toSJS,TEN, and the hypersensitivity syndrome discussed above, the aromatic anticonvulsants can induce a pseudolymphoma syndrome and induce gingival hyperplasia.

SULFONAMIDES

Sulfonamides have perhaps the highest risk of causing cutaneous eruptions and are the drugs most frequently implicated inSJS andTEN. The combination of sulfamethoxazole and trimethoprim frequently induces adverse cutaneous reactions in patients with AIDS (Chap. 309). Desensitization is often successful in AIDS patients with morbilliform eruptions but is a high-risk procedure in AIDS patients who manifest erythroderma, fever, or a bullous reaction in response to their earlier sulfonamide exposure.

AGENTS USED IN CANCER CHEMOTHERAPY

Since many agents used in cancer chemotherapy inhibit cell division, rapidly proliferating elements of the skin, including hair, mucous membranes, and appendages, are sensitive to their effects; as a result, stomatitis and alopecia are among the most frequent dose-dependent side effects of chemotherapy. Onychodystrophy (dystrophic changes in nails) is also seen with bleomycin, hydroxyurea (hydroxycarbamide), and 5-fluorouracil. Sterile cellulitis and phlebitis and ulceration of pressure areas occur with many of these agents. Urticaria, angioedema, exfoliative dermatitis, and erythema of the palms and soles have also been seen, as has local and diffuse hyperpigmentation.

GLUCOCORTICOIDS

Both systemic and topical glucocorticoids cause a variety of skin changes, including acneiform eruptions, atrophy, striae, and other stigmata of Cushing's syndrome, and in sufficiently high doses can retard wound healing. Patients using glucocorticoids are at higher risk for bacterial, yeast, and fungal skin infections that may be misinterpreted as drug eruptions but are instead drug side effects.

CYTOKINE THERAPY

Alopecia is a common complication ofIFN-a. Induction or exacerbation of various immune-mediated disorders (psoriasis, lichen planus, lupus erythematosus) has been also reported with this agent. IFN-b injection has been associated with local necrosis of the skin. Granulocyte colony stimulating factor may induce various neutrophilic dermatosis, including Sweet's syndrome, pyoderma gangrenosum, neutrophilic eccrine hidradenitis, and vasculitis, and can exacerbate psoriasis.

IL-2 is associated with frequent cutaneous reactions including exanthema, facial edema, xerosis, and pruritus. Cases of pemphigus vulgaris, linear IgA disease, psoriasis, and vitiligo have also been described in association with this drug.

ANTIMALARIAL AGENTS

Antimalarial agents are used as therapy for several skin diseases, including the skin manifestations of lupus and polymorphous light eruption, but they can also induce cutaneous reactions. Although also used to treat porphyria cutanea tarda at low doses, in patients with asymptomatic porphyria cutanea tarda, higher doses of chloroquine increase porphyrin levels to such an extent that they may exacerbate the disease.

Pigmentation disturbances, including black pigmentation of the face, mucous membranes, and pretibial and subungual areas, occur with antimalarials. Quinacrine (mepacrine) causes generalized, cutaneous yellow discoloration.

GOLD

Chrysotherapy has been associated with a variety of dose-related dermatologic reactions (including maculopapular eruptions), which can develop as long as 2 years after initiation of therapy and require months to resolve. Erythema nodosum, psoriasiform dermatitis, vaginal pruritus, eruptions similar to those of pityriasis rosea, hyperpigmentation, and lichenoid eruptions resembling those seen with antimalarial agents have been reported. After a cutaneous reaction, it is sometimes possible to reinstitute gold therapy at lower doses without recurrence of the dermatitis.

DIAGNOSIS OF DRUG REACTIONS

Possible causes of an adverse reaction can be assessed as definite, probable, possible, or unlikely based on six variables: (1) previous experience with the drug in the general population, (2) alternative etiologic candidates, (3) timing of events, (4) drug levels or evidence of overdose, (5) patient reaction to drug discontinuation, and (6) patient reaction to rechallenge.

PREVIOUS EXPERIENCE

Tables of relative reaction rates are available and are useful to assess the likelihood that a given drug is responsible for a given cutaneous reaction. The specific morphologic pattern of a drug reaction, however, may modify these reaction rates by increasing or decreasing the likelihood that a given drug is responsible for a given reaction. For example, since fixed eruptions due to drugs are more often seen with barbiturates than with penicillin, a fixed drug reaction in a patient taking both types of agents is more likely to be due to the barbiturate, even though penicillins have a higher overall drug reaction rate.

ALTERNATIVE ETIOLOGIC CANDIDATES

A cutaneous eruption may be due to exacerbation of preexisting disease or to development of new disease unrelated to drugs. For example, a patient with psoriasis may have a flare-up of disease coincidental with administration of penicillin for streptococcal infection; in this case, infection is a more likely cause for the flare-up than drug reaction.

TIMING OF EVENTS

Most drug reactions of the skin occur within 1 to 2 weeks of initiation of therapy. Hypersensitivity syndrome may occur later (up to 8 weeks) after initiating drug therapy. Fixed drug reactions and generalized exanthematous pustulosis often occur earlier (within 48 h), as do reactions of all types in persons with prior sensitization to that drug or a cross-sensitizing agent.

DRUG LEVELS

Some cutaneous reactions are dependent on dosage or cumulative toxicity. For example, lichenoid dermatoses due to gold administration appear more often in patients taking high doses.

DISCONTINUATION

Most adverse cutaneous reactions to drugs remit with discontinuation of the suspected agent. A reaction is considered unlikely to be drug-related if improvement occurs while the drug is continued or if a patient fails to improve after stopping the drug and appropriate therapy.

RECHALLENGE

Rechallenge provides the most definitive information concerning adverse cutaneous reactions to drugs, since a reaction failing to recur on rechallenge with a drug is unlikely to be due to that agent. Rechallenge is usually impractical, however, because the need to ensure patient safety and comfort outweighs the value of the possible information derived from rechallenge.

Of special importance is the rapid recognition of reactions that may become serious or life-threatening.Table 59-2 lists clinical and laboratory features that, if present, suggest the reaction may be serious.Table 59-3 provides key features of the most serious adverse cutaneous reactions.

DIAGNOSIS OF DRUG ALLERGY

Tests for IgE responses include in vivo and in vitro methods, but such tests are available for only a limited number of drugs, including penicillins and cephalosporins, some peptide and protein drugs (insulin, xenogeneic sera), and some agents used for general anesthesia. In vivo testing is accomplished by prick puncture and/or by intradermal skin testing. A wheal-and-flare response 2´ 2 mm greater than that seen with a saline control within 20 min is considered indicative of IgE-mediated mast cell degranulation, provided (1) the patient is not dermographic, (2) the drug does not nonspecifically degranulate mast cells, (3) the drug concentration is not high enough to be irritating, and (4) the buffer itself does not cause wheal-and-flare responses.

Skin testing with major and minor determinants of penicillins or cephalosporins has proved useful for identifying patients at risk of anaphylactic reactions to these agents. However, skin tests themselves carry a small risk of anaphylaxis. Negative skin tests do not rule out IgE-mediated reactivity, and the risk of anaphylaxis in response to penicillin administration in patients with negative skin tests is about 1%; about two-thirds of patients with a positive skin test and history of a previous adverse reaction to penicillin experience an allergic response on rechallenge. Skin tests may be negative in allergic patients receiving antihistamines or in those whose allergy is to determinants not present in the test reagent. Although less well studied, similar techniques can identify patients who are sensitive to protein drugs and to agents such as gallamine and succinylcholine. Most other drugs are small molecules, and skin testing with them is unreliable.

There are no generally available and reliable tests for assessing causality of non-IgE-mediated reactions, except possibly patch tests for assessment of fixed drug reactions. Therefore, diagnosis usually relies on clinical factors rather than test results.

(Bibliography omitted in Palm version)

Back to Table of Contents

60. PHOTOSENSITIVITY AND OTHER REACTIONS TO LIGHT - *David R. Bickers*

SOLAR RADIATION

Sunlight is the most visible and obvious source of comfort in the environment. This natural proclivity for the sun has the beneficial results of warmth and vitamin D synthesis but also can produce pathologic consequences. Few effects of sun exposure beyond those affecting the skin have been identified, but cutaneous exposure to sunlight can evoke immunosuppressive responses and genetic changes that may be relevant to the pathogenesis of nonmelanoma skin cancer and perhaps infections such as herpes simplex.

The sun's energy encompasses a broad range from ultrashort highly energetic ionizing radiation (10^{-2}um) to ultralong radiowaves of very low photon energy (10^{7}um). Thus, the emission spectrum ranges over nine orders of magnitude, but that reaching the earth's surface is narrow and is limited to components of the ultraviolet (UV), visible light, and portions of the infrared. The cutoff at the short end of the UV is at approximately 290 nm, because stratospheric ozone is formed by ionizing radiation of wavelengths less than 100 nm and absorbs solar energy between 120 and 310 nm, thereby preventing penetration to the earth's surface of the shorter, more energetic, potentially more harmful wavelengths of solar radiation. Indeed, concern about destruction of the ozone layer by chlorofluorocarbons released into the atmosphere has led to international agreements to reduce production of these chemicals.

Measurements of solar flux indicate that there is a twentyfold regional variation in the amount of energy at 300 nm that reaches the earth's surface. This variability relates to seasonal effects, the path of sunlight transmission through ozone and air, the altitude (4% increase for each 300 m of elevation), the latitude (increasing intensity with decreasing latitude), and the amount of cloud cover, fog, and pollution.

The major components of the photobiologic action spectrum include theUV and visible wavelengths between 290 and 700 nm. In addition, the wavelengths beyond 700 nm in the infrared primarily evoke heat, but warming of the skin may enhance biologic responses to wavelengths in the UV and visible spectrum.

TheUVspectrum is arbitrarily divided into three major segments: C, B, and A. This includes the wavelengths between 10 and 400 nm. Ultraviolet C (UV-C) consists of wavelengths between 10 and 290 nm and does not reach the earth because of its absorption by stratospheric ozone. These wavelengths are not a cause of photosensitivity except in occupational settings where artificial sources of this energy are employed -- e.g., for germicidal effects. Ultraviolet B (UV-B) consists of wavelengths between 290 and 320 nm. This portion of the photobiologic action spectrum is the most efficient in producing redness or erythema in human skin and hence is sometimes known as the "sunburn spectrum." Ultraviolet A (UV-A) represents those wavelengths between 320 and 400 nm and is approximately 1000-fold less efficient in producing skin hyperemia than is UV-B. The UV-A has also been divided into two parts known as UV-A 1 (340 to 400 nm) and UV-A 2 (320 to 340 nm).

The visible wavelengths between 400 and 700 nm include the familiar white light which when directed through a prism can be shown to consist of various colors including violet, indigo, blue, green, yellow, orange, and red. The energy possessed by photons in the visible spectrum is not capable of damaging human skin in the absence of a photosensitizing chemical. The absorption of energy is critical to the development of photosensitivity. Thus the *absorption spectrum* of a molecule is defined as the range of wavelengths absorbed by it, whereas the *action spectrum* for an effect of incident radiation is defined as the range of wavelengths that evoke the response.

Photosensitivity occurs when a photon-absorbing chemical (chromophore) present in the skin absorbs incident energy, becomes excited, and transfers the absorbed energy to various structures or to oxygen. The absorbed energy must be dissipated by processes including heat, fluorescence, and phosphorescence. It is important to emphasize that absorption spectra and action spectra need not be superimposable, but there must be overlap at some point to produce photosensitization.

STRUCTURE AND FUNCTION OF SKIN

The skin's exposure to sunlight permits the absorption of some wavelengths and the transmission of others. Essentially, human skin is a sandwich of two distinctive compartments, the epidermis and dermis, separated by a basement membrane. The outer epidermis is a stratified squamous epithelium comprising the surface stratum corneum (a protein- and lipid-rich compact membrane), the stratum granulosum, stratum spinosum, and the basal cell layer. The basal cell layer contains a heterogeneous population of cells, a subset of which migrate upward in the process of terminal differentiation that results in the expression of specific keratin genes and the formation of the stratum corneum. Epidermal cells include resident keratinocytes and melanocytes and immigrant cells, including the immunologically active Langerhans cells, lymphocytes, polymorphonuclear leukocytes, monocytes, and macrophages, making the epidermis a major component of the immune system. Branches of sensory nerve endings also reach into this compartment.

The second major component of skin is the dermis, which is relatively large and less densely populated with cells that include fibroblasts, endothelial cells within dermal vessels, and mast cells. Tissue macrophages and sparsely distributed inflammatory cells are also present. All these cells exist within an extracellular matrix of collagen, elastin, and glycosaminoglycans. In contrast to the epidermis, rich vascularization of the dermis allows it to play an important role in temperature regulation and in inflammatory responses to skin injury.

UV RADIATION (UVR) AND SKIN

The epidermis and the dermis contain several chromophores capable of interacting with incident solar energy. These interactions include reflection, refraction, absorption, and transmission. The stratum corneum is a major impediment to the transmission ofUV-B, and less than 10% of incident wavelengths in this region penetrate the basement membrane. Approximately 3% of radiation below 300 nm, 20% of radiation below 360 nm, and 33% of short visible radiation reaches the basal cell layer in untanned human skin. Proteins and nucleic acids absorb intensely in the short UV-B. In contrast, UV-A 1 and 2 penetrate the epidermis efficiently to reach the dermis, where they likely produce changes in structural and matrix proteins that contribute to the aged appearance of chronically sun-exposed skin, particularly in individuals of light complexion.

One of the consequences ofUV-B absorption by DNA is the production of pyrimidine dimers. These structural changes can be repaired by mechanisms that result in their recognition and excision, and the reestablishment of normal base sequences. The efficient repair of these structural aberrations is crucial, since individuals with defective DNA repair are at high risk for the development of cutaneous cancer. For example, patients with xeroderma pigmentosum, an autosomal recessive disorder, are characterized by variably decreased repair of UV-induced photoproducts, and their skin may develop the xerotic appearance of photoaging as well as basal cell and squamous cell carcinomas and melanoma in the first two decades of life. Studies in mice using knockout gene technology have verified the importance of genes regulating these repair pathways in preventing the development of UV-induced cancer.

Cutaneous Optics and Chromophores Chromophores are endogenous or exogenous chemical components that can absorb physical energy. Endogenous chromophores of skin are of two types: (1) chemicals that are normally present, including nucleic acids, proteins, lipids, and 7-dehydrocholesterol, the precursor of vitamin D; and (2) chemicals, such as porphyrins, synthesized elsewhere in the body that circulate in the bloodstream and diffuse into the skin. Normally, only trace amounts of porphyrins are present in the skin, but in the diseases known as the porphyrias, increased amounts are released into the circulation and are transported to the skin, where they absorb incident energy both in the Soret band around 400 nm (short visible) and to a lesser extent in the red portion of the visible spectrum (580 to 660 nm). This results in structural damage to the skin that may be manifest as erythema, edema, urticaria, or blister formation (Chap. 346).

Acute Effects of Sun Exposure The immediate cutaneous consequences of sun exposure include sunburn and vitamin D synthesis.

Sunburn This very common affliction of human skin is caused by exposure toUVR. Generally speaking, the individual's ability to tolerate sunlight is inversely proportional to his or her melanin pigmentation. Melanin is a complex polymer of tyrosine that functions as an efficient neutral-density filter with broad absorbance within theUVportion of the solar spectrum. Melanin is synthesized in specialized epidermal dendritic cells termed *melanocytes* and is packaged into *melanosomes* that are transferred via dendritic processes into *keratinocytes*, where they provide photoprotection. Sun-induced melanogenesis is a consequence of increased tyrosinase activity in melanocytes that in turn may be due to a combination of eicosanoid and endothelin-1 release. Tolerance of sun exposure is a function of the efficiency of the epidermal-melanin unit and can usually be ascertained by asking an individual two questions: (1) Do you burn after sun exposure? and (2) Do you tan after sun exposure? By the answers to these questions, it is usually possible to divide the population into six skin types varying from type I (always burn, never tan) to type VI (never burn, always tan) (Table 60-1).

There are two general theories about the pathogenesis of the sunburn response. First, the lag phase in time between skin exposure and the development of visible redness (usually 4 to 12 h) suggests an epidermal chromophore that causes delayed production and/or release of vasoactive mediator(s), or cytokines, that diffuse to the dermal vasculature to evoke vasodilatation. Indeed,UVRstimulates the release of numerous proinflammatory cytokines and nitric oxide by keratinocytes. Second, it is possible that the small amount of incidentUV-B radiation (10% or less) that penetrates to the dermis can be absorbed directly by endothelial cells in the vasculature, thereby resulting in vasodilatation. The issue remains unresolved.

The action spectrum for sunburn erythema includes theUV-B and UV-A regions. Photons in the shorter UV-B are at least 1000-fold more efficient than photons in the longer UV-B and the UV-A in evoking the response. However, UV-A may contribute to sunburn erythema at midday when much more UV-A than UV-B is present.

The mechanism of injury remains poorly defined, but the action spectrum forUV-B erythema closely resembles the absorption spectrum for DNA after adjusting for the absorbance of incident energy by the stratum corneum. Apoptotic keratinocytes (so-called sunburn cells) are visible histologically within an hour of exposure and are maximal within 24 h. UV-A is less effective than UV-B in producing sunburn cells. Mast cells may release inflammatory mediators after exposure to UV-B and UV-A. For example, erythema doses of both UV-B and UV-A increase histamine levels in experimentally induced suction blisters of human skin that return to normal after 24 h (before visible erythema has subsided). Prostaglandin E_2increases to approximately 150% of control levels after 24 h and then diminishes. Since prostaglandins evoke both pain and redness when injected intradermally, their presence in suction blisters after UV-B exposure suggests a role in UV-B erythema. Age-related declines occur in the amount of inflammatory mediators detectable in human skin after UV-B irradiation. UV-A erythema results in few epidermal sunburn cells, but vascular endothelial injury is greater than with UV-B. In addition, there are increased levels of arachidonic acid and of prostaglandins D_2, E_2, and I_2 that peak within 5 to 9 h and then subside before peak redness occurs. Despite evidence for the role of prostaglandins in both UV-B- and UV-A-irradiated skin, administration of nonsteroidal anti-inflammatory drugs is more effective in reducing erythema evoked by UV-B than by UV-A. UV-B also induces cutaneous matrix-degrading metalloproteinases within hours of exposure.

Vitamin D Photochemistry Cutaneous exposure toUV-B causes photolysis of epidermal previtamin D_3(7-dehydrocholesterol) to previtamin D_3, which then undergoes a temperature-dependent isomerization to form the stable hormone vitamin D_3. This compound then diffuses to the dermal vasculature and circulates systemically where it is converted to the functional hormone 1,25-dihydroxy vitamin D_3[1,25$(OH)_2D_3$]. Vitamin D metabolites from the circulation or those produced in the skin itself can augment epidermal differentiation signaling. Aging substantially decreases the ability of human skin to produce vitamin D_3. This, coupled with the widespread use of sunscreens that filter outUV-B, has led to concern that vitamin D deficiency may become a significant clinical problem in the elderly. Indeed, studies have shown that the use of sunscreens can diminish the production of vitamin D_3 in human skin.

Chronic Effects of Sun Exposure: Nonmalignant The clinical features of photodamaged sun-exposed skin consist of wrinkling, blotchiness, telangiectasia, and a roughened, irregular, "weather-beaten" appearance. Whether these changes, which some refer to as *photoaging* or *dermatoheliosis*, represent accelerated chronologic aging or a separate and distinct process is not clear.

Within chronically sun-exposed epidermis, there is thickening (acanthosis) and morphologic heterogeneity within the basal cell layer. Higher but irregular melanosome content may be present in some keratinocytes, indicating prolonged residence of the cells in the basal cell layer. These structural changes may help to explain the leathery texture and the blotchy discoloration of sun-damaged skin.

The dermis is the major site for sun-associated chronic damage, manifest as a massive increase in thickened irregular masses of tangled elastic fibers resulting from enhanced expression of elastin genes. Collagen fibers are also abnormally clumped in the deeper dermis. Fibroblasts are increased in number and show morphologic signs suggesting activation. Degraded mast cells may be present in the dermis, the relevance of which remains unclear.

These morphologic changes, both gross and microscopic, are features of chronically sun-exposed skin. The chromophore(s), the action spectra, and the specific biochemical events orchestrating these changes are unknown.

Chronic Effects of Sun Exposure: Malignant One of the major known consequences of chronic skin exposure to sunlight is nonmelanoma skin cancer. The two types of nonmelanoma skin cancer are basal cell and squamous cell carcinoma (Chap. 86). There are three major steps for cancer induction: initiation, promotion, and progression. Chronic exposure of animal skin to artificial light sources that mimic solarUVRresults in *initiation*, a step whereby structural (mutagenic) changes in DNA evoke an irreversible alteration in the target cell (keratinocyte) that begins the tumorigenic process. Exposure to a tumor initiator is believed to be a necessary but not sufficient step in the malignant process, since initiated skin cells not exposed to tumor promoters do not generally develop tumors. The second stage in tumor development is *promotion*, a multistep process whereby initiated cells are exposed to chemical and physical agents that evoke epigenetic changes that culminate in the clonal expansion of initiated cells and cause the development, over a period of weeks to months, of benign growths known as *papillomas*. Again, using transgenic animals, the importance ofUVeffects on the expression of additional oncogenes such as *fos* and *jun* in developing papillomas has been demonstrated. UV-B is a *complete carcinogen*, meaning that it can function as both an initiator and a promoter, leading to tumor induction. *Incomplete carcinogens* can initiate tumorigenesis but require additional skin exposure to tumor promoters to elicit tumors. The prototype tumor promoter is the phorbol ester 12-*O*-tetradecanoyl phorbol-13-acetate. Tumor promotion usually requires multiple exposures over time to evoke a neoplasm.

The final step in the malignant process is the conversion of benign precursors into malignant lesions, a process thought to require additional genetic alterations in already transformed cells. Indeed, *ras* gene mutations have been detected in a minority of human nonmelanoma skin cancers. Mutations of the tumor suppressor gene p53 also occur in sun-damaged human skin.

Sun exposure causes nonmelanoma and melanoma cancers of the skin, although the evidence is far more direct for its role in nonmelanoma (basal cell and squamous cell carcinoma) than in melanoma. Approximately 80% of nonmelanoma skin cancers develop on exposed body area, including the face, the neck, and the hands. Men of fair complexion who work outdoors are twice as likely as women to develop these types of cancers. Whites of darker complexions (e.g., Hispanics) have one-tenth the risk of developing such cancers as do light-skinned individuals. Blacks are at lowest risk for all forms of skin cancer. Between 600,000 and 800,000 individuals in the United States develop nonmelanoma skin cancer annually, and the lifetime risk for a white individual to develop such a neoplasm is estimated at approximately 15%. A consensus exists that the incidence of nonmelanoma skin cancer in the population is rising, for reasons that are unclear.

The relationship of sun exposure to melanoma is less clear-cut, but suggestive evidence supports an association. Melanomas occasionally develop by the teenage years, indicating that the latent period for tumor growth is less than that of nonmelanoma skin cancer. Melanomas are among the most rapidly increasing of all human malignancies (Chap. 86). Epidemiologic studies of immigrants of similar ethnic stock indicate that individuals born in one area or who migrated to the same locale before age 10 have higher age-specific melanoma rates than individuals arriving later. It is thus reasonable to conclude that life in a sunny climate from birth or early childhood increases the risk of melanoma. In general, risk does not correlate with cumulative sun exposure but may relate to sequelae of sun exposure in childhood. Thus, a blistering sunburn is associated with a doubling of melanoma risk at the site of the reaction.

Immunologic Effects Exposure to solar radiation influences both local and systemic immune responses.UV-B appears to be most efficient in altering immune responses, likely related to the capacity of such energy to affect antigen presentation in skin by interacting with epidermal Langerhans cells. These bone marrow-derived dendritic cells possess surface markers characteristic of monocytes and macrophages. Following skin exposure to erythema doses of UV-B, Langerhans cells undergo both morphologic and functional changes that result in decreased contact allergic responses when haptens are applied to the irradiated site. This diminished capacity for sensitization is due to the induction of antigen-specific suppressor T lymphocytes. Indeed, while the immunosuppressive effect of irradiation is limited to haptens applied to the irradiated site, the net result is systemic immune suppression to that antigen because of the induction of suppressor T cells.

Higher doses of radiation evoke diminished immunologic responses to antigens introduced either epicutaneously or intracutaneously at sites distant from the irradiated site. These suppressed responses are also associated with the induction of antigen-specific suppressor T lymphocytes and may be mediated by as yet undefined factors that are released from epidermal cells at the irradiated site. The implications of this generalized immune suppression in terms of altered susceptibility to cutaneous cancer or to infection remain to be defined.

It is known thatUV-induced tumors in murine skin are antigenic and are rapidly rejected when transplanted into normal syngeneic animals. If the tumors are transplanted into animals previously exposed to subcarcinogenic doses of UV-B, they are not rejected and instead grow progressively in the recipients. This failure of irradiated animals to reject the transplanted tumors is due to the development of T suppressor cells that prevent the rejection response. While the mechanism of suppression of tumor rejection is unknown, such a response might be a critical determinant of cancer risk in human skin.

PHOTOSENSITIVITY DISEASES

The diagnosis of photosensitivity requires a careful history to define the duration of the signs and symptoms, the length of time between exposure to sunlight and the development of subjective complaints, and visible changes in the skin. The age of onset also can be a helpful clue; for example, the acute photosensitivity of erythropoietic protoporphyria almost always begins in childhood, whereas the chronic photosensitivity of porphyria cutanea tarda typically begins in the fourth and fifth decades. A history of exposure to topical and systemic drugs and chemicals may provide important information. Many classes of drugs can cause photosensitivity on the basis of either phototoxicity or photoallergy. Fragrances such as musk ambrette that were previously present in numerous cosmetic products are also potent photosensitizers.

Examination of the skin may also offer important clues. Anatomic areas that are naturally protected from direct sunlight such as the hairy scalp, the upper eyelids, the retroauricular areas, and the infranasal and submental regions may be spared, whereas exposed areas show characteristic features of the pathologic process. These anatomic localization patterns are often helpful but not infallible in making the diagnosis. For example, airborne contact sensitizers that are blown onto the skin may produce dermatitis that can be difficult to distinguish from photosensitivity, despite the fact that such material may trigger skin reactivity in areas shielded from direct sunlight.

Many dermatologic conditions may be caused or aggravated by light (Table 60-2). The role of light in evoking these responses may be dependent on genetic abnormalities ranging from well-described defects in DNA repair that occur in xeroderma pigmentosum to the inherited abnormalities in heme synthesis that characterize the porphyrias. In certain photosensitivity diseases, the chromophore has been identified, whereas in the majority, the energy-absorbing agent is unknown.

Polymorphous Light Eruption After sunburn, the most common type of photosensitivity disease is *polymorphous light eruption*, the mechanism of which is unknown. Many affected individuals never seek medical attention because the condition is often transient, becoming manifest each spring with initial sun exposure but then subsiding spontaneously with continuing exposure, a phenomenon known as "hardening." The major manifestations of polymorphous light eruption include pruritic (often intensely so) erythematous papules that may coalesce into plaques on exposed areas of the face and arms or other areas as well, making the distribution spotty and uneven.

The diagnosis can be confirmed by skin biopsy and by performing phototest procedures in which skin is exposed to multiple erythema doses ofUV-A and UV-B. The action spectrum for polymorphous light eruption is usually within these portions of the solar spectrum.

Treatment of this disease includes the induction of hardening by the cautious administration ofUVlight, either alone or in combination with photosensitizers such as the psoralens (see below).

Phototoxicity and Photoallergy These photosensitivity disorders are related to the topical or systemic administration of drugs and other chemicals. Both reactions require the absorption of energy by a drug or chemical resulting in the production of an excited-state photosensitizer that can transfer its absorbed energy to a bystander molecule or to molecular oxygen, thereby generating tissue-destructive chemical species.

Phototoxicity is a nonimmunologic reaction caused by drugs and chemicals, a few of which are listed in Table 60-3. The usual clinical manifestations include erythema resembling a sunburn that quickly desquamates or "peels" within several days. In addition, edema, vesicles, and bullae may occur.

Photoallergy is distinct in that the immune system participates in the pathologic process. The excited-state photosensitizer may create highly unstable haptenic free radicals that bind covalently to macromolecules to form a functional antigen capable of evoking a delayed hypersensitivity response. Some of the drugs and chemicals that produce photoallergy are listed in Table 60-4. The clinical manifestations typically differ from those of phototoxicity in that an intensely pruritic eczematous dermatitis tends to predominate and evolves into lichenified, thickened, "leathery" changes in sun-exposed areas. A small subset (perhaps 5 to 10%) of patients with photoallergy may develop a persistent exquisite hypersensitivity to light even when the offending drug or chemical is identified and eliminated. Known as *persistent light reaction*, this may be incapacitating for years. Some have used the term *chronic actinic dermatitis* to encompass these chronic hyperresponsive states.

Diagnostic confirmation of phototoxicity and photoallergy often can be obtained using phototest procedures. In patients with suspected phototoxicity, determination of the minimal erythema dose (MED) while the patient is exposed to a suspected agent and then repeating the MED after discontinuation of the agent may provide a clue to the causative drug or chemical. Photopatch testing can be performed to confirm the diagnosis of photoallergy. This is a simple variant of ordinary patch testing in which a series of known photoallergens is applied to the skin in duplicate and one set is irradiated with a suberythema dose ofUV-A. Development of eczematous changes at sites exposed to sensitizer and light is a positive result. The characteristic abnormality in patients with persistent light reaction is a diminished threshold to erythema evoked by UV-B. Patients with chronic actinic dermatitis may have a broad spectrum of UV hyperresponsiveness.

The management of drug photosensitivity is first and foremost to eliminate exposure to the chemical agents responsible for the reaction and to minimize sun exposure. The acute symptoms of phototoxicity may be ameliorated by cool, moist compresses, topical glucocorticoids, and systemically administered nonsteroidal antiinflammatory agents. In severely affected individuals, a rapidly tapered course of systemic glucocorticoids may be useful. Judicious use of analgesics may be necessary.

Photoallergic reactions require a similar management approach. Furthermore, individuals suffering from persistent light reactivity must be meticulously protected against light exposure. In selected patients in whom chronic systemic high-dose glucocorticoids pose unacceptable risks, it may be necessary to employ cytotoxic agents such as azathioprine or cyclophosphamide.

Porphyria The porphyrias (Chap. 346) are a group of diseases that have in common various derangements in the synthesis of heme. Heme is an iron-chelated tetrapyrrole or porphyrin, and the nonmetal chelated porphyrins are potent photosensitizers that absorb light intensely in both the short (400 to 410 nm) and the long (580 to 650 nm) portions of the visible spectrum.

Heme cannot be reutilized and must be continuously synthesized, and the two body compartments with the largest capacity for its production are the bone marrow and the liver. Accordingly, the porphyrias originate in one or the other of these organs, with the end result of excessive endogenous production of potent photosensitizing porphyrins. The porphyrins circulate in the bloodstream and diffuse into the skin, where they absorb solar energy, become photoexcited, and evoke cutaneous photosensitivity. The mechanism of porphyrin photosensitization is known to be photodynamic or oxygen-dependent and is mediated by reactive oxygen species such as superoxide anions.

Porphyria cutanea tarda is the most common type of human porphyria and is associated with decreased activity of the enzyme uroporphyrinogen decarboxylase associated with a number of gene mutations. There are two basic types of porphyria cutanea tarda: the sporadic or acquired type, generally seen in individuals ingesting ethanol or receiving estrogens; and the inherited type, in which there is autosomal dominant transmission of deficient enzyme activity. Both forms are associated with increased hepatic iron stores.

In both types of porphyria cutanea tarda, the predominant feature is a chronic photosensitivity characterized by increased fragility of sun-exposed skin, particularly areas subject to repeated trauma such as the dorsa of the hands, the forearms, the face, and the ears. The predominant skin lesions are vesicles and bullae that rupture, producing moist erosions, often with a hemorrhagic base, that heal slowly with crusting and purplish discoloration of the affected skin. Hypertrichosis, mottled pigmentary change, and scleroderma-like induration are associated features. Biochemical confirmation of the diagnosis can be obtained by measurement of urinary porphyrin excretion, plasma porphyrin assay, and by assay of erythrocyte and/or hepatic uroporphyrinogen decarboxylase. Multiple mutations of the uroporphyrinogen decarboxylase gene have been identified in human populations, including exon skipping and base substitutions.

Treatment consists of repeated phlebotomies to diminish the excessive hepatic iron stores and/or intermittent low doses of the antimalarial drugs chloroquine and hydroxychloroquine. Long-term remission of the disease can be achieved if the patient eliminates exposure to porphyrinogenic agents.

Erythropoietic protoporphyria originates in the bone marrow and is due to a decrease in the mitochondrial enzyme ferrochelatase secondary to numerous gene mutations. The major clinical features include an acute photosensitivity characterized by subjective burning and stinging of exposed skin that often develops during or just after exposure.

There may be associated skin swelling and, after repeated episodes, a waxlike scarring.

The diagnosis is confirmed by demonstration of measurement of free elevated erythrocyte protoporphyrin. Detection of increased plasma protoporphyrin helps to differentiate lead poisoning and iron-deficiency anemia, in both of which elevated erythrocyte protoporphyrin occurs in the absence of cutaneous photosensitivity and of elevated plasma protoporphyrin.

Treatment consists of reducing sun exposure and the oral administration of the carotenoidb-carotene, which is an effective scavenger of free radicals. This drug increases tolerance to sun exposure in many affected individuals, although it has no effect on deficient ferrochelatase.

An algorithm for the approach to a patient with photosensitivity is illustrated inFig. 60-1.

PHOTOPROTECTION

Since photosensitivity of the skin results from exposure to sunlight, it follows that avoidance of the sun would eliminate these disorders. Unfortunately, social pressures make this an impractical alternative for most individuals, and this has led to a search for better approaches to photoprotection.

Natural photoprotection is provided by structural proteins in the epidermis, particularly keratins and melanin. The amount of melanin and its distribution in cells is genetically regulated, and individuals of darker complexion (skin types IV to VI) are at decreased risk for the development of cutaneous malignancy.

Other forms of photoprotection include clothing and sunscreens. Clothing constructed of tightly woven sun-protective fabrics, irrespective of color, affords substantial protection. Wide-brimmed hats, long sleeves, and trousers all reduce direct exposure. Sunscreens are of two major types -- chemical and physical. Chemical sunscreens are chromophores that absorb energy in theUV-B and/or UV-A regions, thereby diminishing photon absorption by the skin (Table 60-5). Sunscreens are rated for their photoprotective effect by their *sun protective factor* (SPF). The SPF is simply a ratio of the time required to produce sunburn erythema with and without sunscreen application. SPF ratings of 15 or higher provide effective protection against UV-B and, to a lesser extent, UV-A. The major categories of chemical sunscreens include *p*-aminobenzoic acid and its esters, benzophenones, anthranilates, cinnamates, and salicylates. Physical sunscreens are light-opaque mixtures containing metal particles such as titanium oxide and zinc oxide that scatter light, thereby reducing photon absorption by the skin.

In addition to light absorption, a critical determinant of the photoprotective effect of sunscreens is their ability to remain on the skin, a property known as *substantivity*. In general, the *p*-aminobenzoic acid esters formulated in moisturizing vehicles provide the greatest substantivity.

Photoprotection can also be achieved by limiting the time of exposure during the day. Since the majority of an individual's total lifetime sun exposure may occur by the age of

18, it is important to educate parents and young children about the hazards of sunlight. Simply eliminating exposure at midday will substantially reduce lifetimeUV-B exposure.

PHOTOTHERAPY AND PHOTOCHEMOTHERAPY

UVRcan also be used therapeutically. The administration ofUV-B alone or in combination with topically applied agents can induce remissions of psoriasis and atopic dermatitis.

Photochemotherapy in which topically applied or systemically administered *psoralens* are combined withUV-A (PUVA) is also effective in treating psoriasis and in the early stages of cutaneous T cell lymphoma and vitiligo. Psoralens are tricyclic furocoumarins that, when intercalated into DNA and exposed to UV-A, form adducts with pyrimidine bases and eventually form DNA cross-links. These structural changes are thought to decrease DNA synthesis and relate to improvement that occurs in psoriasis. The reason that PUVA photochemotherapy is effective in cutaneous T cell lymphoma is not clear.

In addition to its effects on DNA, PUVA photochemotherapy also stimulates melanin synthesis, and this provides the rationale for its use in the depigmenting disease vitiligo. Oral 8-methoxypsoralen andUV-A appear to be most effective in this regard, but as many as 100 treatments extending over 12 to 18 months may be required to promote satisfactory repigmentation.

The major side effects ofUV-B phototherapy and PUVA photochemotherapy are due to the cumulative effects of photon absorption and include skin dryness, actinic keratoses, and an increased risk of nonmelanoma skin cancer. Despite these risks, the therapeutic index of these modalities is quite acceptable.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 10 -HEMATOLOGIC ALTERATIONS

61. ANEMIA AND POLYCYTHEMIA - *John W. Adamson*, *Dan L. Longo*

HEMATOPOIESIS AND THE PHYSIOLOGIC BASIS OF RED CELL PRODUCTION

Hematopoiesis is the process by which the formed elements of the blood are produced. The process is regulated through a series of steps beginning with the pluripotent hematopoietic stem cell. Stem cells are capable of producing red cells, all classes of granulocytes, monocytes, platelets, and the cells of the immune system. Commitment of the stem cell to the specific cell lineages appears not to be regulated by known exogenous growth factors or cytokines. Rather, stem cells develop into differentiated cell types through incompletely defined molecular events that are intrinsic to the stem cell itself (Chap. 104). Following lineage commitment (or differentiation), hematopoietic progenitor and precursor cells come increasingly under the regulatory influence of growth factors and hormones, such as erythropoietin (EPO) for red cell production. EPO is required for the maintenance of committed erythroid progenitor cells which, in the absence of the hormone, undergo programmed cell death (*apoptosis*). The regulated process of red cell production is *erythropoiesis*, and its key elements are illustrated in Fig. 61-1.

In the bone marrow, the first morphologically recognizable erythroid precursor is the pronormoblast. This cell can undergo 4 to 5 cell divisions that result in the production of 16 to 32 mature red cells. With increasedEPOproduction, or the administration of EPO as a drug, early progenitor cell numbers are amplified and, in turn, give rise to increased numbers of erythrocytes. The regulation of EPO production itself is linked to O_2transport.

In mammals, O_2 is transported to tissues bound to the hemoglobin contained within circulating red cells. The mature red cell is 8um in diameter, anucleate, discoid in shape, and extremely pliable in order for it to traverse the microcirculation successfully; its membrane integrity is maintained by the intracellular generation of ATP. Normal red cell production results in the daily replacement of 0.8 to 1% of all circulating red cells in the body. The average red cell lives 100 to 120 days. The machinery responsible for red cell production is called the *erythron*. The erythron is a dynamic organ made up of a rapidly proliferating pool of marrow erythroid precursor cells and a large mass of mature circulating red blood cells. The size of the red cell mass reflects the balance of red cell production and destruction. The physiologic basis of red cell production and destruction provides an understanding of the mechanisms that can lead to anemia.

The physiologic regulator of red cell production, the glycoprotein hormoneEPO, is produced and released by peritubular capillary lining cells within the kidney. These cells are highly specialized epithelial-like cells. A small amount of EPO is produced by hepatocytes. The fundamental stimulus for EPO production is the availability of O_2 for tissue metabolic needs. Impaired O_2delivery to the kidney can result from a decreased red cell mass (*anemia*), impaired O_2loading of the hemoglobin molecule (*hypoxemia*), or, rarely, impaired blood flow to the kidney (renal artery stenosis). EPO governs the day-to-day production of red cells, and ambient levels of the hormone can be measured in the plasma by sensitive immunoassays -- the normal level being 10 to 25 U/L. When the hemoglobin concentration falls below 100 to 120 g/L (10 to 12 g/dL), plasma EPO levels increase logarithmically in inverse proportion to the severity of the anemia. In circulation, EPO has a half-clearance time of 6 to 9 h. EPO acts by binding to specific receptors on the surface of marrow erythroid precursors, inducing them to proliferate and to mature. Under the stimulus of EPO, red blood cell production can increase four- to fivefold within a 1- to 2-week period but only in the presence of adequate nutrients, especially iron. The functional capacity of the erythron, therefore, requires normal renal production of EPO, a functioning erythroid marrow, and an adequate supply of substrates for hemoglobin synthesis. A defect in any of these key components can lead to anemia. Generally, anemia is recognized in the laboratory when a patient's hemoglobin level or hematocrit is reduced below an expected value (the normal range). The likelihood and severity of anemia are defined based on the deviation of the patient's hemoglobin/hematocrit from values expected for age- and sex-matched normal subjects. The lower ranges of distribution of hemoglobin/hematocrit values for adult males and females are shown in Fig. 61-2. The hemoglobin concentration in adults has a Gaussian distribution. The mean hematocrit value for adult males is 47% (± SD 7) and that for adult females is 42% (± 5). Any individual hematocrit or hemoglobin value carries with it a likelihood of associated anemia. Thus, a hematocrit of£39% in an adult male or<35% in an adult female has only about a 25% chance of being normal. Suspected low hemoglobin or hematocrit values are more easily interpreted if there are historic values for the same patient for comparison.

The critical elements of erythropoiesis -- EPOproduction, iron availability, the proliferative capacity of the bone marrow, and effective maturation of red cell precursors -- are used for the initial classification of anemia (see "Definition and Classification, below).

ANEMIA

CLINICAL PRESENTATION OF ANEMIA

Signs and Symptoms Anemia is most often recognized by abnormal screening laboratory tests. Patients only occasionally present with advanced anemia and its attendant signs and symptoms. Acute anemia is nearly always due to blood loss or hemolysis. In fact, with acute blood loss, hypovolemia dominates the clinical picture and the hematocrit and hemoglobin levels do not reflect the volume of blood lost. Signs of vascular instability dominate with acute losses of 10 to 15% of the total blood volume. In such patients, the issue is not anemia but hypotension and decreased organ perfusion. When>30% of the blood volume is lost suddenly, patients are unable to compensate with the usual mechanisms of vascular contraction and changes in regional blood flow. The patient prefers to remain supine and will show postural hypotension and tachycardia if upright. If the volume of blood lost is >40% (i.e., >2 L in the average-sized adult), signs of hypovolemic shock including confusion, air hunger, diaphoresis, hypotension, and tachycardia appear (Chap. 108). Such patients have significant deficits in vital organ perfusion and require immediate volume replacement. With mild blood loss, enhanced O_2delivery is achieved through changes in the O_2-hemoglobin dissociation curve mediated by a decreased pH or increased CO_2(*Bohr effect*).

With acute hemolytic disease, the signs and symptoms depend on the mechanism that leads to red cell destruction. Intravascular hemolysis with release of free hemoglobin may be associated with acute back pain, free hemoglobin in the plasma and urine, and renal failure. Symptoms associated with more chronic or progressive anemia depend on the age of the patient and the adequacy of blood supply to critical organs. Symptoms associated with moderate anemia include fatigue, loss of stamina, breathlessness, and tachycardia (particularly with physical exertion). However, because of the intrinsic compensatory mechanisms that govern the O_2-hemoglobin dissociation curve, the gradual onset of anemia -- particularly in young patients -- may not be associated with signs or symptoms until the anemia is severe [hemoglobin <70 to 80 g/L (7 to 8 g/dL)]. When anemia develops over a period of days or weeks, the total blood volume is normal to slightly increased and changes in cardiac output and regional blood flow help compensate for the overall loss in O_2-carrying capacity. Changes in the position of the O_2-hemoglobin dissociation curve account for some of the compensatory response to anemia. With chronic anemia, intracellular levels of 2,3-bisphosphoglycerate (BPG) rise, shifting the dissociation curve to the right and facilitating O_2unloading. This compensatory mechanism can only maintain normal tissue O_2delivery in the face of a 20 to 30 g/L (2 to 3 g/dL) deficit in hemoglobin concentration. Finally, further protection of O_2delivery to vital organs is achieved by the shunting of blood away from organs that are relatively rich in blood supply, particularly the kidney, gut, and skin.

Certain disorders are commonly associated with anemia. Chronic inflammatory states (e.g., infection, rheumatoid arthritis) are associated with mild to moderate anemia, whereas lymphoproliferative disorders, such as chronic lymphocytic leukemia and certain other B cell neoplasms, may be associated with autoimmune hemolysis.

Approach to the Patient

The evaluation of the patient with anemia requires a careful history and physical examination. Historic information that may be useful includes exposure to certain toxic agents or drugs and symptoms related to other disorders commonly associated with anemia. These include symptoms and signs such as bleeding, fatigue, malaise, fever, weight loss, night sweats, and other systemic symptoms. Clues to the mechanisms of anemia may be provided on physical examination by findings of infection, blood in the stool, lymphadenopathy, splenomegaly, or petechiae. Splenomegaly and lymphadenopathy suggest an underlying lymphoproliferative disease, while petechiae suggest platelet dysfunction. If it is uncertain whether a mild anemia represents an extreme normal value or an abnormal finding, past laboratory measurements may be helpful. Nutritional history related to drugs or alcohol intake and family history of anemia should always be assessed. Certain geographic backgrounds and ethnic origins are associated with an increased likelihood of an inherited disorder of the hemoglobin molecule or intermediary metabolism. Glucose-6-phosphate dehydrogenase deficiency and certain hemoglobinopathies are seen more commonly in those of middle-Eastern or African origin.

In the anemic patient, physical examination may demonstrate a forceful heartbeat, strong peripheral pulses, and a systolic "flow" murmur. The skin and mucous membranes may be pale if the hemoglobin is <80 to 100 g/L (8 to 10 g/dL). This part of the physical examination should focus on areas where vessels are close to the surface such as the mucous membranes, nail beds, and palmar creases. If the palmar creases are lighter in color than the surrounding skin when the hand is hyperextended, the hemoglobin level is usually<80 g/L (8 g/dL).

Laboratory Evaluation Table 61-1 lists the tests used in the initial workup of anemia. A routine complete blood count (CBC) is required as part of the evaluation and includes the hemoglobin, hematocrit, and red cell indices: the mean cell volume (MCV) in femtoliters, mean cell hemoglobin (MCH) in picograms per cell, and mean concentration of hemoglobin per volume of red cells (MCHC) in grams per liter (non-SI: grams per deciliter). The red cell indices are calculated as shown in Table 61-2, and the normal variations in the CBC with age are shown in Table A-7. A number of physiologic factors affect the normal CBC values including age, gender, pregnancy, smoking, and altitude. High-normal hemoglobin values may be seen in men and women who live at altitude or smoke heavily. The elevations in smokers reflect normal compensation due to the displacement of O_2 by CO in hemoglobin binding. Other important information is provided by the reticulocyte count and measurements of iron supply including the *serum iron*, the *total iron-binding capacity* (TIBC; an indirect measure of the transferrin level), and *serum ferritin*. Marked alterations in the red cell indices usually reflect disorders of maturation or iron deficiency. Clinical laboratories also provide a description of both the red and white cells, a white cell differential count, and the platelet count. In patients with severe anemia and abnormalities in red blood cell morphology, a bone marrow aspirate or biopsy may be important to assist in the diagnosis. Other tests of value in the diagnosis of specific anemias are discussed in chapters on specific disease states.

The components of theCBC also help in the classification of anemia. *Microcytosis* is reflected by a lower than normalMCV(<80), whereas high values (>100) reflect *macrocytosis*. TheMCH andMCHCreflect defects in hemoglobin synthesis (*hypochromia*). Automated cell counters describe the red cell volume distribution width (RDW). The MCV (representing the peak of the distribution curve) is insensitive to the appearance of small populations of macrocytes or microcytes. An experienced laboratory technician will be able to identify minor populations of large or small cells or hypochromic cells before the red cell indices change.

PERIPHERAL BLOOD SMEAR The peripheral blood smear provides important information about defects in red cell production. As a complement to the red cell indices, the blood smear also reveals variations in cell size (*anisocytosis*) and shape (*poikilocytosis*). The degree of anisocytosis usually correlates with increases in theRDW or the range of cell sizes. Poikilocytosis suggests a defect in the maturation of red cell precursors in the bone marrow or fragmentation of circulating red cells. The blood smear may also reveal *polychromasia* -- red cells that are slightly larger than normal and grayish blue in color on the Wright-Giemsa stain. These cells are reticulocytes that have been prematurely released from the bone marrow, and their color represents residual amounts of ribosomal RNA. These cells appear in circulation in response toEPOstimulation or to architectural damage of the bone marrow (fibrosis, infiltration of the marrow by malignant cells, etc.) that results in their disordered release from the marrow. The appearance of nucleated red cells, Howell-Jolly bodies, target cells, sickle cells, and others may provide clues to specific disorders (see Plates V-2,V-3,V-8,V-9,V-16,V-21,V-24,V-26,V-27,V-28, andV-39).

RETICULOCYTE COUNT An accurate reticulocyte count is key to the initial classification of anemia. Normally, reticulocytes are red cells that have been recently released from the bone marrow. They are identified by staining with a supravital dye that precipitates the residual ribosomal RNA. These precipitates appear as blue or black punctate spots. This residual RNA is metabolized over the first 24 to 36 h of the reticulocyte's lifespan in circulation. Normally, the reticulocyte count ranges from 1 to 2% and reflects the daily replacement of 0.8 to 1.0% of the circulating red cell population. A correctly interpreted reticulocyte count provides a reliable measure of red cell production.

In the initial classification of anemia, the patient's reticulocyte count is compared with the expected reticulocyte response. In general, if theEPO and erythroid marrow responses to moderate anemia [hemoglobin<100 g/L (10 g/dL)] are intact, the red cell production rate increases to two to three times normal within 10 days following the onset of anemia. In the face of established anemia, a reticulocyte response less than two to three times normal indicates an inadequate marrow response.

In order to use the reticulocyte count to estimate marrow response, two corrections are necessary. The first correction adjusts the reticulocyte count based on the reduced number of circulating red cells. With anemia, the percentage of reticulocytes may be increased while the absolute number is unchanged. To correct for this effect, the reticulocyte percentage is multiplied by the ratio of the patient's hemoglobin or hematocrit to the expected hemoglobin/hematocrit for the age and gender of the patient (Table 61-3). This provides an estimate of the absolute reticulocyte count. In order to convert the corrected reticulocyte count to an index of marrow production, a further correction is required, depending on whether some of the reticulocytes in circulation have been released from the marrow prematurely. For this second correction, the peripheral blood smear is examined to see if there are polychromatophilic macrocytes present. These cells, representing prematurely released reticulocytes, are referred to as "shift" cells, and the relationship between the degree of shift (and the necessary shift correction factor) is shown in Fig. 61-3. The correction is necessary because these prematurely released cells survive as reticulocytes in circulation for>1 day, thereby providing a falsely high estimate of daily red cell production. If polychromasia is increased, the reticulocyte count, already corrected for anemia, should be divided again by a factor of 2 to account for the prolonged reticulocyte maturation time. The second correction factor varies from 1 to 3 depending upon the severity of anemia. In general, a correction of 2 is commonly used. An appropriate correction is shown inTable 61-3. If polychromatophilic cells are not seen on the blood smear, the second correction is not required. The now doubly corrected reticulocyte count is the *reticulocyte production index*, and it provides an estimate of marrow production relative to normal.

Premature release of reticulocytes is normally due to increasedEPOstimulation. However, if the integrity of the bone marrow release process is lost through tumor infiltration, fibrosis, or other disorders, the appearance of nucleated red cells or polychromatophilic macrocytes should still invoke the second reticulocyte correction. The shift correction should always be applied to a patient with anemia and a very high reticulocyte count to provide a true index of effective red cell production. Patients with severe chronic hemolytic anemia may increase red cell production as much as six- to sevenfold. This measure alone, therefore, confirms the fact that the patient has an appropriate EPO response, a normally functioning bone marrow, and sufficient iron available to meet the demands for new red cell formation. If the reticulocyte production index is <2 in the face of established anemia, a defect in erythroid marrow proliferation or maturation must be present.

TESTS OF IRON SUPPLY AND STORAGE The laboratory measurements that reflect the availability of iron for hemoglobin synthesis include the serum iron, theTIBC, and the percent transferrin saturation. The percent transferrin saturation is derived by dividing the serum iron level (´ 100) by the TIBC. The normal serum iron ranges from 9 to 27 umol/L (50 to 150 ug/dL), while the normal TIBC is 54 to 64 umol/L (300 to 360 ug/dL); the transferrin saturation ranges from 25 to 50%. A diurnal variation in the serum iron leads to a variation in the percent transferrin saturation. The serum ferritin is used to evaluate total-body iron stores. Adult males have serum ferritin levels that average about 100 ug/L, corresponding to iron stores of about 1 g. Adult females have lower serum ferritin levels averaging 30 ug/L, reflecting lower iron stores. A serum ferritin level of 10 to 15 ug/L represents depletion of body iron stores. However, ferritin is also an acute-phase reactant and, in the presence of acute or chronic inflammation, may rise severalfold above baseline levels. As a rule, a serum ferritin >200 ug/L means there is at least some iron in tissue stores.

BONE MARROW EXAMINATION A bone marrow aspirate and smear or a needle biopsy may be useful in the diagnosis of a marrow disorder such as myelofibrosis, a red cell maturation defect, or an infiltrative disease (Plates V-5,V-13,V-14,V-15,V-19,V-29, andV-33). The increase or decrease of one cell lineage (myeloid vs. erythroid) compared to another is obtained by a differential count of nucleated cells in a bone marrow smear [the erythroid/granulocytic (E/G) ratio]. A patient with a hypoproliferative anemia (see below) and a reticulocyte production index<2 will demonstrate an E/G ratio of 1:2 or 1:3. In contrast, patients with hemolytic disease and a production index>3 will have an E/G ratio of at least 1:1. Maturation disorders are identified from the discrepancy between a high E/G ratio and a low reticulocyte production index (see below). Either the marrow smear or biopsy can be stained for the presence of iron stores or iron in developing red cells. The storage iron is in the form of *ferritin* or *hemosiderin*. On carefully prepared bone marrow smears, small ferritin granules can normally be seen in 10 to 20% of developing erythroblasts. Such cells are called *sideroblasts*.

OTHER LABORATORY MEASUREMENTS

Additional laboratory tests may be of value in confirming specific diagnoses.**For details of these tests and how they are applied in individual disorders, see Chaps. 105 to 109.*

DEFINITION AND CLASSIFICATION OF ANEMIA

Initial Classification of Anemia Classifying an anemia according to the functional defect in red cell production helps organize the subsequent use of laboratory studies. The three major classes of anemia are: (1) marrow production defects (*hypoproliferation*), (2) red cell maturation defects (*ineffective erythropoiesis*), and (3) decreased red cell survival (*blood loss/hemolysis*). This functional classification of anemia then guides the selection of specific clinical and laboratory studies designed to complete the differential diagnosis and to plan appropriate therapy. The classification is shown in Fig. 61-4. A hypoproliferative anemia is typically seen with a low reticulocyte production index together with little or no change in red cell morphology (a normocytic, normochromic anemia) (Chap. 105). Maturation disorders typically have a slight to moderately elevated reticulocyte production index that is accompanied by either macrocytic (Chap. 107) or microcytic (Chaps. 105,106) red cell indices. Increased red blood cell destruction secondary to hemolysis results in an increase in the reticulocyte production index to at least three times normal (Chap. 108), provided sufficient iron is available for hemoglobin synthesis. Hemorrhagic anemia does not typically result in production indices of more than 2.5 times normal because of the limitations placed on expansion of the erythroid marrow by iron availability.

In the first branch point of the classification of anemia, a reticulocyte production index>2.5 indicates that hemolysis is most likely. A reticulocyte production index of <2 indicates either a hypoproliferative anemia or maturation disorder. The latter two possibilities can often be distinguished by the red cell indices, by examination of the peripheral blood smear, or by a marrow examination. If the red cell indices are normal, the anemia is almost certainly hypoproliferative in nature. Maturation disorders are characterized by ineffective red cell production and a low reticulocyte production index with bizarre red cell shapes -- macrocytes or hypochromic microcytes on the peripheral blood smear. With a hypoproliferative anemia, no erythroid hyperplasia is noted in the marrow, whereas patients with ineffective red cell production have erythroid hyperplasia and an E/G ratio^{3}1:1.

Hypoproliferative Anemias At least 75% of all cases of anemia are hypoproliferative in nature. A hypoproliferative anemia reflects absolute or relative marrow failure in which the erythroid marrow has not proliferated appropriately for the degree of anemia. The majority of hypoproliferative anemias are due to mild to moderate iron deficiency or inflammation. A hypoproliferative anemia can result from marrow damage, iron deficiency, or inadequateEPOstimulation. The last may reflect impaired renal function, suppression of EPO production by inflammatory cytokines such as interleukin 1, or reduced tissue needs for O_2 from metabolic disease such as hypothyroidism. Only occasionally is the marrow unable to produce red cells at a normal rate, and this is most prevalent in patients with renal failure. In general, hypoproliferative anemias are characterized by normocytic, normochromic red cells, although microcytic, hypochromic cells may be observed with mild iron deficiency or long-standing chronic inflammatory disease. The key laboratory tests in distinguishing between the various forms of hypoproliferative anemia include the serum iron and iron-binding capacity, evaluation of renal and thyroid function, a marrow biopsy or aspirate to detect marrow damage or infiltrative disease, and serum ferritin to assess iron stores. Occasionally, an iron stain of the marrow will be needed to determine the pattern of iron distribution. Patients with the anemia of acute or chronic inflammation show a distinctive pattern of serum iron (low),TIBC(normal or low), percent transferrin saturation (low), and serum ferritin (normal or high). A distinct pattern of results is noted in mild to moderate iron deficiency (low serum iron, high TIBC, low percent transferrin saturation, low serum ferritin) (Chap. 105). Marrow damage by a drug, infiltrative disease such as leukemia or lymphoma, or marrow aplasia can usually be diagnosed from the peripheral blood and bone marrow morphology. With infiltrative disease or fibrosis, a marrow biopsy will likely be required.

Maturation Disorders The presence of anemia with an inappropriately low reticulocyte production index, macro- or microcytosis on smear, and abnormal red cell indices suggests a maturation disorder. Maturation disorders are divided into two categories: nuclear maturation defects, associated with macrocytosis and abnormal marrow development, and cytoplasmic maturation defects, associated with microcytosis and hypochromia usually from defects in hemoglobin production. The low reticulocyte production index is a reflection of the ineffective erythropoiesis that results from the destruction within the marrow of developing erythroblasts. Marrow morphology shows an E/G ratio of ^{3} 1:1, diagnostic of erythroid hyperplasia.

Nuclear maturation defects result from vitamin B_{12} or folic acid deficiency, drug damage, or myelodysplasia. Drugs that interfere with cellular DNA metabolism, such as methotrexate or alkylating agents, can produce a nuclear maturation defect. Alcohol, alone, is also capable of producing macrocytosis and a variable degree of anemia, but this is usually associated with coincident folic acid deficiency. Measurements of folic acid and vitamin B_{12} are key not only in identifying the specific vitamin deficiency but also because they reflect different pathogenetic mechanisms.

Cytoplasmic maturation defects result from severe iron deficiency or abnormalities in globin or heme synthesis. Iron deficiency occupies an unusual position in the classification of anemia. If the iron-deficiency anemia is mild to moderate, erythroid marrow proliferation is decreased and the anemia is classified as hypoproliferative. However, if the anemia is severe and prolonged, the erythroid marrow will become hyperplastic despite the inadequate iron supply, and the anemia will be classified as ineffective erythropoiesis with a cytoplasmic maturation defect. In either case, a reduced reticulocyte production index, microcytosis, and a classic pattern of iron values make the diagnosis clear and easily distinguish iron deficiency from other cytoplasmic maturation defects such as the thalassemias. Defects in heme synthesis, in contrast to globin synthesis, are less common and may be acquired or inherited (Chap. 346). Acquired abnormalities are usually associated with myelodysplasia, may lead to either a macro- or microcytic anemia, and are frequently associated with mitochondrial iron loading. In these cases, iron is taken up by the mitochondria of the developing erythroid cell but not incorporated into heme. The iron-encrusted mitochondria surround the nucleus of the erythroid cell, forming a ring. Based on the distinctive finding of so-called ringed sideroblasts on the marrow iron stain (seePlate V-37), patients are diagnosed as having a sideroblastic anemia -- almost always reflecting myelodysplasia. Again, studies of iron parameters are helpful in the differential diagnosis and management of these patients.

Blood Loss/Hemolytic Anemia In contrast to anemias associated with an inappropriately low reticulocyte production index, blood loss or hemolysis is associated with red cell production indices of ^{3}2.5 times normal. The stimulated erythropoiesis is reflected in the blood smear by the appearance of increased numbers of polychromatophilic macrocytes. A marrow examination is rarely indicated if the reticulocyte production index is increased appropriately. The red cell indices are typically normocytic or slightly macrocytic, reflecting the increased number of reticulocytes. Acute blood loss is not associated with an increased reticulocyte production index because of the time required to increaseEPOproduction and, subsequently, marrow proliferation. Subacute blood loss may be associated with modest reticulocytosis because iron is lost along with the red cells. Anemia from chronic blood loss more often presents as iron deficiency than with the picture of increased red cell production.

The evaluation of blood loss anemia is usually not difficult. Most problems arise when a patient presents with an increased red cell production index from an episode of acute blood loss that went unrecognized. The cause of the anemia and increased red cell production may not be obvious. The confirmation of a recovering state may require observations over a period of 2 to 3 weeks, during which the hemoglobin concentration will be seen to rise and the reticulocyte production index fall.

Hemolytic disease, while dramatic, is among the least common forms of anemia. The ability to sustain a high reticulocyte production index reflects the ability of the erythroid marrow to compensate for hemolysis and the efficient recycling of iron from the destroyed red cells to support new hemoglobin synthesis. The level of response will depend on the severity of the anemia and the nature of the underlying disease process.

Hemolytic anemias present in different ways. Some appear suddenly as an acute, self-limited episode of intravascular or extravascular hemolysis, a presentation pattern often seen in patients with autoimmune hemolysis or with inherited defects of the Embden-Myerhof pathway or the glutathione reductase pathway. Patients with inherited disorders of the hemoglobin molecule or red cell membrane generally have a lifelong clinical history typical of the disease process. Those with chronic hemolytic disease, such as hereditary spherocytosis, may actually present not with anemia but with a complication stemming from the prolonged increase in red cell destruction such as aplastic crisis, symptomatic bilirubin gallstones, or splenomegaly.

The differential diagnosis of an acute or chronic hemolytic event requires the careful integration of family history, pattern of clinical presentation, and a number of highly specific laboratory studies (Chap. 108). Some of the more common congenital hemolytic anemias may be identified from the red cell morphology, a routine laboratory test such as hemoglobin electrophoresis, or a screen for red cell enzymes. Acquired defects in red cell survival are often immunologically mediated and require the immunoglobulin test or a cold agglutinin titer to detect the presence of hemolytic antibodies or complement-mediated red cell destruction.

TREATMENT

An overriding principle is to not initiate treatment of mild to moderate anemia without a specific diagnosis. Rarely, in the acute setting, anemia may be so severe that red cell transfusions are required before a specific diagnosis is made. Whether the anemia is of acute or gradual onset, the selection of the appropriate treatment is determined by the documented cause(s) of the anemia. Often, the cause of the anemia may be multifactorial. For example, a patient with severe rheumatoid arthritis who has been taking anti-inflammatory drugs may have a hypoproliferative anemia associated with chronic inflammation as well as chronic blood loss associated with intermittent gastrointestinal bleeding. In every circumstance, it is important to evaluate the patient's iron status fully before and during the treatment of any anemia.**Transfusion is discussed in Chap. 114; iron therapy is discussed in Chap. 105; treatment of megaloblastic anemia is discussed in Chap. 107; treatment of other entities is discussed* *in their respective chapters (sickle cell anemia, Chap. 106; hemolytic anemias, Chap. 108; aplastic anemia and myelodysplasia, Chap. 109).*

Therapeutic options for the treatment of anemias have expanded dramatically during the past 25 years. Blood component therapy is available and safe. RecombinantEPO as an adjunct to anemia management has transformed the lives of patients with chronic renal failure on dialysis. Improvements in the management of sickle cell crises and sickle cell anemia have also occurred. Eventually, patients with inherited disorders of globin synthesis or mutations in the globin gene, such as sickle cell disease, may benefit from the successful introduction of targeted genetic therapy (Chap. 69).

POLYCYTHEMIA

Polycythemia is defined as an increase in circulating red blood cells above normal. This increase may be real or only apparent (spurious or relative) because of a decrease in plasma volume. The term *erythrocytosis* may be used interchangeably with polycythemia, but some draw a distinction between them; erythrocytosis implies documentation of increased red cell mass, whereas polycythemia refers to any increase in red cells. Often patients with polycythemia are detected through an incidental finding of elevated hemoglobin or hematocrit levels. Concern that the hemoglobin level may be abnormally high is usually triggered at 170 g/L (17 g/dL) for men and 150 g/L (15 g/dL) for women. Hematocrit levels >50% in men or>45% in women may be abnormal. Hematocrits>60% in men and >55% in women are almost invariably associated with increased red cell mass.

Historic features useful in the differential diagnosis include smoking history; living at high altitude; or a history of congenital heart disease, peptic ulcer disease, sleep-apnea, chronic lung disease, or renal disease.

Patients with polycythemia may be asymptomatic or experience symptoms related to the increased red cell mass or an underlying disease process that leads to increased red cell production. The dominant symptoms from increased red cell mass are thrombotic (both venous and arterial), because the blood viscosity increases logarithmically at hematocrits >55%. Manifestations range from digital ischemia to Budd-Chiari syndrome with hepatic vein thrombosis. Abdominal thromboses are particularly common. Neurologic symptoms such as vertigo, tinnitus, headache, and visual disturbances may occur. Hypertension is often present. Patients with *polycythemia vera* may have aquagenic pruritus and symptoms related to hepatosplenomegaly. Patients may have easy bruising, epistaxis, or bleeding from the gastrointestinal tract. Patients with hypoxemia may develop cyanosis on minimal exertion or have headache, impaired mental acuity, and fatigue.

The physical examination usually reveals a ruddy complexion. Splenomegaly favors polycythemia vera as the diagnosis (Chap. 110). The presence of cyanosis or evidence of a right-to-left shunt suggests congenital heart disease presenting in the adult, particularly tetralogy of Fallot or Eisenmenger syndrome (Chap. 234). Increased blood viscosity raises pulmonary artery pressure; hypoxemia can lead to increased pulmonary vascular resistance. Together these factors can produce cor pulmonale.

Polycythemia can be spurious (related to a decrease in plasma volume; Gaisbock's syndrome), primary, or secondary in origin. The secondary causes are all associated with increases inEPOlevels: either a physiologically adapted appropriate elevation based upon tissue hypoxia (lung disease, high altitude, CO poisoning, high-affinity hemoglobinopathy) or an abnormal overproduction (renal disease, tumors with ectopic EPO production). A rare familial form of polycythemia is associated with normal EPO levels but mutations producing hyperresponsive EPO receptors.

Approach to the Patient

As shown inFig. 61-5, the first step is to document the presence of an increased red cell mass using the principle of isotope dilution by administering^{51}Cr-labeled autologous red blood cells to the patient and sampling blood radioactivity over a 2-h period. If the red cell mass is normal (<36 mL/kg in men, <32 mL/kg in women), the patient has spurious polycythemia. If the red cell mass is increased (>36 mL/kg in men, >32 mL/kg in women), serumEPOlevels should be measured. If EPO levels are low or absent, the patient most likely has polycythemia vera. Ancillary tests that support this diagnosis include elevated white blood cell count, increased absolute basophil count, thrombocytosis, elevated leukocyte alkaline phosphatase levels, and elevated serum vitamin B_{12} and vitamin B_{12}-binding protein levels.

If serumEPOlevels are elevated, one attempts to distinguish whether the elevation is a physiologic response to hypoxia or is related to autonomous production. Patients with low arterial O_2saturation (<92%) should be further evaluated for the presence of heart or lung disease, if they are not living at high altitude. Patients with normal O_2saturation who are smokers may have elevated EPO levels because of CO displacement of O_2. If carboxyhemoglobin (COHb) levels are high, the diagnosis is smoker's polycythemia. Such patients should be urged to stop smoking. Those who cannot stop smoking require phlebotomy to control their polycythemia. Patients with normal O_2saturation who do not smoke either have an abnormal hemoglobin that does not deliver O_2 to the tissues (evaluated by finding elevated O_2-hemoglobin affinity) or have a source of EPO production that is not responding to the normal feedback inhibition. Further workup is dictated by the differential diagnosis of EPO-producing neoplasms. Hepatoma, uterine leiomyoma, and renal disease or cysts are all detectable with abdominopelvic computed tomography scans. Cerebellar hemangiomas may produce EPO, but they nearly always present with localizing neurologic signs and symptoms rather than polycythemia-related symptoms.

ACKNOWLEDGEMENT

Dr. Robert S. Hillman wrote this chapter in the 14th edition, and elements of his chapter were retained here.

(Bibliography omitted in Palm version)

Back to Table of Contents

62. BLEEDING AND THROMBOSIS - *Robert I. Handin*

Hemorrhage, intravascular thrombosis, and embolism are common clinical manifestations of many diseases. The normal hemostatic system limits blood loss by precisely regulated interactions between components of the vessel wall, blood platelets, and plasma proteins. However, when disease or trauma damages large arteries and veins, excessive bleeding may occur, despite a normal hemostatic system. Less frequently, hemorrhage is caused by an inherited or acquired disorder of the hemostatic machinery itself. A large number of such bleeding disorders have been identified.

In addition, unregulated activation of the hemostatic system may cause thrombosis and embolism, which can reduce blood flow to critical organs such as the brain and myocardium. Although we understand less about the pathophysiology of thrombosis than of hemostatic failure, certain patient groups have been identified that are particularly prone to thrombosis and embolism. These include patients who (1) are immobilized after surgery, (2) have chronic congestive heart failure, (3) have atherosclerotic vascular disease, (4) have a malignancy, or (5) are pregnant. Most of these "thrombosis-prone" patients have inherited or acquired "hypercoagulable" or "prethrombotic" disorders.

Certain information in the patient's history, such as the mode of onset and sites of bleeding, a family bleeding tendency, and a record of drug ingestion, helps establish the correct diagnosis. Physical examination can identify bleeding in the skin or joint deformities due to previous hemarthroses. Ultimately, however, bleeding disorders are diagnosed by laboratory tests. General screening tests are used first, to document a systemic disorder, and are then supplemented by specific tests of coagulation protein or platelet function to arrive at an accurate diagnosis.

The hypercoagulable or prethrombotic patient can also be identified by a careful history. Three important clues to this diagnosis are: (1) repeated episodes of thromboembolism without an obvious predisposing condition, (2) a family history of thrombosis, and (3) well-documented thromboembolism in adolescents and young adults. All of the known inherited prethrombotic disorders can be diagnosed with specific immunologic, functional, and, in some cases, genetic tests.

NORMAL HEMOSTASIS

Accurate diagnosis and treatment of patients with either bleeding or thrombosis require knowledge of the pathophysiology of hemostasis. The process can be divided into primary and secondary components and is initiated when trauma, surgery, or disease disrupts the vascular endothelial lining and blood is exposed to subendothelial connective tissue. *Primary hemostasis* is the name given to the process of platelet plug formation at sites of injury. It occurs within seconds of injury and is of prime importance in stopping blood loss from capillaries, small arterioles, and venules (Fig. 62-1). *Secondary hemostasis* consists of the reactions of the plasma coagulation system that result in fibrin formation. It requires several minutes for completion. The fibrin strands that are produced strengthen the primary hemostatic plug. This reaction is particularly important in larger vessels and prevents bleeding from recurring hours or days after the injury. Although presented here as separate events, primary and secondary hemostasis are closely linked. For example, activated platelets accelerate plasma coagulation, and products of the plasma coagulation reaction, such as thrombin, induce platelet activation.

Effective primary hemostasis requires three critical events -- platelet adhesion, granule release, and platelet aggregation. Within a few seconds of injury, platelets adhere to collagen fibrils in vascular subendothelium via a specific platelet collagen receptor, glycoprotein Ia/IIa, which is a member of the integrin family. As shown in Fig. 62-2, this interaction is stabilized by the von Willebrand factor, an adhesive glycoprotein that allows platelets to remain attached to the vessel wall despite the high shear forces generated within the vascular lumen. The von Willebrand factor accomplishes this task by forming a link between a platelet receptor site on glycoprotein Ib/IX and subendothelial collagen fibrils. The adherent platelets then release preformed granule constituents and generate de novo mediators like those depicted inFig. 62-1.

As in other cells, platelet activation and secretion are regulated by changes in the level of cyclic nucleotides, the influx of calcium, hydrolysis of membrane phospholipids, and phosphorylation of critical intracellular proteins. The relevant pathways are depicted inFigs. 62-3 and62-4. The binding of agonists such as epinephrine, collagen, or thrombin to platelet surface receptors activates two membrane enzymes -- phospholipase C and phospholipase A_2. These enzymes catalyze the release of arachidonic acid from two of the major membrane phospholipids, phosphatidylinositol and phosphatidylcholine. Initially, a small quantity of the released arachidonic acid is converted to thromboxane A_2(TXA_2), which, in turn, can activate phospholipase C. The formation of $TXA_2$from arachidonic acid is mediated by the enzyme cyclooxygenase (Fig. 62-3). This enzyme is inhibited by aspirin and nonsteroidal anti-inflammatory drugs. Inhibition of TXA_2synthesis is a cause of mild bleeding in some patients and is the same way some antithrombotic drugs work.

Hydrolysis of the membrane phospholipid phosphatidylinositol 4,5-bisphosphate produces diacylglycerol (DAG) and inositol triphosphate (IP_3), both of which play critical roles in platelet metabolism. IP_3mediates the movement of calcium into the platelet cytosol and stimulates the phosphorylation of myosin light chains. The latter interact with actin to facilitate granule movement and platelet shape change. DAG activates protein kinase C, which, in turn, phosphorylates several substrates, including myosin light chain kinase and a 47-kDa protein (plekstrin). Phosphorylation of these or other proteins may regulate platelet granule secretion.

A finely balanced mechanism controls the rate and extent of platelet activation (Fig. 62-3).TXA_2, a platelet product of arachidonic acid, stimulates platelet activation and secretion. In contrast, prostacyclin, an endothelial cell product of arachidonic acid metabolism, inhibits platelet activation by raising intraplatelet levels of cyclic adenosine monophosphate.

Following activation, platelets secrete their granule contents into plasma. Endoglycosidases and a heparin-cleaving enzyme are released from lysosomes; calcium, serotonin, and adenosine diphosphate (ADP) are released from the dense granules; and several proteins, including the von Willebrand factor, fibronectin, thrombospondin, the platelet-derived growth factor (PDGF), and a heparin-neutralizing protein (platelet factor 4), are released froma granules. Released ADP binds to purinergic receptors, which, when activated, change the conformation of the glycoprotein IIb/IIIa complex so that it binds fibrinogen, linking adjacent platelets into a hemostatic plug (Fig. 62-2). Released PDGF stimulates the growth and migration of fibroblasts and smooth-muscle cells within the vessel wall, an important part of the repair process.

As the primary hemostatic plug is being formed, plasma coagulation proteins are activated to initiate secondary hemostasis. An overall picture of the coagulation scheme, including the role of various inhibitors, is shown inFig. 62-5. The coagulation pathway can be broken down into a series of reactions (Fig. 62-6) that culminate in the production of enough thrombin to convert a small amount of plasma fibrinogen to fibrin. Each of the reactions requires the formation of a surface-bound complex and the conversion of inactive precursor proteins into active proteases by limited proteolysis, and each is regulated by both plasma and cellular cofactors and calcium.

In *reaction 1*, the intrinsic or contact phase of coagulation, three plasma proteins, Hageman factor (factor XII), high-molecular-weight kininogen (HMWK), and prekallikrein (PK), form a complex on vascular subendothelial collagen. After binding to HMWK, factor XII is slowly converted to an active protease (XIIa), which then activates PK to kallikrein and factor XI to XIa. Kallikrein accelerates the conversion of XII to XIIa, while XIa participates in subsequent coagulation reactions. An alternative mechanism for the activation of factor XI may exist, as patients who are deficient in either factor XII, HMWK, or PK have apparently normal hemostasis and no clinical bleeding.

Reaction 2 provides a second pathway to initiate coagulation by activating factor VII to a protease. In this extrinsic or tissue-factor-dependent pathway, a complex is formed between factor VII, calcium, and tissue factor, a ubiquitous lipoprotein present in cellular membranes and exposed by cellular injury. The tissue factor-VII pathway is continuously active and makes a major contribution to basal coagulation. Factor VII and three other coagulation proteins -- factors II (prothrombin), IX, and X -- require calcium and vitamin K for biologic activity. These proteins are synthesized in the liver, where a vitamin K-dependent carboxylase catalyzes a unique posttranslational modification that adds a second carboxyl group to certain glutamic acid residues. Pairs of these di-g-carboxyglutamic acid (Gla) residues bind calcium, which alters protein comformation for binding to phospholipid surfaces and confers biologic activity. Inhibition of this modification by vitamin K antagonists (e.g., warfarin) is the basis of one of the most common forms of anticoagulant therapy.

In *reaction 3*, factor X is activated by the proteases generated in the two previous reactions. In one reaction, a calcium- and lipid-dependent complex is formed between factors VIII, IX, and X. Within this complex, factor IX is first converted to IXa by factor XIa that was generated within the intrinsic pathway (reaction 1). Factor X is then activated by factor IXa in concert with factor VIII. Alternatively, both factors IX and X can be activated more directly by factor VIIa, generated via the extrinsic pathway (reaction 2). Activation of factors IX and X provides a link between the intrinsic and extrinsic coagulation pathways (Fig. 62-5).

Reaction 4, the final step, converts prothrombin to thrombin in the presence of factor V, calcium, and phospholipid. Although prothrombin conversion can take place on various natural and artificial phospholipid-rich surfaces, it proceeds several thousand times faster on the surface of activated platelets or endothelial cells. Thrombin has multiple functions in hemostasis. Although its principal role in hemostasis is the conversion of fibrinogen to fibrin, it also activates factors V, VIII, and XIII and stimulates platelet aggregation and secretion. Following the release of fibrinopeptides A and B from thea and b chains of fibrinogen, the modified molecule, now called *fibrin monomer*, polymerizes into an insoluble gel. The fibrin polymer is then stabilized by the cross-linking of individual chains by factor XIIIa, a plasma transglutaminase (Fig. 62-5).

Although the classic view of coagulation had clinical utility, it left several important questions unanswered: (1) Why does factor XII deficiency dramatically prolong partial thromboplastin time (PTT) but not cause bleeding? (2) Why is there heterogeneity in the bleeding symptoms of patients with factor XI deficiency? (3) Why do deficiencies in factors VIII or IX produce such dramatic bleeding even though the "extrinsic" pathway remains intact? Activation of factors IX and X by the tissue factor-VIIa complex is thought to play a major role in the initiation of hemostasis. Once coagulation is initiated by this interaction, the tissue factor pathway inhibitor (TFPI) blocks the pathway, and elements of the intrinsic pathway, particularly factors VIII and IX, become the dominant regulators of thrombin generation. This step in the pathway explains why factor XII-deficient patients are asymptomatic and why factor XI-deficient patients have only a mild to moderate bleeding diathesis (Fig. 62-7).

Clot lysis and vessel repair begin immediately after the formation of the definitive hemostatic plug. Three potential activators of the fibrinolytic system are: Hageman factor fragments, urinary plasminogen activator (uPA) or urokinase, and tissue plasminogen activator (tPA). The principal physiologic activators, tPA and uPA, diffuse from endothelial cells and convert plasminogen, adsorbed to the fibrin clot, into plasmin (Fig. 62-8). Plasmin then degrades fibrin polymer into small fragments, which are cleared by the monocyte-macrophage scavenger system. Although plasmin can also degrade fibrinogen, the reaction remains localized because (1) tPA and some forms of uPA activate plasminogen more effectively when it is adsorbed to fibrin clots; (2) any plasmin that enters the circulation is rapidly bound and neutralized by thea$_2$plasmin inhibitor (patients who lack this factor have unchecked fibrinolysis and bleed); and (3) endothelial cells release a plasminogen activator inhibitor (PAI-1), which blocks the action of tPA.

Only a small quantity of each coagulation enzyme is converted to its active form. As a consequence, the hemostatic plug does not propagate beyond the site of injury. Precise regulation is important, since each milliliter of blood contains enough clotting potential to clot all the fibrinogen in the body in 10 to 15 s. Blood fluidity is maintained by the flow of blood, the adsorption of coagulation factors to surfaces and their trapping in the emerging clot, and by multiple inhibitors in plasma. Antithrombin, proteins C and S, andTFPI are important inhibitors that maintain blood fluidity.

These inhibitors have distinct modes of action. Antithrombin forms complexes with all serine protease coagulation factors except factor VII (Fig. 62-5). Rates of complex formation are accelerated by heparin and heparin-like molecules on the surface of the endothelial cells. Heparin's ability to accelerate antithrombin activity is the basis for its anticoagulant action. Protein C is converted to an active protease by thrombin after it is bound to an endothelial cell protein called *thrombomodulin*. Activated protein C then inactivates the two plasma cofactors V and VIII by limited proteolysis, which slows down two critical coagulation reactions. Protein C may also stimulate the release oftPA from endothelial cells. The inhibitory function of protein C is enhanced by protein S. Reduced levels of antithrombin or proteins C and S, or dysfunctional forms of these molecules, result in a hypercoagulable or prethrombotic state. In addition, a particularly common heritable defect associated with a hypercoagulable state is the presence of a form of factor V (factor V Leiden) that is resistant to protein C inhibition. Between 20 and 50% of patients with unexplained venous thromboembolism have this defect.

Blood coagulation is not uniform throughout the body. The composition of the blood clot varies with the site of injury. Hemostatic plugs or thrombi that form in veins where blood flow is slow are rich in fibrin and trapped red blood cells and contain relatively few platelets. They are often called *red thrombi* because of their appearance in surgical and pathologic specimens. The friable ends of these red thrombi, which most often form in leg veins, can break off and embolize to the pulmonary circulation. Conversely, clots that form in arteries under conditions of high flow are predominantly composed of platelets and have little fibrin. These *white thrombi* may readily dislodge from the arterial wall and embolize to distant sites, causing temporary or permanent ischemia. These clots are a particularly common cause of embolism in the cerebral and retinal circulation, where they may lead to transient neurologic dysfunction (transient ischemic attacks), including temporary monocular blindness (amaurosis fugax), or to strokes. In addition, most episodes of myocardial infarction are due to thrombi that form after the rupture of atherosclerotic plaques within diseased coronary arteries. Hemostatic plugs, which are a physiologic response to injury, are very similar to pathologic thrombi. Thrombosis has been described as coagulation occurring in the wrong place or at the wrong time.

CLINICAL EVALUATION

HISTORY

Certain elements of the history are particularly useful in determining whether bleeding is caused by an underlying hemostatic disorder or by a local anatomic defect. One clue is a history of bleeding following common hemostatic stresses such as dental extraction, childbirth, or minor surgery. Bleeding that is sufficiently severe to require a blood transfusion merits special attention. A family history of bleeding and bleeding from multiple sites that cannot be linked to trauma or surgery also suggest a systemic disorder. Since bleeding can be mild, lack of a family history of bleeding does not exclude an inherited hemostatic disorder.

Bleeding from a platelet disorder is usually localized to superficial sites such as the skin and mucous membranes, comes on immediately after trauma or surgery, and is readily controlled by local measures (Table 62-1). In contrast, bleeding from secondary hemostatic or plasma coagulation defects occurs hours or days after injury and is unaffected by local therapy. Such bleeding most often occurs in deep subcutaneous tissues, muscles, joints, or body cavities. A careful and thorough history may establish the presence of a hemostatic disorder and guide initial laboratory testing.

PHYSICAL EXAMINATION

The most common site to observe bleeding is in the skin and mucous membranes. Collections of blood in the skin are called *purpura* and may be subdivided on the basis of the site of bleeding in the skin. Small pinpoint hemorrhages into the dermis due to the leakage of red cells through capillaries are called *petechiae* and are characteristic of platelet disorders -- in particular, severe thrombocytopenia. Larger subcutaneous collections of blood due to leakage of blood from small arterioles and venules are called *ecchymoses* (common bruises) or, if somewhat deeper and palpable, *hematomas*. They are also common in patients with platelet defects and result from minor trauma. Dilated capillaries, or *telangiectasia*, may cause bleeding without any hemostatic defect. In addition, the loss of connective tissue support for capillaries and small veins that accompanies aging increases the fragility of superficial vessels, such as those on the dorsum of the hand, leading to extravasation of blood into subcutaneous tissue -- *senile purpura*. Menorrhagia is sometimes a serious problem in women with severe thrombocytopenia or platelet dysfunction. Some patients with primary hemostatic defects, especially von Willebrand's disease, may have recurrent gastrointestinal hemorrhage, often associated with angiodysplasia, a common vascular malformation in the gastrointestinal tract.

Bleeding into body cavities, the retroperitoneum, or joints is a common manifestation of plasma coagulation defects. Repeated joint bleeding may cause synovial thickening, chronic inflammation, and fluid collections and may erode articular cartilage and lead to chronic joint deformity and limited mobility. Such deformities are particularly common in deficiencies of factors VIII and IX, the two sex-linked coagulation disorders referred to as the *hemophilias*. For unclear reasons, hemarthroses are much less common in patients with other plasma coagulation defects. Blood collections in various body cavities or soft tissues can cause secondary necrosis of tissues or nerve compression. Retroperitoneal hematomas can cause femoral nerve compression, and large collections of poorly coagulated blood in soft tissues occasionally mimic malignant growths -- the pseudotumor syndrome. Two of the most life-threatening sites of bleeding are in the oropharynx, where bleeding can compromise the airway, and in the central nervous system. Intracerebral hemorrhage is one of the leading causes of death in patients with severe coagulation disorders. Because of their need for plasma and factor concentrates derived from multiple donors, many patients with hemophilia were infected with HIV before effective testing of donors was in place.

LABORATORY TESTS

The most important screening tests of the primary hemostatic system are (1) a *bleeding time* (a sensitive measure of platelet function), and (2) a *platelet count*. The latter correlates well with the propensity to bleed. The normal platelet count is 150,000 to 450,000/uL of blood. As long as the count is>100,000/uL, patients are usually not symptomatic and the bleeding time remains normal. Platelet counts of 50,000 to 100,000/uL cause mild prolongation of the bleeding time; bleeding occurs only from severe trauma or other stress. Patients with platelet counts<50,000/uL have easy bruising, manifested by skin purpura after minor trauma and bleeding after mucous membrane surgery. Patients with a platelet count <20,000/uL have an appreciable incidence of spontaneous bleeding, usually have petechiae, and may have intracranial or other spontaneous internal bleeding. The major causes of thrombocytopenia are outlined in Table 62-2.

Patients with qualitative platelet abnormalities have a normal platelet count and a prolonged bleeding time (Table 62-3). The bleeding time is ascertained by making a small, superficial skin incision and timing the duration of blood flow from the wounded area. By careful standardization, bleeding time is a reliable and sensitive test of platelet function. A template or an automated scalpel controls the length and depth of the incision (usually 1 mm deep by 9 mm long), and a sphygmomanometer inflated to 40 mmHg distends the capillary bed of the forearm uniformly. The bleeding time test must be performed by an experienced technician, as small differences in technique have a big effect on outcome. Any patient with a bleeding time >10 min has an increased risk of bleeding, but the risk does not become great until the bleeding time >15 or 20 min. As shown inFig. 62-9, the relationship between the platelet count and the bleeding time is roughly linear. When a defect in primary hemostasis is uncovered, specialized testing is needed to determine the cause of the platelet dysfunction (Table 62-3). A precise diagnosis is important in determining the proper treatment. Occasional patients with a strong history of bleeding, particularly those with mild von Willebrand's disease, may have a normal bleeding time when initially tested, owing to cyclical variations in the level of the von Willebrand factor. Repeated testing may be necessary to establish an accurate diagnosis. Bleeding time is not an effective screening test for preoperative patients.

Plasma coagulation function is readily assessed with thePTT, prothrombin time (PT), thrombin time (TT), and quantitative fibrinogen determination (Fig. 62-5,Table 62-4). The PTT screens the intrinsic limb of the coagulation system and tests for the adequacy of factors XII,HMWK,PK, XI, IX, and VIII. The PT screens the extrinsic or tissue factor-dependent pathway. Both tests also evaluate the common coagulation pathway involving all the reactions that occur after the activation of factor X. Prolongation of the PT and PTT that does not resolve after the addition of normal plasma suggests a coagulation inhibitor. A specific test for the conversion of fibrinogen to fibrin is needed when both the PTT and PT are prolonged -- either a TT or a clottable fibrinogen level can be employed. When abnormalities are noted in any of the screening tests, more specific coagulation factor assays can be ordered to determine the nature of the defect.

Several rare coagulation abnormalities that may be missed as they do not affect these screening tests: factor XIII deficiency,$a_2$plasmin inhibitor deficiency,PAI-1deficiency (PAI-1 is the major inhibitor of plasminogen activators), and Scott's syndrome, a platelet coagulant defect. A test for factor XIII-dependent fibrin cross-linking, such as clot solubility in 5 *M* urea, should be ordered when thePT andPTT are both normal but the history of bleeding is strong. The fibrinolytic system can be assessed by measuring the rate of clot lysis with the euglobulin lysis or whole blood clot lysis tests and by measuring the levels of$a_2$plasmin inhibitor and PAI-1. Scott's syndrome can be detected by measuring the serum PT, which assesses the amount of residual prothrombin.

Conditions associated with thrombosis are listed in Table 62-5. Patients suspected of having a hypercoagulable or prethrombotic disorder on the basis of clinical information should be tested with specific assays to screen for the known defects. Currently available tests can identify 50 to 60% of the cases of familial or recurrent venous thrombosis.

Inhibitor syndromes or circulating anticoagulants are usually due to antibodies that impair coagulation factor activity. They are an infrequent cause of bleeding and require specialized diagnostic testing. Inhibitors are likely when screening test abnormalities cannot be reversed by adding normal plasma to patient plasma. Antibodies against specific coagulation factors may develop in (1) postpartum women, (2) patients with autoimmune disorders such as systemic lupus erythematosus, (3) patients taking drugs such as penicillin and streptomycin, and (4) otherwise healthy elderly individuals. In addition, between 10 to 20% of patients with severe hemophilia who have received multiple plasma infusions develop inhibitory antibodies. Some patients, especially those with systemic lupus erythematosus, may also have a nonspecific form of anticoagulant antibody that interferes with phospholipid binding of coagulation factors and prolongs the PTand PTTbut does not cause clinical bleeding. The presence of the lupus anticoagulant may increase the risk of thromboembolism and may cause placental infarction, recurrent midtrimester abortion, and venous and arterial thrombosis. The lupus-like anticoagulant is one manifestation of the anticardiolipin antibody syndrome. Patients may have anticardiolipin antibodies that do not prolong the PTT, but patients are still at risk from thrombosis. Occasionally, patients develop inhibitors that are not antibodies. For example, several patients with clinical bleeding have been found to have circulating mucopolysaccharides that have heparin-like activity.

(Bibliography omitted in Palm version)

Back to Table of Contents

63. ENLARGEMENT OF LYMPH NODES AND SPLEEN - *Patrick H. Henry*, *Dan L. Longo*

This chapter is intended to serve as a guide to the evaluation of patients who present with enlargement of the lymph nodes (*lymphadenopathy*) or the spleen (*splenomegaly*). Lymphadenopathy is a rather common clinical finding in primary care settings, whereas palpable splenomegaly is less so.

LYMPHADENOPATHY

Lymphadenopathy may be an incidental finding in patients being examined for various reasons or it may be a presenting sign or symptom of the patient's illness. The physician must eventually decide whether the lymphadenopathy is a normal finding or one that requires further study, up to and including biopsy. Soft, flat, submandibular nodes (<1 cm) are often palpable in healthy children and young adults, and healthy adults may have palpable inguinal nodes of up to 2 cm, which are considered normal. Further evaluation of these normal nodes is not warranted. In contrast, if the physician believes the node(s) to be abnormal, then pursuit of a more precise diagnosis is needed.

Approach to the Patient

Lymphadenopathy may be a primary or secondary manifestation of numerous disorders, as shown inTable 63-1. Many of these disorders are infrequent causes of lymphadenopathy. Analysis of lymphadenopathy in primary care practice has shown that more than two-thirds of patients have nonspecific causes or upper respiratory illnesses (viral or bacterial), and fewer than 1% have a malignancy. In one study, researchers reported that 186 of 220 patients (84%) referred for evaluation of lymphadenopathy had a "benign" diagnosis. The remaining 34 patients (16%) had a malignancy (lymphoma or metastatic adenocarcinoma). Sixty-three percent (112) of the 186 patients with benign lymphadenopathy had a nonspecific or reactive etiology (no causative agent found), and the remainder had a specific cause demonstrated, most commonly infectious mononucleosis, toxoplasmosis, or tuberculosis. Thus, the vast majority of patients with lymphadenopathy will have a nonspecific etiology requiring few diagnostic tests.

Clinical Assessment The physician will be aided in the pursuit of an explanation for the lymphadenopathy by a careful medical history, physical examination, selected laboratory tests, and perhaps an excisional lymph node biopsy.

The *medical history* should reveal the setting in which lymphadenopathy is occurring. Symptoms such as sore throat, cough, fever, night sweats, fatigue, weight loss, or pain in the nodes should be sought. The patient's age, sex, occupation, exposure to pets, sexual behavior, and use of drugs such as diphenylhydantoin are other important historic points. For example, children and young adults usually have benign (i.e., nonmalignant) disorders, such as viral or bacterial upper respiratory infections, infectious mononucleosis, toxoplasmosis, and, in some countries, tuberculosis, which account for the observed lymphadenopathy. In contrast, after age 50 the incidence of malignant disorders increases and that of benign disorders decreases.

The *physical examination* can provide useful clues such as the extent of lymphadenopathy (localized or generalized), size of nodes, texture, presence or absence of nodal tenderness, signs of inflammation over the node, skin lesions, and splenomegaly. A thorough ear, nose, and throat (ENT) examination is indicated in adult patients with cervical adenopathy and a history of tobacco use. Localized or regional adenopathy implies involvement of a single anatomic area. Generalized adenopathy has been defined as involvement of three or more noncontiguous lymph node areas. Many of the causes of lymphadenopathy (Table 63-1) can produce localized *or* generalized adenopathy, so this distinction is of limited utility in the differential diagnosis. Nevertheless, generalized lymphadenopathy is frequently associated with nonmalignant disorders such as infectious mononucleosis [Epstein-Barr virus (EBV) or cytomegalovirus (CMV)], toxoplasmosis, AIDS, other viral infections, systemic lupus erythematosus (SLE), and mixed connective tissue disease. Acute and chronic lymphocytic leukemias and malignant lymphomas also produce generalized adenopathy in adults.

The site of localized or regional adenopathy may provide a useful clue about the cause. Occipital adenopathy often reflects an infection of the scalp, and preauricular adenopathy accompanies conjunctival infections and cat-scratch disease. The most frequent site of regional adenopathy is the neck, and most of the causes are benign -- upper respiratory infections, oral and dental lesions, infectious mononucleosis, other viral illnesses. The chief malignant causes include metastatic cancer from head and neck, breast, lung, and thyroid primaries. Enlargement of supraclavicular and scalene nodes is always abnormal. Because these nodes drain regions of the lung and retroperitoneal space, they can reflect either lymphomas, other cancers, or infectious processes arising in these areas. Virchow's node is an enlarged left supraclavicular node infiltrated with metastatic cancer from a gastrointestinal primary. Metastases to supraclavicular nodes also occur from lung, breast, testis, or ovarian cancers. Tuberculosis, sarcoidosis, and toxoplasmosis are nonneoplastic causes of supraclavicular adenopathy. Axillary adenopathy is usually due to injuries or localized infections of the ipsilateral upper extremity. Malignant causes include melanoma or lymphoma and, in women, breast cancer. Inguinal lymphadenopathy is usually secondary to infections or trauma of the lower extremities and may accompany sexually transmitted diseases such as lymphogranuloma venereum, primary syphilis, genital herpes, or chancroid. These nodes may also be involved by lymphomas and metastatic cancer from primary lesions of the rectum, genitalia, or lower extremities (melanoma).

The size and texture of the lymph node(s) and the presence of pain are useful parameters in evaluating a patient with lymphadenopathy. Nodes <1.0 cm^2 in area (1.0 ´ 1.0 cm or less) are almost always secondary to benign, nonspecific reactive causes. In one retrospective analysis of younger patients (9 to 25 years) who had a lymph node biopsy, a maximum diameter of >2 cm served as one discriminant for predicting that the biopsy would reveal malignant or granulomatous disease. Another study showed that a lymph node size of 2.25 cm^2 (1.5 cm ´ 1.5 cm) was the best discriminating limit for distinguishing malignant or granulomatous lymphadenopathy from other causes of lymphadenopathy. Patients with node(s) £1.0 cm^{2}should be observed after excluding infectious mononucleosis and/or toxoplasmosis unless there are symptoms and signs of an underlying systemic illness.

The texture of lymph nodes may be described as soft, firm, rubbery, hard, discrete, matted, tender, movable, or fixed. Tenderness is found when the capsule is stretched during rapid enlargement, usually secondary to an inflammatory process. Some malignant diseases such as acute leukemia may produce rapid enlargement and pain in the nodes. Nodes involved by lymphoma tend to be large, discrete, symmetric, rubbery, firm, mobile, and nontender. Nodes containing metastatic cancer are often hard, nontender, and nonmovable because of fixation to surrounding tissues. The coexistence of splenomegaly in the patient with lymphadenopathy implies a systemic illness such as infectious mononucleosis, lymphoma, acute or chronic leukemia,SLE, sarcoidosis, toxoplasmosis, cat-scratch disease, or other less common hematologic disorders. The patient's story should provide helpful clues about the underlying systemic illness.

Nonsuperficial presentations (thoracic or abdominal) of adenopathy are usually detected as the result of a symptom-directed diagnostic workup. Thoracic adenopathy may be detected by routine chest roentgenography or during the workup for superficial adenopathy. It may also be found because the patient complains of a cough or wheezing from airway compression; hoarseness from recurrent laryngeal nerve involvement; dysphagia from esophageal compression; or swelling of the neck, face, or arms secondary to compression of the superior vena cava or subclavian vein. The differential diagnosis of mediastinal and hilar adenopathy includes primary lung disorders and systemic illnesses that characteristically involve mediastinal or hilar nodes. In the young, mediastinal adenopathy is associated with infectious mononucleosis and sarcoidosis. In endemic regions, histoplasmosis can cause unilateral paratracheal lymph node involvement that mimics lymphoma. Tuberculosis can also cause unilateral adenopathy. In older patients, the differential diagnosis includes primary lung cancer (especially among smokers), lymphomas, metastatic carcinoma (usually lung), tuberculosis, fungal infection, and sarcoidosis.

Enlarged intraabdominal or retroperitoneal nodes are usually malignant. Although tuberculosis may present as mesenteric lymphadenitis, these masses usually contain lymphomas or, in young men, germ cell tumors.

Laboratory Investigation The laboratory investigation of patients with lymphadenopathy must be tailored to elucidate the etiology suspected from the patient's history and physical findings. One study from a family practice clinic evaluated 249 younger patients with "enlarged lymph nodes, not infected" or "lymphadenitis." No laboratory studies were obtained in 51%. When studies were performed, the most common were a complete blood count (33%), throat culture (16%), chest x-ray (12%), or monospot test (10%). Only eight patients (3%) had a node biopsy, and half of those were normal or reactive. The complete blood count can provide useful data for the diagnosis of acute or chronic leukemias,EBV orCMVmononucleosis, lymphoma with a leukemic component, pyogenic infections, or immune cytopenias in illnesses such asSLE. Serologic studies may demonstrate antibodies specific to components of EBV, CMV, HIV, and other viruses; *Toxoplasma gondii*; *Brucella*; etc. If SLE is suspected, then antinuclear and anti-DNA antibody studies are warranted.

The chest x-ray is usually negative, but the presence of a pulmonary infiltrate or mediastinal lymphadenopathy would suggest tuberculosis, histoplasmosis, sarcoidosis, lymphoma, primary lung cancer, or metastatic cancer and demands further investigation.

A variety of imaging techniques [computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, color Doppler ultrasonography] have been employed to differentiate benign from malignant lymph nodes, especially in patients with head and neck cancer. CT and MRI are comparably accurate (65 to 90%) in the diagnosis of metastases to cervical lymph nodes. Ultrasonography has been used to determine the long (L) axis, short (S) axis, and a ratio of long to short axis in cervical nodes. An L/S ratio of <2.0 has a sensitivity and a specificity of 95% for distinguishing benign and malignant nodes in patients with head and neck cancer. This ratio has greater specificity and sensitivity than palpation or measurement of either the long or the short axis alone.

The indications for lymph node biopsy are imprecise, yet it is a valuable diagnostic tool. The decision to biopsy may be made early in a patient's evaluation or delayed for up to 2 weeks. Prompt biopsy should occur if the patient's history and physical findings suggest a malignancy; examples include a solitary, hard, nontender cervical node in an older patient who is a chronic user of tobacco; supraclavicular adenopathy; and solitary or generalized adenopathy that is firm, movable, and suggestive of lymphoma. If a primary head and neck cancer is suspected as the basis of a solitary, hard cervical node, then a carefulENTexamination should be performed. Any mucosal lesion that is suspicious for a primary neoplastic process should be biopsied first. If no mucosal lesion is detected, an excisional biopsy of the largest node should be performed. Fine-needle aspiration should not be performed as the first diagnostic procedure. Most diagnoses require more tissue than such aspiration can provide and it often delays a definitive diagnosis. Fine-needle aspiration should be reserved for thyroid nodules and for confirmation of relapse in patients whose primary diagnosis is known. If the primary physician is uncertain about whether to proceed to biopsy, consultation with a hematologist or medical oncologist should be helpful. In primary care practices, fewer than 5% of lymphadenopathy patients will require a biopsy. That percentage will be considerably larger in referral practices, i.e., hematology, oncology, or otolaryngology (ENT).

Two groups have reported algorithms that they claim will identify more precisely those lymphadenopathy patients who should have a biopsy. Both reports were retrospective analyses in referral practices. The first study involved patients 9 to 25 years of age who had a node biopsy performed. Three variables were identified that predicted those young patients with peripheral lymphadenopathy who should undergo biopsy; lymph node size >2 cm in diameter and abnormal chest x-ray had positive predictive value, whereas recentENTsymptoms had negative predictive values. The second study evaluated 220 lymphadenopathy patients in a hematology unit and identified five variables [lymph node size, location (supraclavicular or non-supraclavicular), age (>40 years or <40 years), texture (nonhard or hard), and tenderness] that were utilized in a mathematical model to identify those patients requiring a biopsy. Positive predictive value was found for age >40 years, supraclavicular location, node size >2.25 cm$_2$, hard texture, and lack of pain or tenderness. Negative predictive value was evident for age <40 years, node size<1.0 cm$_2$, nonhard texture, and tender or painful nodes. Ninety-one percent of those who required biopsy were correctly classified by this model. Since both of these studies were retrospective analyses and one was limited to young patients, it is not known how useful these models would be if applied prospectively in a primary care setting.

Most lymphadenopathy patients do not require a biopsy, and at least half require no laboratory studies. If the patient's history and physical findings point to a benign cause for lymphadenopathy, then careful follow-up at a 2- to 4-week interval can be employed. The patient should be instructed to return for reevaluation if the node(s) increase in size. Antibiotics are not indicated for lymphadenopathy unless there is strong evidence of a bacterial infection. Glucocorticoids should not be used to treat lymphadenopathy because their lympholytic effect obscures some diagnoses (lymphoma, leukemia, Castleman's disease) and they contribute to delayed healing or activation of underlying infections. An exception to this statement is the life-threatening pharyngeal obstruction by enlarged lymphoid tissue in Waldeyer's ring that is occasionally seen in infectious mononucleosis.

SPLENOMEGALY

STRUCTURE AND FUNCTION OF THE SPLEEN

The spleen is a reticuloendothelial organ that has its embryologic origin in the dorsal mesogastrium at about 5 weeks' gestation. It arises in a series of hillocks, migrates to its normal adult location in the left upper quadrant (LUQ), and is attached to the stomach via the gastrolienal ligament and to the kidney via the lienorenal ligament. When the hillocks fail to unify into a single tissue mass, accessory spleens may develop in around 20% of persons. The function of the spleen has been elusive. Galen believed it was the source of "black bile" or melancholia, and the word *hypochondria* (literally, beneath the ribs) and the idiom "to vent one's spleen" attest to the beliefs that the spleen had an important influence on the psyche and emotions. In humans, its normal physiologic roles seem to be the following:

1. Maintenance of quality control over erythrocytes in the red pulp by removal of senescent and defective red blood cells. The spleen accomplishes this function through a unique organization of its parenchyma and vasculature (Fig. 63-1).

2. Synthesis of antibodies in the white pulp.

3. The removal of antibody-coated bacteria and antibody-coated blood cells from the circulation.

An increase in these normal functions may result in splenomegaly.

The spleen is composed of red pulp and white pulp, which are Malpighi's terms for the red blood-filled sinuses and reticuloendothelial cell-lined cords and the white lymphoid follicles arrayed within the red pulp matrix. The spleen is in the portal circulation. The reason for this is unknown but may relate to the fact that lower blood pressure allows less rapid flow and minimizes damage to normal erythrocytes. Blood flows into the spleen at a rate of about 150 mL/min through the splenic artery, which ultimately ramifies into central arterioles. Some blood goes from the arterioles to capillaries and then to splenic veins and out of the spleen, but the majority of blood from central arterioles flows into the macrophage-lined sinuses and cords. The blood entering the sinuses reenters the circulation through the splenic venules, but the blood entering the cords is subjected to an inspection of sorts. In order to return to the circulation, the blood cells in the cords must squeeze through slits in the cord lining to enter the sinuses that lead to the venules. Old and damaged erythrocytes are less deformable and are retained in the cords, where they are destroyed and their components recycled. Red cell inclusion bodies such as parasites, nuclear residua (Howell-Jolly bodies), or denatured hemoglobin (Heinz bodies) are pinched off in the process of passing through the slits, a process called *pitting*. The culling of dead and damaged cells and the pitting of cells with inclusions appear to occur without significant delay since the blood transit time through the spleen is only slightly slower than in other organs.

The spleen is also capable of assisting the host in adapting to its hostile environment. It has at least three adaptational functions: (1) clearance of bacteria and particulates from the blood, (2) the generation of immune responses to certain invading pathogens, and (3) the generation of cellular components of the blood under circumstances in which the marrow is unable to meet the needs (i.e., extramedullary hematopoiesis). The latter adaptation is a recapitulation of the blood-forming function the spleen plays during gestation. In some animals, the spleen also serves a role in the vascular adaptation to stress because it stores red blood cells (often hemoconcentrated to higher hematocrits than normal) under normal circumstances and contracts under the influence ofb-adrenergic stimulation to provide the animal with an autotransfusion and improved oxygen-carrying capacity. However, the normal human spleen does not sequester or store red blood cells and does not contract in response to sympathetic stimuli. The normal human spleen contains approximately one-third of the total body platelets and a significant number of marginated neutrophils. These sequestered cells are available when needed to respond to bleeding or infection.

Approach to the Patient

Clinical Assessment The most common *symptoms* produced by diseases involving the spleen are pain and a heavy sensation in theLUQ. Massive splenomegaly may cause early satiety. Pain may result from acute swelling of the spleen with stretching of the capsule, infarction, or inflammation of the capsule. For many years, it was believed that splenic infarction was clinically silent, which at times is true. However, Soma Weiss, in his classic 1942 report of the self-observations by a Harvard medical student on the clinical course of subacute bacterial endocarditis, documented that severe LUQ and pleuritic chest pain may accompany thromboembolic occlusion of splenic blood flow. Vascular occlusion, with infarction and pain, is commonly seen in children with sickle cell crises. Rupture of the spleen, either from trauma or infiltrative disease that breaks the capsule, may result in intraperitoneal bleeding, shock, and death. The rupture itself may be painless.

A palpable spleen is the major *physical sign* produced by diseases affecting the spleen and suggests enlargement of the organ. The normal spleen is said to weigh less than 250 g, decreases in size with age, normally lies entirely within the rib cage, has a maximum cephalocaudad diameter of 13 cm by ultrasonography or maximum length of 12 cm and/or width of 7 cm by radionuclide scan, and is usually not palpable. However, a palpable spleen was found in 3% of 2200 asymptomatic, male, freshman college students. Follow-up at 3 years revealed that 30% of those students still had a palpable spleen without any increase in disease prevalence. Ten-year follow-up found no evidence for lymphoid malignancies. Furthermore, in some tropical countries (e.g., New Guinea) the incidence of splenomegaly may reach 60%. Thus, the presence of a palpable spleen does not always equate with presence of disease. Even when disease is present, splenomegaly may not reflect the primary disease, but rather a reaction to it. For example, in patients with Hodgkin's disease, only two-thirds of the palpable spleens show involvement by the cancer.

Physical examination of the spleen utilizes primarily the techniques of palpation and percussion. Inspection may reveal a fullness in theLUQ that descends on inspiration, a finding associated with a massively enlarged spleen. Auscultation may reveal a venous hum or a friction rub.

Palpation can be accomplished by bimanual palpation, ballotment, and palpation from above (Middleton maneuver). For bimanual palpation, which is at least as reliable as the other techniques, the patient is supine with flexed knees. The examiner's left hand is placed on the lower rib cage and pulls the skin toward the costal margin, allowing the fingertips of the right hand to feel the tip of the spleen as it descends while the patient inspires slowly, smoothly, and deeply. Palpation is begun with the right hand in the left lower quadrant with gradual movement toward the left costal margin, thereby identifying the lower edge of a massively enlarged spleen. When the spleen tip is felt, the finding is recorded as centimeters below the left costal margin at some arbitrary point, i.e., 10 to 15 cm, from the midpoint of the umbilicus or the xiphisternal junction. This allows other examiners to compare findings or the initial examiner to determine changes in size over time. Bimanual palpation in the right lateral decubitus position adds nothing to the supine examination.

Percussion for splenic dullness is accomplished with any of three techniques described by Nixon, Castell, or Barkun:

1. *Nixon's method*: The patient is placed on the right side so that the spleen lies above the colon and stomach. Percussion begins at the lower level of pulmonary resonance in the posterior axillary line and proceeds diagonally along a perpendicular line toward the lower midanterior costal margin. The upper border of dullness is normally 6 to 8 cm above the costal margin. Dullness greater than 8 cm in an adult is presumed to indicate splenic enlargement.

2. *Castell's method*: With the patient supine, percussion in the lowest intercostal space in the anterior axillary line (8th or 9th) produces a resonant note if the spleen is normal in size. This is true during expiration or full inspiration. A dull percussion note on full inspiration suggests splenomegaly.

3. *Percussion of Traube's semilunar space*: The borders of Traube's space are the sixth rib superiorly, the left midaxillary line laterally, and the left costal margin inferiorly. The patient is supine with the left arm slightly abducted. During normal breathing, this space is percussed from medial to lateral margins, yielding a normal resonant sound. A dull percussion note suggests splenomegaly.

Studies comparing methods of percussion and palpation with a standard of ultrasonography or scintigraphy have revealed sensitivity of 56 to 71% for palpation and 59 to 82% for percussion. Reproducibility among examiners is better for palpation than percussion. Both techniques are less reliable in obese patients or patients who have just eaten. Thus, the physical examination techniques of palpation and percussion are imprecise at best. It has been suggested that the examiner perform percussion first and, if positive, proceed to palpation; if the spleen is palpable, then one can be reasonably confident that splenomegaly exists. However, not allLUQmasses are enlarged spleens; gastric or colon tumors and pancreatic or renal cysts or tumors can mimic splenomegaly.

The presence of an enlarged spleen can be more precisely determined, if necessary, by liver-spleen radionuclide scan,CT,MRI, or ultrasonography. The latter technique is the current procedure of choice for routine assessment of spleen size (normal = a maximum cephalocaudad diameter of 13 cm) because it has high sensitivity and specificity and is safe, noninvasive, quick, mobile, and less costly. Nuclear medicine scans are accurate, sensitive, and reliable but are costly, require greater time to generate data, and utilize immobile equipment. They have the advantage of demonstrating accessory splenic tissue. CT and MRI provide accurate determination of spleen size, but the equipment is immobile and the procedures are expensive. MRI appears to offer no advantage over CT. Changes in spleen structure such as mass lesions, infarcts, inhomogeneous infiltrates, and cysts are more readily assessed by CT, MRI, or ultrasonography. None of these techniques is very reliable in the detection of patchy infiltration (e.g., Hodgkin's disease).

Differential Diagnosis Many of the diseases associated with splenomegaly are listed in Table 63-2. They are grouped according to the presumed basic mechanisms responsible for organ enlargement:

1. Hyperplasia or hypertrophy related to a particular splenic function such as reticuloendothelial hyperplasia (work hypertrophy) in diseases such as hereditary spherocytosis or thalassemia syndromes that require removal of large numbers of defective red blood cells; immune hyperplasia in response to systemic infection (infectious mononucleosis, subacute bacterial endocarditis) or to immunologic diseases (immune thrombocytopenia,SLE, Felty's syndrome).

2. Passive congestion due to decreased blood flow from the spleen in conditions that produce portal hypertension (cirrhosis, Budd-Chiari syndrome, congestive heart failure).

3. Infiltrative diseases of the spleen (lymphomas, metastatic cancer, amyloidosis, Gaucher's disease, myeloproliferative disorders with extramedullary hematopoiesis).

The differential diagnostic possibilities are much fewer when the spleen is "massively enlarged," that is, it is palpable more than 8 cm below the left costal margin or its drained weight is ^{3}1000 g (Table 63-3). The vast majority of such patients will have non-Hodgkin's lymphoma, chronic lymphocytic leukemia, hairy cell leukemia, chronic myelogenous leukemia, myelofibrosis with myeloid metaplasia, or polycythemia vera.

Laboratory Assessment The major laboratory abnormalities accompanying splenomegaly are determined by the underlying systemic illness. Erythrocyte counts may be normal, decreased (thalassemia major syndromes,SLE, cirrhosis with portal hypertension), or increased (polycythemia vera). Granulocyte counts may be normal, decreased (Felty's syndrome, congestive splenomegaly, leukemias), or increased (infections or inflammatory disease, myeloproliferative disorders). Similarly, the platelet count may be normal, decreased when there is enhanced sequestration or destruction of platelets in an enlarged spleen (congestive splenomegaly, Gaucher's disease, immune thrombocytopenia), or increased in the myeloproliferative disorders such as polycythemia vera.

The complete blood count may reveal cytopenia of one or more blood cell types, which should suggest *hypersplenism*. This condition is characterized by splenomegaly, cytopenia(s), normal or hyperplastic bone marrow, and a response to splenectomy. The latter characteristic is less precise because reversal of cytopenia, particularly granulocytopenia, is sometimes not sustained after splenectomy. The cytopenias result from increased destruction of the cellular elements secondary to reduced flow of blood through enlarged and congested cords (congestive splenomegaly) or to immune-mediated mechanisms. In hypersplenism, various cell types usually have normal morphology on the peripheral blood smear, although the red cells may be spherocytic due to loss of surface area during their longer transit through the enlarged spleen. The increased marrow production of red cells should be reflected as an increased reticulocyte production index, although the value may be less than expected due to increased sequestration of reticulocytes in the spleen.

The need for additional laboratory studies is dictated by the differential diagnosis of the underlying illness of which splenomegaly is a manifestation.

SPLENECTOMY

Splenectomy is infrequently performed for diagnostic purposes, especially in the absence of clinical illness or other diagnostic tests that suggest underlying disease. More often splenectomy is performed for staging the extent of disease in patients with Hodgkin's disease, for symptom control in patients with massive splenomegaly, for disease control in patients with traumatic splenic rupture, or for correction of cytopenias in patients with hypersplenism or immune-mediated destruction of one or more cellular blood elements. Splenectomy is necessary for routine staging of patients with Hodgkin's disease only in those with clinical stage I or II disease in whom radiation therapy alone is contemplated as the treatment. Noninvasive staging of the spleen in Hodgkin's disease is not a sufficiently reliable basis for treatment decisions because one-third of normal-sized spleens will be involved with Hodgkin's disease and one-third of enlarged spleens will be tumor-free. Although splenectomy in chronic myelogenous leukemia does not affect the natural history of disease, removal of the massive spleen usually makes patients significantly more comfortable and simplifies their management by significantly reducing transfusion requirements. Splenectomy is an effective secondary or tertiary treatment for two chronic B cell leukemias, hairy cell leukemia and prolymphocytic leukemia, and for the very rare splenic mantle cell or marginal zone lymphoma. Splenectomy in these diseases may be associated with significant tumor regression in bone marrow and other sites of disease. Similar regressions of systemic disease have been noted after splenic irradiation in some types of lymphoproliferative disease, especially chronic lymphocytic leukemia and prolymphocytic leukemia. This has been termed the *abscopal effect*. Such systemic tumor responses to local therapy directed at the spleen suggest that there may be some hormone or growth factor produced by the spleen that affects tumor cell proliferation, but this conjecture is not yet substantiated. The most common indication for splenectomy is traumatic or iatrogenic splenic rupture. In a fraction of patients with splenic rupture, peritoneal seeding of splenic fragments can lead to splenosis -- the presence of multiple rests of spleen tissue not connected to the portal circulation. This ectopic spleen tissue may cause pain or gastrointestinal obstruction, as in endometriosis. A large number of hematologic, immunologic, and congestive causes of splenomegaly can lead to destruction of one or more cellular blood elements. In the majority of such cases, splenectomy can correct the cytopenias, particularly anemia and thrombocytopenia. Perhaps the only contraindication to splenectomy is the presence of marrow failure, in which the enlarged spleen is the only source of hematopoietic tissue.

The absence of the spleen has minimal long-term effects on the hematologic profile. In the immediate postsplenectomy period, there may be some leukocytosis (up to 25,000/uL) and thrombocytosis (up to 1 ´10$_6$/uL), but within 2 to 3 weeks, blood cell counts and survival of each cell lineage are usually normal. The chronic manifestations of splenectomy are marked variation in size and shape of erythrocytes (anisocytosis, poikilocytosis) and the presence of Howell-Jolly bodies (nuclear remnants), Heinz bodies (denatured hemoglobin), basophilic stippling, and an occasional nucleated erythrocyte in the peripheral blood. When such erythrocyte abnormalities appear in a patient whose spleen has not been removed, one should suspect splenic infiltration by tumor that has interfered with its normal culling and pitting function.

The most serious consequence of splenectomy is increased susceptibility to bacterial infections, particularly those with capsules such as *Streptococcus pneumoniae*, *Haemophilus influenzae*, and some gram-negative enteric organisms. Patients under age 20 years are particularly susceptible to overwhelming sepsis with *S. pneumoniae*, and the overall actuarial risk of sepsis in patients who have had their spleens removed is about 7% in 10 years. The case-fatality rate for pneumococcal sepsis in splenectomized patients is 50 to 80%. About 25% of patients without spleens will develop a serious infection at some time in their life. The frequency is highest within the first 3 years after splenectomy. About 15% of the infections are polymicrobial, and lung, skin, and blood are the most common sites. No increased risk of viral infection has been noted in patients who have no spleen. The susceptibility to bacterial infections relates to the inability to remove opsonized bacteria from the bloodstream and a defect in making antibodies to T cell-independent antigens such as the polysaccharide components of bacterial capsules. Pneumococcal vaccine (23-valent polysaccharide vaccine) should be administered to all patients 2 weeks before elective splenectomy. The Advisory Committee on Immunization Practices recommends that even splenectomized patients receive pneumococcal vaccine with a repeat vaccination 5 years later. Efficacy has not been proven in this setting, and the recommendation discounts the possibility that administration of the vaccine may actually lower the titer of specific pneumococcal antibodies. A more effective pneumococcal vaccine that involves T cells in the response is in development. The vaccine to *H. influenzae* should also be given to patients in whom elective splenectomy is planned. No other vaccines are routinely recommended in this setting.

Splenectomized patients should be educated to consider any unexplained fever as a medical emergency. Prompt medical attention with evaluation and treatment of suspected bacteremia may be life-saving. Routine chemoprophylaxis with oral penicillin can result in the emergence of drug-resistant strains and is not recommended.

In addition to an increased susceptibility to bacterial infections, splenectomized patients are also more susceptible to the parasitic disease babesiosis. The splenectomized patient should avoid areas where the parasite *Babesia* is endemic (e.g., Cape Cod, MA).

Surgical removal of the spleen is an obvious cause of *hyposplenism*. Patients with sickle cell disease often suffer from autosplenectomy as a result of splenic destruction by the numerous infarcts associated with sickle cell crises during childhood. Indeed, the presence of a palpable spleen in a patient with sickle cell disease after age 5 suggests a coexisting hemoglobinopathy, e.g., thalassemia or hemoglobin C. In addition, patients who receive splenic irradiation for a neoplastic or autoimmune disease are also functionally hyposplenic. The term *hyposplenism* is preferred to *asplenism* in referring to the physiologic consequences of splenectomy because asplenia is a rare, specific, and fatal congenital abnormality in which there is a failure of the left side of the coelomic cavity (which includes the splenic anlagen) to develop normally. Infants with asplenia have no spleens, but that is the least of their problems. The right side of the developing embryo is duplicated on the left so there is liver where the spleen should be, there are two right lungs, and the heart comprises two right atria and two right ventricles.

(Bibliography omitted in Palm version)

Back to Table of Contents

64. DISORDERS OF GRANULOCYTES AND MONOCYTES - ***Steven M. Holland*, *John I. Gallin***

Leukocytes are the major cells comprising inflammatory and immune responses and include neutrophils, T and B lymphocytes, natural killer (NK) cells, monocytes, eosinophils, and basophils. These cells have specific functions, such as antibody production by B lymphocytes or destruction of bacteria by neutrophils, but in no single infectious disease is the exact role of the cell types completely established. Thus, whereas neutrophils are classically thought to be critical to host defense against bacteria, they may also play important roles in defense against viral infections.

The blood delivers leukocytes to the various tissues from the bone marrow, where they are produced. Normal blood leukocyte counts are given in the Appendix (Tables A-7 and A-8). The various leukocytes are derived from a common stem cell in the bone marrow. Three-fourths of the nucleated cells of bone marrow are committed to the production of leukocytes. Leukocyte maturation in the marrow is under the regulatory control of a number of different factors, known as colony stimulating factors and interleukins (Chap. 104). Because an alteration in the number and type of leukocytes is often associated with disease processes, total white blood count (WBC) (cells per microliter) and differential counts are informative. The lymphocytes and basophils are discussed inChaps. 305 and310, respectively. This chapter focuses on the neutrophils, monocytes, and eosinophils.

NEUTROPHILS

MATURATION

Important events in neutrophil life are summarized inFig. 64-1. In normal humans, neutrophils are produced only in the bone marrow. The minimum number of stem cells necessary to support hematopoiesis is estimated to be 400 to 500. Human blood monocytes, tissue macrophages, and stromal cells produce colony stimulating factors, hormones required for the growth of monocytes and neutrophils in the bone marrow. The hematopoietic system not only produces enough neutrophils (~1.3´10^{11}cells per 80-kg person per day) to carry out physiologic functions but also has a large reserve stored in the marrow which can be mobilized in response to inflammation or infection. An increase in the number of blood neutrophils is called neutrophilia, and the presence of immature cells is termed a shift to the left. A decrease in the number of blood neutrophils is called neutropenia.

Neutrophils and monocytes evolve from pluripotent stem cells under the influence of cytokines and colony stimulating factors (Fig. 64-2). The proliferation phase through the metamyelocyte takes about 1 week, while the maturation phase from metamyelocyte to mature neutrophil takes another week. The myeloblast is the first recognizable precursor cell and is followed by the *promyelocyte* (Plate V-23). The promyelocyte evolves when the classic lysosomal granules, called the *primary* or *azurophil granules*, are produced. The primary granules contain hydrolases, elastase, myeloperoxidase, cationic proteins, and bactericidal/permeability-increasing protein, which is important for killing gram-negative bacteria. Azurophil granules also contain *defensins*, a family of cysteine-rich polypeptides with broad antimicrobial activity against bacteria, fungi, and certain enveloped viruses. The promyelocyte divides to produce the *myelocyte*, a cell responsible for the synthesis of the *specific* or *secondary granules* which contain unique (specific) constituents such as lactoferrin, vitamin B_{12}-binding proteins, membrane components of the nicotinamide-adenine dinucleotide phosphate (NADPH) oxidase required for hydrogen peroxide production, histaminase, and receptors for certain chemoattractants and adherence-promoting factors (CR3) as well as receptors for the basement membrane component, laminin. The secondary granules do not contain acid hydrolases and therefore are not classic lysosomes. Packaging of secondary granule contents during myelopoiesis is controlled by CCAAT/enhancer binding protein-e. Secondary granule contents are readily released extracellularly, and their mobilization is important in modulating inflammation. During the final stages of maturation no cell division occurs, and the cell passes through the *metamyelocyte* stage and then to the *band* neutrophil with a sausage-shaped nucleus (Plate V-35). As the band cell matures, the nucleus assumes a lobulated configuration. The nucleus of neutrophils normally contains up to four segments. Excessive segmentation (more than five nuclear lobes) may be a manifestation of folate or vitamin B_{12}deficiency (Plate V-38). The Pelger-Huet anomaly (Plate V-34B), an infrequent dominant benign inherited trait, results in neutrophils with distinctive bilobed nuclei that must be distinguished from band forms. The physiologic role of the multilobed nucleus of neutrophils is unknown, but it may allow great deformation of neutrophils during migration into tissues at sites of inflammation.

In severe acute bacterial infection, prominent neutrophil cytoplasmic granules called *toxic granulations* are occasionally seen (Plate V-11). Toxic granulations are immature or abnormally staining azurophil granules. Cytoplasmic inclusions, also called *Dohle bodies* (Plate V-35), can be seen during infection and are fragments of ribosome-rich endoplasmic reticulum. Large neutrophil vacuoles are often present in acute bacterial infection and probably represent pinocytosed (internalized) membrane.

Neutrophils are heterogeneous in function. Monoclonal antibodies have been developed that recognize only a subset of mature neutrophils. The meaning of neutrophil heterogeneity is not known.

MARROW RELEASE AND CIRCULATING COMPARTMENTS

Specific signals, including interleukin (IL) 1, tumor necrosis factor-a (TNF-a), the colony stimulating factors, complement fragment C3e, and perhaps other cytokines mobilize leukocytes from the bone marrow and deliver them to the blood in an unstimulated state. Under normal conditions, about 90% of the neutrophil pool is in the bone marrow, 2 to 3% in the circulation, and the remainder in the tissues (Fig. 64-3).

The circulating pool exists in two dynamic compartments: one freely flowing and one marginated. The freely flowing pool is about one-half the neutrophils in the basal state and is composed of those cells that are in the blood and not in contact with the endothelium. Marginated leukocytes are those that are in close physical contact with the endothelium (Fig. 64-4). In the pulmonary circulation, where an extensive capillary bed (~1000 capillaries per alveolus) exists, margination occurs because the capillaries are about the same size as a mature neutrophil. Therefore, neutrophil fluidity and deformability are necessary to make the transit through the pulmonary bed. Increased neutrophil rigidity and decreased deformability lead to augmented neutrophil trapping and margination in the lung. In contrast, in the systemic postcapillary venules, margination is mediated by the interaction of specific cell-surface molecules. *Selectins* are glycoproteins expressed on neutrophils and endothelial cells, among others, that cause a low-affinity interaction, resulting in "rolling" of the neutrophil along the endothelial surface. On neutrophils, the molecule L-selectin [cluster determinant (CD) 62L] binds to glycosylated proteins on endothelial cells [e.g., glycosylation-dependent cell adhesion molecule (GlyCAM1) and CD34]. Glycoproteins on neutrophils, most importantly sialyl-Lewis$_x$(SLe$_x$, CD15s), are targets for binding of selectins expressed on endothelial cells [E-selectin (CD62E) and P-selectin (CD62P)] and other leukocytes. In response to chemotactic stimuli from injured tissues (e.g., complement product C5a, leukotriene B$_4$,IL-8) or bacterial products [e.g., *N*-formylmethionylleucylphenylalanine (f-metleuphe)], neutrophil adhesiveness increases, and the cells "stick" to the endothelium through *integrins*. The integrins are leukocyte glycoproteins that exist as complexes of a common CD18 b-chain with CD11a (LFA-1), CD11b (also called either Mac-1, CR3, or the C3bi receptor), and CD11c (p150,95). CD11a/CD18 and CD11b/CD18 bind to specific endothelial receptors [intercellular adhesion molecules (ICAM) 1 and 2].

On cell stimulation, L-selectin is shed; receptors for chemoattractants and opsonins are mobilized; the phagocytes orient toward the chemoattractant source in the extravascular space, increase their motile activity (chemokinesis), and migrate directionally (chemotaxis) into tissues. The process of migration into tissues is called *diapedesis* and involves the crawling of neutrophils between postcapillary endothelial cells that open junctions between adjacent cells to permit leukocyte passage. Diapedesis involves platelet/endothelial cell adhesion molecule (PECAM) 1 (CD31), which is expressed on both the emigrating leukocyte and the endothelial cells. The endothelial responses (increased blood flow from increased vasodilation and permeability) are mediated by anaphylatoxins (e.g., C3a and C5a) as well as vasodilators such as histamine, bradykinin, serotonin, nitric oxide, vascular endothelial growth factor (VEGF), and prostaglandins E and I. Cytokines regulate some of these processes [e.g.,TNF-ainduction of VEGF, interferon (IFN) g inhibition of prostaglandin E].

In the healthy adult, most neutrophils leave the body by migration through the mucous membrane of the gastrointestinal tract. Normally, neutrophils spend a short time in the circulation (half-life, 6 to 7 h). Senescent neutrophils are cleared from the circulation by macrophages in the lung and spleen. Once in the tissues, neutrophils release enzymes, such as collagenase and elastase, that help establish abscess cavities. Neutrophils ingest pathogenic materials that have been opsonized by IgG and C3b. Fibronectin and the tetrapeptide tuftsin facilitate phagocytosis.

With phagocytosis comes a burst of oxygen consumption and activation of the hexose-monophosphate shunt. A membrane-associatedNADPHoxidase, consisting of membrane and cytosolic components, is assembled and catalyzes the reduction of oxygen to superoxide anion, which is then converted to hydrogen peroxide and other toxic oxygen products (e.g., hydroxyl radical). Hydrogen peroxide + chloride+ neutrophil myeloperoxidase generates hypochlorous acid (bleach), hypochlorite, and chlorine. These products oxidize and halogenate microorganisms and tumor cells and, when uncontrolled, can damage host tissue. Strongly cationic proteins, defensins, and probably nitric oxide also participate in microbial killing. Other enzymes, such as lysozyme and acid proteases, help digest microbial debris. After 1 to 4 days in tissues neutrophils die. The apoptosis of neutrophils is also cytokine regulated; granulocyte colony stimulating factor (G-CSF) andIFN-gprevent their death. Under certain conditions, such as in delayed-type hypersensitivity, monocyte accumulation occurs within 6 to 12 h of initiation of inflammation. Neutrophils, monocytes, microorganisms in various states of digestion, and altered local tissue cells make up the inflammatory exudate, pus. Myeloperoxidase confers the characteristic green color to pus and may participate in turning off the inflammatory process by inactivating chemoattractants and immobilizing phagocytic cells.

Neutrophils respond to certain cytokines [IFN-g, granulocyte-macrophage colony stimulating factor (GM-CSF),IL-8] and produce cytokines and chemotactic signals [TNF-a, IL-8, macrophage inflammatory protein (MIP) 1] that modulate the inflammatory response. In the presence of fibrinogen, f-metleuphe or leukotriene B_4induces IL-8 production by neutrophils, providing autocrine amplification of inflammation. *Chemokines* (*chemo*attractant cyto*kines*) are small proteins produced by many different cell types, including endothelial cells, fibroblasts, epithelial cells, neutrophils, and monocytes, that regulate neutrophil and monocyte recruitment and activation. The chemokines transduce their signals through heterotrimeric G protein-linked receptors that have seven cell membrane-spanning domains, the same type of cell-surface receptor that mediates the response to the classical chemoattractants *N*-f-metleuphe and C5a. Four major groups of chemokines are recognized based on the cysteine structure near the N terminus: C, CC, CXC, and CXXXC. The CXC cytokines like IL-8 mainly attract neutrophils; CC chemokines likeMIP-1aattract lymphocytes, monocytes, eosinophils, and basophils; the C chemokine lymphotactin is T cell tropic; the CXXXC chemokine fractalkine attracts neutrophils, monocytes, and T cells. These molecules and their receptors not only regulate the trafficking and activation of inflammatory cells, but chemokine receptors serve as co-receptors for HIV infection (Chap. 309).

NEUTROPHIL ABNORMALITIES

A defect in the neutrophil life cycle can lead to dysfunction and compromised host defenses. Inflammation is often depressed, and the clinical result is often recurrent and severe bacterial and fungal infections. Aphthous ulcers of mucous membranes (gray ulcers without pus) and gingivitis and periodontal disease suggest a phagocytic cell disorder. Patients with congenital phagocyte defects can have infections within the first few days of life. Skin, ear, upper and lower respiratory tract, and bone infections are common. Sepsis and meningitis are rare. In some disorders the frequency of infection is variable, and patients can go for months or even years without major infection. Aggressive management of these congenital diseases has extended the life span of patients beyond 30 years.

Neutropenia The consequences of absent neutrophils are dramatic. Susceptibility to infectious diseases increases sharply when neutrophil counts fall below 1000 cells/uL. When the absolute neutrophil count (ANC; band forms and mature neutrophils combined) falls below 500 cells/uL, control of endogenous microbial flora (e.g., mouth, gut) is impaired; when the ANC is < 200/uL, the inflammatory process is absent. Neutropenia can be due to depressed production, increased peripheral destruction, or excessive peripheral pooling. A falling neutrophil count or a significant decrease in the number of neutrophils below steady state levels, together with a failure to increase neutrophil counts in the setting of infection or other challenge, requires investigation. Acute neutropenia, such as that caused by cancer chemotherapy, is more likely to be associated with increased risk of infection than neutropenia of long duration (months to years) that reverses in response to infection or carefully controlled administration of endotoxin (see "Laboratory Diagnosis," below).

Some causes of inherited and acquired neutropenia are listed in Table 64-1. The most common neutropenias are iatrogenic, resulting from the use of cytotoxic or immunosuppressive therapies for malignancy or control of autoimmune disorders. These drugs cause neutropenia because they result in decreased production of rapidly growing progenitor (stem) cells of the marrow. Certain antibiotics such as chloramphenicol, trimethoprim-sulfamethoxazole, flucytosine, vidarabine, and the antiretroviral drug zidovudine may cause neutropenia by inhibiting proliferation of myeloid precursors. The marrow suppression is generally dose-related and dependent on continued administration of the drug. Recombinant humanG-CSFreverses this form of neutropenia.

Another important mechanism for iatrogenic neutropenia is the effect of drugs that serve as immune haptens and sensitize neutrophils or neutrophil precursors to immune-mediated peripheral destruction. This form of drug-induced neutropenia can be seen within 7 days of exposure to the drug; with previous drug exposure, resulting in preexisting antibodies, neutropenia may occur a few hours after administration of the drug. Although any drug can cause this form of neutropenia, the most frequent causes are commonly used antibiotics, such as sulfa-containing compounds, penicillins, and cephalosporins. Fever and eosinophilia also may be associated drug reactions, but often these signs are not present. Drug-induced neutropenia can be severe, but discontinuation of the sensitizing drug is sufficient for recovery, which is usually seen within 5 to 7 days and is complete by 10 days. Readministration of the sensitizing drug should be avoided, since abrupt neutropenia often will result. For this reason, diagnostic challenge should be avoided.

Autoimmune neutropenias caused by circulating antineutrophil antibodies are another form of acquired neutropenia that results in increased destruction of neutrophils. Acquired neutropenia also may be seen with viral infections, including infection with HIV. Acquired neutropenia may be cyclic in nature, occurring at intervals of several weeks. Acquired cyclic or stable neutropenia may be associated with an expansion of large granular lymphocytes (LGL), which may be T cells,NKcells, or NK-like cells. Patients with LGL lymphocytosis may have moderate blood and bone marrow lymphocytosis, neutropenia, polyclonal hypergammaglobulinemia, splenomegaly, rheumatoid arthritis, and absence of lymphadenopathy. Such patients may have a chronic and relatively stable course. Recurrent bacterial infections are frequent. Benign and malignant forms of this syndrome occur. In some patients, a spontaneous regression has occurred even after 11 years, suggesting an immunoregulatory defect as the basis for at least one form of the disorder. Glucocorticoids, cyclosporine,IFN-a, and nucleosides such as 2-chlorodeoxyadenosine each have induced remission.

Hereditary Neutropenias Hereditary neutropenias are rare and may manifest in early childhood as a profound constant neutropenia or agranulocytosis. Congenital forms of neutropenia include Kostmann's syndrome (neutrophil count<100/uL), which is often fatal; more benign chronic idiopathic neutropenia (neutrophil count of 300 to 1500/uL); the cartilage-hair hypoplasia syndrome; Shwachman's syndrome associated with pancreatic insufficiency; myelokathexis, a congenital disorder characterized by neutrophil degeneration, hypersegmentation, and myeloid hyperplasia in the marrow associated with decreased expression of bcl-$X_L$in myeloid precursors and accelerated apoptosis; and neutropenias associated with other immune defects (X-linked agammaglobulinemia, ataxia telangiectasia, IgA deficiency). Mutations in theG-CSFreceptor on chromosome 1 associated with poor response to G-CSF can occur with severe congenital neutropenia and predispose to myeloid malignancy. Hereditary cyclic neutropenia, an autosomal dominant trait, may occur in infancy and is characterized by a remarkably regular 3-week cycle. Hereditary cyclic neutropenia actually is cyclic hematopoiesis, due to mutations in the neutrophil elastase gene. Glucocorticoids and G-CSF blunt the cycling in some patients.

Maternal factors can be associated with neutropenia in the newborn. Transplacental transfer of IgG directed against antigens on fetal neutrophils can result in peripheral destruction. Drugs (e.g., thiazides) ingested during pregnancy can cause neutropenia in the newborn by either depressed production or peripheral destruction.

The presence of immunoglobulin directed toward neutrophils is seen in Felty's syndrome -- a triad of rheumatoid arthritis, splenomegaly, and neutropenia (Chap. 312). Patients with Felty's syndrome who respond to splenectomy with an increase in their neutrophil count also have lower postoperative serum neutrophil-binding IgG. Some of these patients have neutropenia associated with an increased number ofLGL. Splenomegaly with peripheral trapping and destruction of neutrophils is also seen in lysosomal storage diseases and in portal hypertension.

Neutrophilia Neutrophilia results from increased neutrophil production, increased marrow release, or defective margination (Table 64-2). The most important acute cause of neutrophilia is infection. Neutrophilia from acute infection represents both increased production and increased marrow release. Increased production is also associated with chronic inflammation and certain myeloproliferative diseases. Increased marrow release and mobilization of the marginated leukocyte pool are induced by glucocorticoids. Release of epinephrine, as with vigorous exercise, excitement, or stress, will demarginate neutrophils in the spleen and lungs and double the neutrophil count in minutes. Leukocytosis with cell counts of 10,000 to 25,000/uL occurs in response to infection and other forms of acute inflammation and results from both release of the marginated pool and mobilization of marrow reserves. Persistent neutrophilia with cell counts of 30,000 to 50,000/uL or higher is called a *leukemoid reaction*, a term often used to distinguish this degree of neutrophilia from leukemia. In a leukemoid reaction, the circulating neutrophils are usually mature and not clonally derived.

Abnormal Neutrophil Function Inherited and acquired abnormalities of phagocyte function are listed in Table 64-3. The resulting diseases are best considered in terms of the functional defects of adherence, chemotaxis, and microbicidal activity. The distinguishing features of the important inherited disorders of phagocyte function are shown in Table 64-4.

Disorders of Adhesion Two types of leukocyte adhesion deficiency (LAD) have been described. Both are autosomal recessive traits and result in the inability of neutrophils to exit the circulation to sites of infection, leading to leukocytosis and increased susceptibility to infection (Fig. 64-4). Patients with LAD 1 have mutations in CD18, the common component of the integrins LFA-1, Mac-1, and p150,95, leading to a defect in tight adhesion between neutrophils and the endothelium. The heterodimer formed by CD18/CD11b (Mac-1) is also the receptor for the complement-derived opsonin C3bi (CR3). The CD18 gene is located on distal chromosome 21q. Variable expression of the defect determines the severity of clinical disease. Complete lack of expression of the leukocyte adhesion proteins results in the severe phenotype in which inflammatory cytokines do not increase the expression of leukocyte adhesion proteins on neutrophils or activated T and B cells. Neutrophils (and monocytes) from patients with LAD 1 adhere poorly to endothelial cells and protein-coated surfaces and exhibit defective spreading, aggregation, and chemotaxis. Patients with LAD 1 have recurrent bacterial and fungal infections involving skin, oral and genital mucosa, and respiratory and intestinal tracts; persistent leukocytosis (neutrophil counts of 15,000 to 20,000/uL) because cells do not marginate; and, in severe cases, a history of delayed separation of the umbilical stump. Infections, especially of the skin, may become necrotic with progressively enlarging borders, slow healing, and development of dysplastic scars. The most common bacteria are *Staphylococcus aureus* and enteric gram-negative bacteria. LAD 2 is caused by an abnormality of SLe$_x$(CD15s), the ligand on neutrophils that interacts with selectins on endothelial cells.

Disorders of Neutrophil Granules The most common neutrophil defect is *myeloperoxidase deficiency*, a primary granule defect inherited as an autosomal recessive trait; the incidence is ~1 in 2000 persons. Isolated myeloperoxidase deficiency is not associated with clinically compromised defenses, because other defense systems such as hydrogen peroxide generation are amplified. Microbicidal activity of neutrophils is delayed but not absent. Myeloperoxidase deficiency may make other acquired host defense defects more serious. An acquired form of myeloperoxidase deficiency occurs in myelomonocytic leukemia and acute myeloid leukemia.

Chediak-Higashi syndrome (CHS) is a rare disease with autosomal recessive inheritance due to defects in the lysosomal transport protein LYST, encoded by the gene *CHS1* at 1q42. This protein is required for normal packaging and disbursement of granules. Neutrophils (and all cells containing lysosomes) from patients with CHS characteristically have large granules (Plate V-34A). Patients with CHS have an increased number of infections resulting from many agents. CHS neutrophils and monocytes have impaired chemotaxis and abnormal rates of microbial killing due to slow rates of fusion of the lysosomal granules with phagosomes.NK cell function is also impaired.

Specific granule deficiency is a rare autosomal recessive disease in which the production of secondary granules and their contents, as well as primary granule component defensins, is defective. The defect in bacterial killing leads to severe bacterial infections. One type of specific granule deficiency is due to a mutation in the CCAAT/enhancer binding protein-e, a regulator of expression of granule components.

Chronic granulomatous disease Chronic granulomatous disease (CGD) is a group of disorders of granulocyte and monocyte oxidative metabolism. Although CGD is rare, with an incidence of 1 in 200,000 individuals, it is an important model of defective neutrophil oxidative metabolism. Most often CGD is inherited as an X-linked recessive trait; 30% of patients inherit the disease in an autosomal recessive pattern. Mutations in the genes for the four proteins that assemble at the plasma membrane account for all patients with CGD. Two proteins (a 91-kDa protein, abnormal in X-linked CGD, and a 22-kDa protein, absent in one form of autosomal recessive CGD) form the heterodimer cytochrome b-558 in the plasma membrane. Two other proteins (47 and 67 kDa, abnormal in the other autosomal recessive forms of CGD) are cytoplasmic in origin and interact with the cytochrome after cell activation to formNADPHoxidase, required for hydrogen peroxide production. Leukocytes from patients with CGD have severely diminished hydrogen peroxide production. The genes involved in each of the defects have been cloned and sequenced and the chromosome locations identified. Patients with CGD characteristically have increased numbers of infections due to catalase-positive microorganisms (organisms that destroy their own hydrogen peroxide). When patients with CGD become infected, they often have extensive inflammatory reactions, and lymph node suppuration is common despite the administration of appropriate antibiotics. Aphthous ulcers and chronic inflammation of the nares are often present. Granulomas are frequent and can obstruct the gastrointestinal or genitourinary tracts. The excessive inflammation probably reflects failure to degrade chemoattractants and antigens, leading to persistent neutrophil accumulation. Impaired killing of intracellular microorganisms by macrophages may lead to persistent cell-mediated immunity and granuloma formation.

MONONUCLEAR PHAGOCYTES

The mononuclear phagocyte system is composed of monoblasts, promonocytes, and monocytes in addition to the structurally diverse tissue macrophages that make up what was previously referred to as the reticuloendothelial system. Macrophages are long-lived phagocytic cells capable of many of the functions of neutrophils. They are also secretory cells that participate in many immunologic and inflammatory processes distinct from neutrophils. Monocytes leave the circulation by diapedesis more slowly than neutrophils and have a half-life in the blood of 12 to 24 h.

After blood monocytes arrive in the tissues, they differentiate into macrophages ("big eaters") with specialized functions suited for specific anatomic locations. Macrophages are particularly abundant in capillary walls of the lung, spleen, liver, and bone marrow, where they function to remove microorganisms and other noxious elements from the blood. Alveolar macrophages, liver Kupffer cells, splenic macrophages, peritoneal macrophages, bone marrow macrophages, lymphatic macrophages, brain microglial cells, and dendritic macrophages all have specialized functions. Macrophage-secreted products include lysozyme, neutral proteases, acid hydrolases, arginase, complement components, enzyme inhibitors (plasmin,a$_2$-macroglobulin), binding proteins (transferrin, fibronectin, transcobalamin II), nucleosides, and cytokines (TNF-a;IL-1, -8, -12, and -18). IL-1 (Chaps. 17 and305) has many functions, including initiating fever in the hypothalamus, mobilizing leukocytes from the bone marrow, activating lymphocytes and neutrophils. TNF-a is a pyrogen that duplicates many of the actions of IL-1 and plays an important role in the pathogenesis of gram-negative shock (Chap. 124). TNF-a stimulates production of hydrogen peroxide and related toxic oxygen species by macrophages and neutrophils. In addition, TNF-a induces catabolic changes that contribute to the profound wasting (cachexia) associated with many chronic diseases.

Other macrophage-secreted products include reactive oxygen and nitrogen metabolites, bioactive lipids (arachidonic acid metabolites and platelet-activating factors), chemokines, colony stimulating factors, and factors stimulating fibroblast and vessel proliferation. Macrophages help regulate the replication of lymphocytes and participate in the killing of tumors, viruses, and certain bacteria (*Mycobacterium tuberculosis* and *Listeria monocytogenes*). Macrophages are key effector cells in the elimination of intracellular microorganisms. Their ability to fuse to form giant cells that coalesce into granulomas in response to some inflammatory stimuli is important in the elimination of intracellular microbes and is under the control ofIFN-g. Nitric oxide induced by IFN-g is an important effector against intracellular parasites including tuberculosis and *Leishmania*.

Macrophages play an important role in the immune response (Chap. 305). They process and present antigen to lymphocytes and secrete cytokines that modulate and direct lymphocyte development and function. Macrophages participate in autoimmune phenomena by removing immune complexes and other substances from the circulation. Polymorphisms in macrophage receptors for immunoglobulin (FcgRII) determine suceptibility to some infections and autoimmune diseases. In wound healing, they dispose of senescent cells, and they contribute to atheroma development. Macrophage elastase mediates development of emphysema from cigarette smoking.

DISORDERS OF THE MONONUCLEAR PHAGOCYTE SYSTEM

Many disorders of neutrophils extend to mononuclear phagocytes. Thus, drugs that suppress neutrophil production in the bone marrow can cause monocytopenia. Transient monocytopenia occurs after stress or glucocorticoid administration. Monocytosis is associated with tuberculosis, brucellosis, subacute bacterial endocarditis, Rocky Mountain spotted fever, malaria, and visceral leishmaniasis (kala azar). Monocytosis also occurs with malignancies, leukemias, myeloproliferative syndromes, hemolytic anemias, chronic idiopathic neutropenias, and granulomatous diseases such as sarcoidosis, regional enteritis, and some collagen vascular diseases. Patients withLAD, hyperimmunoglobulin E-recurrent infection (Job's) syndrome,CHS, andCGD all have defects in the mononuclear phagocyte system.

Monocyte cytokine production is impaired in some patients with disseminated nontuberculous mycobacterial infection who are not infected with HIV. Genetic defects inIFN-greceptors 1 and 2 impair monocyte killing of intracellular parasites, as do lesions in the potent IFN-g inducer,IL-12 and its receptor (Fig. 64-5).

Certain viral infections impair mononuclear phagocyte function. For example, influenza virus infection causes abnormal monocyte chemotaxis. Mononuclear phagocytes can be infected by HIV using CCR5, the chemokine receptor that acts as a coreceptor with CD4 for HIV. T lymphocytes produceIFN-g, which induces FcR expression and phagocytosis and stimulates hydrogen peroxide production by mononuclear phagocytes and neutrophils. In certain diseases, such as AIDS, IFN-g production may be deficient, while in other diseases, such as T cell lymphomas, excessive release of IFN-g may be associated with erythrophagocytosis by splenic macrophages.

Monocytopenia occurs with acute infections, with stress, and after treatment with glucocorticoids. Monocytopenia also occurs in aplastic anemia, hairy cell leukemia, acute myeloid leukemia, and as a direct result of myelotoxic drugs.

EOSINOPHILS

Eosinophils and neutrophils share similar morphology, many lysosomal constituents, phagocytic capacity, and oxidative metabolism. Eosinophils express a specific chemoattractant receptor and respond to a specific chemokine, eotaxin. Little is known about the role of eosinophils. Eosinophils are much longer lived than neutrophils, and unlike neutrophils, tissue eosinophils can recirculate. During most infections, eosinophils are not important. However, in invasive helminthic infections, such as hookworm, schistosomiasis, strongyloidiasis, toxocariasis, trichinosis, filariasis, echinococcosis, and cysticercosis, the eosinophil plays a central role in host defense. Eosinophils are associated with bronchial asthma, cutaneous allergic reactions, and other hypersensitivity states.

The distinctive feature of the red-staining (Wright's stain) eosinophil granules is its crystalline core consisting of an arginine-rich protein (major basic protein) with histaminase activity, important in host defense against parasites. Eosinophil granules also contain a unique eosinophil peroxidase that catalyzes the oxidation of many substances by hydrogen peroxide and may facilitate killing of microorganisms.

Eosinophil peroxidase, in the presence of hydrogen peroxide and halide, initiates mast cell secretion in vitro and thereby promotes inflammation. Eosinophils contain cationic proteins, some of which bind to heparin and reduce its anticoagulant activity. Eosinophil-derived neurotoxin and eosinophil cationic protein are ribonucleases that can kill respiratory syncytial virus. Eosinophil cytoplasm contains Charcot-Leyden crystal protein, a hexagonal bipyramidal crystal first observed in a patient with leukemia and then in sputum of patients with asthma; this protein is lysophospholipase and may function to detoxify certain lysophospholipids.

Several factors enhance the eosinophil's function in host defense. T cell-derived factors enhance the ability of eosinophils to kill parasites. Mast cell-derived eosinophil chemotactic factor of anaphylaxis (ECFa) increases the number of eosinophil complement receptors and enhances eosinophil killing of parasites. Eosinophil colony stimulating factors (e.g.,IL-5) produced by macrophages increase eosinophil production in the bone marrow and activate eosinophils to kill parasites.

EOSINOPHILIA

Eosinophilia is the presence of >500 eosinophils per microliter of blood and is common in many settings besides parasite infection. Significant tissue eosinophilia can occur without an elevated blood count. The most common cause of eosinophilia is allergic reactions to drugs (iodides, aspirin, sulfonamides, nitrofurantoin, penicillins, and cephalosporins). Allergies such as hay fever, asthma, eczema, serum sickness, allergic vasculitis, and pemphigus are associated with eosinophilia. Eosinophilia also occurs in collagen vascular diseases (e.g., rheumatoid arthritis, eosinophilic fasciitis, allergic angiitis, and periarteritis nodosa) and malignancies (e.g., Hodgkin's disease; mycosis fungoides; chronic myelogenous leukemia; and cancer of the lung, stomach, pancreas, ovary, or uterus), as well as in Job's syndrome andCGD. Eosinophilia commonly is present in the helminthic infections.IL-5 is the dominant eosinophil growth factor. Therapeutic administration of the cytokines IL-2 andGM-CSFfrequently leads to transient eosinophilia. The most dramatic hypereosinophilic syndromes are Loeffler's syndrome, tropical pulmonary eosinophilia, Loeffler's endocarditis, eosinophilic leukemia, and idiopathic hypereosinophilic syndrome (50,000 to 100,000/uL).

The idiopathic hypereosinophilic syndrome represents a heterogeneous group of disorders with the common feature of prolonged eosinophilia of unknown cause and organ system dysfunction, including the heart, central nervous system, kidneys, lungs, gastrointestinal tract, and skin. The bone marrow is involved in all affected individuals, but the most severe complications involve the heart and central nervous system. Clinical manifestations and organ dysfunction are highly variable. Eosinophils are found in the involved tissues and likely cause tissue damage by local deposition of toxic eosinophil proteins such as eosinophil cationic protein and major basic protein. In the heart, the pathologic changes lead to thrombosis, endocardial fibrosis, and restrictive endomyocardiopathy. The damage to tissues in other organ systems is similar. The mechanism for the hypereosinophilia is not known. Glucocorticoids usually induce remission. In patients who do not respond to glucocorticoids, a cytotoxic agent such as hydroxyurea has been used successfully to lower the peripheral blood eosinophil counts and to improve markedly the prognosis.IFN-aalso is effective in some patients, including those unresponsive to hydroxyurea. Aggressive medical and surgical approaches are used to manage patients with cardiovascular complications.

The *eosinophilia-myalgia syndrome* is a multisystem disease with prominent cutaneous, hematologic, and visceral manifestations that frequently evolves into a chronic course and can occasionally be fatal. The syndrome is characterized by eosinophilia (eosinophil count >1000/uL) and generalized disabling myalgias without other recognized causes. Eosinophil fasciitis, pneumonitis, and myocarditis; neuropathy culminating in respiratory failure; and encephalopathy may occur. The disease is caused by ingesting contaminants in L-tryptophan-containing products. Eosinophils, lymphocytes, macrophages, and fibroblasts accumulate in the affected tissues, but their role in pathogenesis is unclear. Activation of eosinophils and fibroblasts and the deposition of eosinophil-derived toxic proteins in affected tissues may contribute.IL-5 and transforming growth factor b have been implicated as potential mediators. Treatment is withdrawal of L-tryptophan-containing products and the administration of glucocorticoids. Most patients recover fully, remain stable, or show slow recovery; but the disease can be fatal in up to 5% of patients.

EOSINOPENIA

Eosinopenia occurs with stress, such as acute bacterial infection, and after treatment with glucocorticoids. The mechanism of eosinopenia of acute bacterial infection is unknown but is independent of endogenous glucocorticoids, since it occurs in animals after total adrenalectomy. There is no known adverse effect of eosinopenia.

HYPERIMMUNOGLOBULIN E-RECURRENT INFECTION SYNDROME

The hyperimmunoglobulin E-recurrent infection (HIE) syndrome or *Job's syndrome* is a rare multisystem disease in which the immune system, bone, teeth, lung and skin are affected. Abnormal chemotaxis is a variable feature. The molecular basis for this syndrome is not known, but some cases show autosomal dominant transmission with linkage to 4q. Patients with this syndrome have characteristic facies with broad nose, kyphoscoliosis and osteoporosis, and eczema. The primary teeth erupt normally but do not deciduate, often requiring extraction. Patients develop recurrent sinopulmonary and cutaneous infections that tend to be much less inflamed than appropriate for the degree of infection and have been referred to as "cold abscesses." A high degree of suspicion is required to diagnose infections in these patients, who may appear well despite extensive disease. The cold abscesses have been considered a reflection of impaired chemotaxis with too few phagocytes arriving too late, perhaps due to a lymphocyte factor inhibiting chemotaxis. However, the chemotactic defect in these patients is variable, and the fundamental basis for the impaired defenses is complex and poorly defined.

LABORATORY DIAGNOSIS AND MANAGEMENT

Initial studies ofWBC and differential and often a bone marrow examination are followed by assessment of bone marrow reserves (steroid challenge test), marginated circulating pool of cells (epinephrine challenge test), and marginating ability (endotoxin challenge test) (Fig. 64-3). In vivo assessment of inflammation is possible with a Rebuck skin window test or an in vivo blister assay, which measures the ability of leukocytes and inflammatory mediators to accumulate locally in the skin. In vitro tests of phagocyte aggregation, adherence, chemotaxis, phagocytosis, degranulation, and microbicidal activity (for *S. aureus*) may help pinpoint cellular or humoral lesions. Deficiencies of oxidative metabolism are detected with the nitroblue tetrazolium (NBT) dye test, which is based on the ability of products of oxidative metabolism to reduce yellow, soluble NBT to blue-black formazan, an insoluble material that can be seen microscopically. Studies of superoxide and hydrogen peroxide production may further define neutrophil oxidative function.

Patients with leukopenias or leukocyte dysfunction often have delayed inflammatory responses. Therefore, clinical manifestations may be minimal despite overwhelming infection, and unusual infections must always be suspected. Early signs of infection demand prompt, aggressive culturing for microorganisms, use of antibiotics, and surgical drainage of abscesses. Prolonged antibiotics are often required. In patients withCGD, prophylactic antibiotics (trimethoprim-sulfamethoxazole) diminish the frequency of life-threatening infections. Short courses of glucocorticoids may relieve gastrointestinal or genitourinary tract obstruction by granulomas in patients with CGD. Recombinant humanIFN-g, which nonspecifically stimulates phagocytic cell function, reduces the frequency of infections in patients with CGD by 70% and reduces the severity of infection. This effect of IFN-g in CGD is additive to the effect of prophylactic antibiotics. The recommended dose is 50 ug/m$_2$subcutaneously three times weekly. IFN-g also has been used successfully in the treatment of leprosy, nontuberculous mycobacteria, and visceral leishmaniasis.

Rigorous oral hygiene reduces but does not eliminate the discomfort of gingivitis, periodontal disease, and aphthous ulcers; chlorhexidine mouthwash and tooth brushing with a hydrogen peroxide-sodium bicarbonate paste helps many patients. Oral antifungal agents (fluconazole) have reduced mucocutaneous candidiasis in patients with Job's syndrome. Androgens, glucocorticoids, lithium, and immunosuppressive therapy have been used to restore myelopoiesis in patients with neutropenia due to impaired production. RecombinantG-CSF is useful in the management of certain forms of neutropenia due to depressed neutrophil production, especially that related to cancer chemotherapy. Patients with chronic neutropenia with evidence of a good bone marrow reserve need not receive prophylactic antibiotics.

Patients with constant or cyclic neutrophil counts <500/uL may benefit from prophylactic antibiotics andG-CSFduring periods of neutropenia. Oral trimethoprim-sulfamethoxazole (160/800 mg) twice daily can prevent infection. Increased numbers of fungal infections are not seen in patients withCGD on this regimen. Oral quinolones such as norfloxacin and ciprofloxacin are alternatives.

In the setting of cytotoxic chemotherapy with severe, persistent neutropenia, trimethoprim-sulfamethoxazole prevents *Pneumocystis carinii* pneumonia. These patients, and patients with phagocytic cell dysfunction, should avoid heavy exposure to airborne soil, dust, or decaying matter (mulch, manure), which are often rich in spores of *Aspergillus* or other fungi. Restriction of activities or social contact has no proven role in reducing risk of infection.

Cure of some congenital phagocyte defects is possible by bone marrow transplantation (Chap. 115). However, complications of bone marrow transplantation are still serious, and with rigorous medical care many patients with phagocytic disorders can go for years without a life-threatening infection. The identification of specific gene defects in patients withLAD 1 and CGDhas led to gene therapy trials in a number of genetic white cell disorders.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART THREE -GENETICS AND DISEASE

65. PRINCIPLES OF HUMAN GENETICS - *J. Larry Jameson*, *Peter Kopp*

IMPACT OF GENETICS ON MEDICAL PRACTICE

New insights into the genetic basis of disease are being generated at an ever-increasing rate. This explosion of information was ignited by technological advances, such as the polymerase chain reaction (PCR) and automated DNA sequencing, and is fueled by rapid progress in the Human Genome Project (HGP). Although its promise is great, the integration of genetics into the everyday practice of medicine remains challenging. To date, the most significant impact of genetics has been to enhance our understanding of disease etiology and pathogenesis. In the near term, we can expect an even greater role for genetics in the diagnosis, prevention, and treatment of disease (Chaps. 68 and69).

Genetic disorders are more common than generally appreciated. It is estimated, for example, that 3% of pregnancies result in a child with a genetic disease or birth defect. About 10% of all pediatric and adult hospitalization admissions involve genetic diseases. This number would increase substantially if one included complex, multifactorial genetic diseases, such as diabetes or cardiovascular disease. The prevalence of genetic diseases, combined with their severity and chronic nature, imposes a great financial, social, and emotional burden on society.

Genetics has historically focused on chromosomal and metabolic disorders, reflecting the long-standing availability of techniques to diagnose these conditions. For example, conditions such as trisomy 21 (Down syndrome) or monosomy X (Turner syndrome) can be diagnosed using cytogenetics (Chap. 66). Likewise, many metabolic disorders (e.g., phenylketonuria, familial hypercholesterolemia) have been diagnosed using biochemical analyses. Recent advances in DNA diagnostics have extended the field of genetics to include virtually all medical specialties. In cardiology, for example, the molecular basis of inherited cardiomyopathies and ion channel defects that predispose to arrhythmias is being defined (Chaps. 230 and238). In neurology, genetics has unmasked the pathophysiology of a startling number of neurodegenerative disorders (Chap. 359). Hematology has evolved dramatically, from its incipient genetic descriptions of hemoglobinopathies to the current understanding of the molecular basis of red cell membrane defects, clotting disorders, and thrombotic disorders (Chaps. 106 and117). It is now abundantly clear that neoplasia and the acquisition of metastatic potential can be described in genetic terms (Chaps. 81,82, and83).

New concepts derived from genetic studies can sometimes clarify topics that were previously opaque. For example, although many different genetic defects can cause peripheral neuropathies, disruption of the normal folding of the myelin sheaths is frequently a common finalpathway (Chap. 379). Several genetic causes of obesity appear to converge on a physiologic pathway that involves products of the proopiomelanocortin polypeptide and the MC4R receptor, thus identifying a key mechanism for appetite control (Chap. 77). A similar situation is emerging for genetically distinct forms of Alzheimer disease, several of which lead to the formation of neurofibrillary tangles (Chap. 362). Increasingly, the identification of defective genes can pinpoint cellular pathways involved in key physiologic processes. Examples include identification of the cystic fibrosis conductance regulator (*CFTR*) gene, the Duchenne's muscular dystrophy (*DMD*) gene, which encodes dystrophin, and the fibroblast growth factor receptor-3 (*FGFR3*) gene, which is responsible for achondroplastic dwarfism. Similarly, transgenic and gene "knockout" models can help to unravel the physiologic function of genes. Genetic approaches have proven invaluable for the detection of infectious pathogens and are used clinically to identify agents that are difficult to culture such as mycobacteria, viruses, and parasites (Chap. 121). In many cases, molecular genetics has improved the feasibility and accuracy of diagnostic testing, enhanced our understanding of pathophysiology, and is beginning to open new avenues for therapy, including gene therapy (Chap. 69).

It is increasingly apparent that genetic background plays some role in virtually every medical condition. This is particularly true when one considers disease susceptibility, the interaction of genetic background with the environment, host responses to illness and to pharmaceutical agents, or the metabolism of drugs. Although genetics has traditionally been viewed through the window of relatively rare single-gene diseases, many disorders such as hypertension, asthma, diabetes, susceptibility to cardiovascular disease, and mental illness are also affected by genetic background, as often evident from a patient's family history. These complex genetic traits involve the contributions of many different genes, as well as environmental factors that can modify disease risk (Chap. 68).

The astounding rate at which new genetic information is being generated creates a major challenge for physicians and other health care providers. The terminology and techniques used for discovery evolve continuously. Much genetic information presently resides in computer databases or is being published in basic science journals. The ongoing development of bioinformatics promises to simplify this seemingly daunting onslaught of new information. It is now possible, for example, to search for genetic testing centers through a web site (http://www.genlink.wustl.edu) that can be accessed conveniently by organ system, disease state, or gene. Monogenic disorders are summarized in a large, continuously evolving compendium, referred to as the *Online Mendelian Inheritance in Man* (OMIM;http://www.ncbi.nlm.nih.gov/omim/). These and other databases (http://www.genebank.com) will expand rapidly in conjunction with advances in theHGP.

CHROMOSOMES AND DNA REPLICATION

ORGANIZATION OF DNA INTO CHROMOSOMES

Size of the Human Genome The human genome is divided into 23 different chromosomes, including 22 autosomes (numbered 1 to 22) and the X and Y sex chromosomes. Adult cells are diploid, meaning they contain two homologous sets of 22 autosomes and a pair of sex chromosomes. Females have two X chromosomes (XX), whereas males have one X and one Y chromosome (XY). As a consequence of meiosis, germ cells (sperm or oocytes) are haploid and contain one set of 22 autosomes and one of the sex chromosomes. At the time of fertilization, the diploid genome is reconstituted by pairing of the homologous chromosomes from the mother and father. With each cell division (mitosis), chromosomes are replicated, paired, segregated, and divided into two daughter cells (Chap. 66).

The genome is estimated to contain about 100,000 genes that are divided among the 23 chromosomes. A *gene* is a functional unit that is regulated by transcription (see below) and encodes a product, either RNA or protein, that exerts activity within the cell. Historically, genes were identified because they conferred specific traits that are transmitted from one generation to the next.

Human DNA is estimated to consist of about 3 billion base pairs (bp) of DNA per haploid genome. DNA length is normally measured in units of 1000 bp (kilobases, kb) or 1,000,000 bp (megabases, Mb). Not all DNA encodes genes. In fact, genes account for only about 10 to 15% of DNA. Much of the remaining DNA consists of highly repetitive sequences, the function of which is poorly understood. These repetitive DNA regions, along with nonrepetitive sequences that do not encode genes, may serve a structural role in the packaging of DNA into chromatin (DNA bound to histone proteins) and chromosomes (Fig. 65-1). If only 10% of DNA is expressed and there are 100,000 genes, the average gene would be about 3 kb in length. Although many genes are about this size, the range is quite broad. For example, some genes are only a few hundred bp, whereas others, like the *DMD* gene, are extraordinarily large (2 million bp).

Structure of DNA Each gene is composed of a linear polymer of DNA. DNA is a double-stranded helix composed of four different bases: adenine (A), thymidine (T), guanine (G), and cytosine (C). Adenine is paired to thymidine, and guanine is paired to cytosine, by hydrogen bond interactions that span the double helix. DNA has several remarkable features that make it ideal for the transmission of genetic information. It is relatively stable, at least in comparison to RNA or proteins. The double-stranded nature of DNA and its feature of strict base-pair complementarity permit faithful replication during cell division. As described below, complementarity also allows the transmission of genetic information from DNA ® RNA ® protein (Fig. 65-2). Messenger RNA (mRNA) is encoded by the so-called sense strand of the DNA double helix and is translated into proteins by ribosomes.

The presence of four different bases provides surprising genetic diversity. In the protein-coding regions of genes, the DNA bases are arranged into codons, a triplet of bases that specifies a particular amino acid. It is possible to arrange the four bases into 64 different triplet codons (4^3). Each codon specifies 1 of the 20 different amino acids, or a regulatory signal, such as stop translation. Because there are more codons than amino acids, the genetic code is degenerate; that is, most amino acids can be specified by several different codons. By arranging the codons in different combinations and in various lengths, it is possible to generate the tremendous diversity of primary protein structure.

REPLICATION OF DNA AND MITOSIS

Genetic information in DNA is transmitted to daughter cells under two different circumstances: (1) somatic cells divide by mitosis, allowing the diploid (2*n*) genome to replicate itself completely in conjunction with cell division; and (2) germ cells (sperm and ova) undergo meiosis, a process that enables the reduction of the diploid (2*n*) set of chromosomes to the haploid state (1*n*) (Chap. 66).

Prior to mitosis, cells exit the resting, or G_0state, and enter the cell cycle (Chap. 82). After traversing a critical checkpoint in G_1, cells undergo DNA synthesis (S phase), during which the DNA in each chromosome is replicated, yielding two pairs of sister chromatids (2*n*® 4*n*). The process of DNA synthesis requires stringent fidelity in order to avoid transmitting errors to subsequent generations of cells. Genetic abnormalities of DNA mismatch/repair include xeroderma pigmentosum, Bloom syndrome, ataxia telangiectasia, and hereditary nonpolyposis colon cancer (HNPCC), among others. Many of these disorders strongly redispose to neoplasia because of the rapid acquisition of additional mutations (Chap. 81). After completion of DNA synthesis, cells enter $G_2$and progress through a second checkpoint before entering mitosis. At this stage, the chromosomes condense and are aligned along the equatorial plate at metaphase. The two identical sister chromatids, held together at the centromere, divide and migrate to opposite poles of the cell (Fig. 66-3). After formation of a nuclear membrane around the two separated sets of chromatids, the cell divides and two daughter cells are formed, thus restoring the diploid (2*n*) state.

ASSORTMENT AND SEGREGATION OF GENES DURING MEIOSIS

Meiosis occurs only in germ cells of the gonads. It shares certain features with mitosis but involves two distinct steps of cell division that reduce the chromosome number to the haploid state. In addition, there is active recombination that generates genetic diversity. During the first cell division, two sister chromatids (2*n*® 4*n*) are formed for each chromosome pair and there is an exchange of DNA between homologous paternal and maternal chromosomes. This process involves the formation of *chiasmata*, structures that correspond to the DNA segments that cross over between the maternal and paternal homologues (Fig. 65-3). Usually there is at least one crossover on each chromosomal arm; recombination occurs more frequently in female meiosis than in male meiosis. Subsequently, the chromosomes segregate randomly. Because there are 23 chromosomes, there exist 2_{23}(>8 million) possible combinations of chromosomes. Together with the genetic exchanges that occur during recombination, chromosomal segregation generates tremendous diversity, and each gamete is genetically unique. The process of recombination, and the independent segregation of chromosomes, provide the foundation for performing linkage analyses, whereby one attempts to correlate the inheritance of certain chromosomal regions (or linked genes) with the presence of a disease or genetic trait (see below).

After the first meiotic division, which results in two daughter cells (2*n*), the two chromatids of each chromosome separate during a second meiotic division to yield four gametes with a haploid state (1*n*). When the egg is fertilized by sperm, the two haploid sets are combined, thereby restoring the diploid state (2*n*) in the zygote.

REGULATION OF GENE EXPRESSION

Mechanisms that regulate gene expression play a critical role in the function of genes. The new field of *functional genomics* is based on the concept that understanding gene regulation and function will provide a better understanding of physiology and offer novel therapeutic opportunities. The transcription of genes is controlled primarily by *transcription factors* that bind to DNA sequences in the regulatory regions of genes. As described below, mutations in transcription factors cause an unexpectedly large number of genetic disorders. Gene expression is also influenced by *epigenetic events*, such as X-inactivation and imprinting, processes in which DNA methylation is associated with the silencing (i.e., suppression) of expression. Several genetic disorders, such as Prader-Willi syndrome (neonatal hypotonia, developmental delay, obesity, short stature, and hypogonadism) and Albright hereditary osteodystrophy (resistance to parathyroid hormone, short stature, brachydactyly, resistance to other hormones in certain subtypes), exhibit the consequences of genomic imprinting. Most studies of gene expression have focused on the regulatory DNA elements of genes that control transcription. However, it should be emphasized that gene expression requires a series of steps including mRNA processing, protein translation, and posttranslational modifications, all of which are actively regulated (Fig. 65-2).

STRUCTURE OF GENES

A gene product is usually a protein but can occasionally consist of RNA that is not translated. *Exons* refer to the portion of genes that are eventually spliced together to form mRNA. *Introns* refer to the spacing regions between the exons that are spliced out of precursor RNAs during RNA processing (Fig. 65-2).

The gene locus also includes regions that are necessary to control its expression. The regulatory regions most commonly involve sequences upstream (5¢) of the transcription start site, although there are also examples of control elements within introns or downstream of the coding regions of a gene. The upstream regulatory regions are also referred to as the *promoter*. The minimal promoter usually consists of a TATA box (which binds TATA-binding protein, TBP) and initiator sequences that enhance the formation of an active transcription complex. Transcriptional termination signals reside downstream, or 3¢, of a gene. Specific sequences, such as the AAUAAA sequence at the 3¢ end of the mRNA, designate the site for polyadenylation (poly-A tail), a process that influences mRNA transport to the cytoplasm, stability, and translation efficiency. A rigorous test of the regulatory region boundaries involves expressing a gene in a transgenic animal to determine whether the isolated DNA flanking sequences are sufficient to recapitulate the normal developmental, tissue-specific, and signal-responsive features of the endogenous gene. This has been accomplished for only a few genes; there are many examples in which large genomic fragments only partially reconstitute normal gene regulation in vivo, implying the presence of distant regulatory sequences. This approach is critical to our understanding of mechanisms that regulate genes and is also relevant for gene therapy strategies that require normal gene regulation (Chap. 69).

As genes are dissected with greater resolution, the number of DNA sequences and transcription factors that regulate transcription is much greater than originally anticipated. Most genes contain at least 15 to 20 discrete regulatory elements within 300 bp of the transcription start site. This densely packed promoter region often contains binding sites for ubiquitous transcription factors such as CAAT box/enhancer binding protein (C/EBP), cyclic AMP response element binding (CREB) protein, selective promoter factor 1 (Sp-1), or activator protein 1 (AP-1). However, factors involved in cell-specific expression may also bind to these sequences. For example, basic helix-loop-helix (bHLH) proteins bind to E-boxes in the promoters of myogenic genes, and steroidogenic factor 1 (SF-1) binds to a specific recognition site in the regulatory region of multiple steroidogenic enzyme genes. Key regulatory elements may also reside at some distance from the proximal promoter. The globin and the immunoglobulin genes, for example, contain *locus control regions* that are several kilobases away from the structural sequences of the gene. Specific groups of transcription factors that bind to these promoter and enhancer sequences provide a combinatorial code for regulating transcription. In this manner, relatively ubiquitous factors interact with more restricted factors to allow each gene to be expressed and regulated in a unique manner. As described below, the transcription factors that bind to DNA actually represent only the first level of regulatory control. Other proteins -- *coactivators* and *corepressors* -- interact with the DNA-binding transcription factors to generate large regulatory complexes. These complexes are subject to control by numerous cell-signaling pathways, including phosphorylation and acetylation. Ultimately, the recruited transcription factors interact with, and stabilize, components of the basal transcription complex that assembles at the site of the TATA box and initiator region. This basal transcription factor complex consists of >30 different proteins. Gene transcription occurs when RNA polymerase begins to synthesize RNA from the DNA template.

TRANSCRIPTIONAL ACTIVATION AND REPRESSION

Every gene is controlled uniquely, whether in its spatial or temporal pattern of expression or in its response to extracellular signals. It is estimated that transcription factors account for about 30% of expressed genes. A growing number of identified genetic diseases involve transcription factors (Table 65-1). The MODY (maturity-onset diabetes of the young) disorders are representative of this group of diseases; mutations in several different islet cell-specific transcription factors cause various forms of MODY (Chap. 333).

Transcriptional activation can be divided into three main mechanisms:

1. Events that alter chromatin structure can enhance the access of transcription factors to DNA. For example, histone acetylation opens chromatin structure and is correlated with transcriptional activation.

2. Posttranslational modifications of transcription factors, such as phosphorylation, can induce the assembly of active transcription complexes. As an example, phosphorylation ofCREBprotein on serine 133 induces a conformational change that allows the recruitment of CREB-binding protein (CBP), a factor that integrates the actions of many transcription factors, including proteins, with histone acetyltransferase activity.

3. Transcriptional activators can displace a repressor protein. This mechanism is particularly common during development when the pattern of transcription factor expression changes dynamically.

Of course, these mechanisms are not mutually exclusive, and most genes are activated by some combination of these events.

In general, mechanisms of transcriptional repression have not been studied to the same extent as mechanisms of transcriptional activation. Nonetheless, suppression of gene expression is as important as gene activation. Some mechanisms of repression are the corollary of activation. For example, repression is often associated with histone deacetylation or protein dephosphorylation. For nuclear hormone receptors, transcriptional silencing involves the recruitment of repression complexes that contain histone deacetylase activity. Aberrant expression of repressor proteins is sometimes associated with neoplasia. The t(15;17) chromosomal translocation that occurs in promyelocytic leukemia fuses the *PML* gene to a portion of the retinoic acid receptora (*RAR a*) gene (Table 65-1). This event causes unregulated transcriptional repression in a manner that precludes normal cellular differentiation. The addition of the RAR ligand, retinoic acid, activates the receptor, thereby relieving repression and allowing cells to differentiate and ultimately undergo apoptosis. This mechanism has therapeutic importance as the addition of retinoic acid to treatment regimens induces a higher remission rate in patients with promyelocytic leukemia (Chap. 111).

CLONING AND SEQUENCING DNA

Since the mid-1970s, eight Nobel prizes have been awarded for research that led, directly or indirectly, to major methodological advances as well as to profound insights into genetics. Examples include the discoveries of reverse transcriptase, restriction enzymes, plasmid cloning vectors, DNA sequencing, andPCR. A description of recombinant DNA techniques, the methodology used for the manipulation, analysis, and characterization of DNA segments, is beyond the scope of this chapter. As these methods are widely used in genetics and molecular diagnostics, however, it is useful to review briefly some of the fundamental principles of cloning and DNA sequencing.

CLONING OF GENES

Cloning refers to the creation of a recombinant DNA molecule that can be propagated indefinitely. The ability to clone genes and cDNAs therefore provides a permanent and renewable source of these reagents. Cloning is essential for DNA sequencing, nucleic acid hybridization studies, expression of recombinant proteins, and other recombinant DNA procedures.

The cloning of DNA involves the insertion of a DNA fragment into a cloning vector, followed by the propagation of the recombinant DNA in a host cell. The most straightforward cloning strategy involves inserting a DNA fragment into bacterial plasmids. Plasmids are small, autonomously replicating, circular DNA molecules that propagate separately from the chromosome in bacterial cells. The process of DNA insertion relies heavily on the use of restriction enzymes, which cleave DNA at highly specific sequences (usually 4 to 6 bp in length). Restriction enzymes generate complementary, cohesive sequences at theends of the DNA fragment, which allow them to be efficiently ligated to the plasmid vector. Because plasmids contain genes that confer resistance to antibiotics, their presence in the host cell can be used for selection and DNA amplification.

A variety of vectors and appropriate hosts are now used for cloning (Table 65-2). Many of these are used for creating *libraries*, a term that refers to a collection of DNA clones. A genomic library represents an array of clones derived from genomic DNA. These overlapping DNA fragments represent the entire genome and can ultimately be arranged according to their linear order. Genomic libraries are propagated using a variety of vectors, such as lambda (l) phage, cosmids, bacterial artificial chromosomes (BACs), and yeast artificial chromosomes (YACs). Phage libraries have been used extensively to isolate specific genes. Cosmids, BACs, and YACs are particularly useful for studying large genes and for defining the order of genes along the chromosomes (Fig. 65-4). cDNA libraries reflect clones derived from mRNA, typically from a particular tissue source. Thus, a cDNA library from the heart contains copies of mRNA expressed specifically in cardiac myocytes, in addition to those that are expressed ubiquitously. For this reason, a heart cDNA library will be enriched with cardiac-specific gene products and will differ from cDNA libraries generated from liver or pituitary mRNAs. As an example of the complexity of a genomic library, consider that the human genome contains 3´10_9 bp and the average genomic insert in a l phage library is about 10_4 bp. Therefore, it requires at least 3 ´10_5clones to represent all of the genomic DNA. Specific clones are isolated from the several hundred thousand clones by using DNA hybridization.

With completion of theHGP, all human genes have been cloned and sequenced. As a result, many of these cloning procedures will be unnecessary or greatly facilitated by the extensive information concerning DNA markers and the sequence of DNA (see below).

NUCLEIC ACID HYBRIDIZATION

Nucleic acid *hybridization* is a fundamental principle in molecular biology that takes advantage of the fact that the two complementary strands of nucleic acids bind, or *hybridize*, to one another with very high specificity. The goal of hybridization is to detect specific nucleic acid (DNA or RNA) sequences in a complex background of other sequences. This technique is used for Southern blotting, northern blotting, and for screening libraries (see above). Further adaptation of hybridization techniques has led to the development of microarray DNA chips.

Southern Blot Southern blotting is used to analyze whether genes have been deleted or rearranged. It is also used to detect restriction fragment length polymorphisms (RFLPs). Genomic DNA is digested with restriction endonucleases and separated by gel electrophoresis. Individual fragments can then be transferred to a membrane and detected after hybridization with specific radioactive DNA probes. Because single base-pair mismatches can disrupt the hybridization of short DNA probes (oligonucleotides), a variation of the Southern blot, termed *oligonucleotide-specific hybridization* (OSH), uses short oligonucleotides to distinguish normal from mutant genes.

Northern Blot Northern blots are used to analyze patterns and levels of gene expression in different tissues. In a northern blot, mRNA is separated on a gel and transferred to a membrane, and specific transcripts are detected using radiolabeled DNA as a probe. This technique is rapidly being supplanted by more sensitive and comprehensive methods such as reverse transcriptase (RT)-PCR and gene expression arrays on DNA chips (see below).

Microarray Technology A rapidly evolving approach to genome-scale studies consists of *microarrays*, or *DNA chips*. These approaches consist of thousands of synthetic nucleic acid sequences aligned on thin glass or silicon surfaces. Fluorescently labeled test sample DNA or RNA is hybridized to the chip, and a computerized scanner detects sequence matches. Microarrays allow the detection of variations in DNA sequence and are used for mutational analysis and genotyping. Alternatively, the expression pattern of large numbers of mRNA transcripts can be determined by hybridization of RNA samples to cDNA or genomic microarrays. This method has tremendous potential in the era of functional genomics. As one example, microarrays can be used to develop genetic fingerprints of different types of lymphomas, providing information useful for classification, pathophysiology, prognosis, and treatment.

THE POLYMERASE CHAIN REACTION

ThePCR, introduced in 1985, has revolutionized the way DNA analyses are performed and has become a cornerstone of molecular biology and genetic analysis. In essence, PCR provides a rapid way of cloning (amplifying) specific DNA fragments in vitro (Fig. 65-5). Exquisite specificity is conferred by the use of PCR primers, which are designed for a given DNA sequence. The geometric amplification of the DNA after multiple cycles yields remarkable sensitivity. As a result, PCR can be used to amplify DNA from very small samples, including single cells. These properties also allow DNA amplification from a variety of tissue sources including blood samples, biopsies, surgical or autopsy specimens, or cells from hair or saliva. PCR can also be used to study mRNA. In this case, the enzymeRT is first used to convert the RNA to DNA, which can then be amplified by PCR. This procedure, commonly known as *RT-PCR*, is useful as a quantitative measure of gene expression.

PCRprovides a key component of molecular diagnostics. It provides a strategy for the rapid amplification of DNA (or mRNA) to search for mutations by a wide array of techniques, including DNA sequencing. PCR is also used for the amplification of highly polymorphic di- or trinucleotide repeat sequences, which allow various polymorphic alleles to be traced in genetic linkage or association studies. PCR is increasingly used to diagnose various microbial pathogens (Chap. 121).

DNA SEQUENCING

DNA sequencing is now an automated procedure. Although many protocols exist, the most commonly used strategy is based on the Sanger method in which dideoxynucleotides are used to randomly terminate DNA polymerization at each of the four bases (A,G,T,C). After separating the array of terminated DNA fragments using high-resolution gel or capillary electrophoresis, it is possible to deduce the DNA sequence by examining the progression of fragment lengths generated in each of the four nucleotide reactions. The use of fluorescently labeled dideoxynucleotides allows automated detection of the different bases and direct computer analysis of the DNA sequence. Efforts are underway to develop faster, more cost-effective DNA sequencing technologies. These include the use of mass spectrometry; detection of fluorescently labeled bases in flow cytometry; direct reading of the DNA sequence by scanning, tunneling, or atomic force microscopy; and sequence analysis using DNA chips.

TRANSGENIC MICE AS MODELS OF GENETIC DISEASE

Several organisms have been studied extensively as genetic models, including *Mus musculus* (mouse), *Drosophila melanogaster* (fruit fly), *Caenorhabditis elegans* (nematode), *Saccharomyces cerevisiae* (baker's yeast), and *Eschericia coli* (colonic bacterium). The ability to use these evolutionarily distant organisms as genetic models that are relevant to human physiology reflects a surprising conservation of genetic pathways and gene function. Transgenic mouse models have been particularly valuable, because many human and mouse genes exhibit similar structure and function and because manipulation of the mouse genome is relatively straightforward compared to those of other mammalian species.

Transgenic strategies in mice can be divided into two main approaches: (1) overexpression of a gene by random insertion into the genome, and (2) deletion or targeted mutagenesis of a gene by homologous recombination with the native endogenous gene (Fig. 65-6). Many variations of these basic approaches now exist that allow genes to be expressed in specific cell types, at different times during development, or at varying levels. Consequently, transgenic technology has emerged as a powerful strategy for defining the physiologic effects of deleting or overexpressing a gene, as well as providing unique genetic models for dissecting pathophysiology or testing therapies.

Examples of transgenic models relevant to human genetic disorders are listed inTable 65-3. Transgenic overexpression of genes is useful for studying disorders that are sensitive to gene dosage. Overexpression of *PMP22*, for example, mimics a common duplication of this gene in type IA Charcot-Marie-Tooth disease (Chap. 379). Duplication of the *PMP22* gene results in high levels of expression of peripheral myelin protein 22, and this dosage effect is responsible for the demyelinating neuropathy. Expression of the Y chromosome-specific gene, *SRY*, in XX females demonstrates that *SRY* is sufficient to induce the formation of testes. This finding confirms the pathogenic role of *SRY* translocations to the X chromosome in sex-reversed XX females. Huntington disease is an autosomal dominant disorder caused by expansion of a CAG trinucleotide repeat that encodes a polyglutamine tract. Targeted deletion of the Huntington disease (*HD*) gene does not induce the neurologic disorder. On the other hand, transgenic expression of the entire gene or of the first exon containing the expanded polyGlu repeat is sufficient to cause many features of the neurologic disorder, indicating a gain-of-function property for the expanded polyGlu-containing protein. Transgenic strategies can also be used as a precursor to gene therapy. Expression of dystrophin, the protein that is deleted in Duchenne muscular dystrophy, partially corrects the disorder in a mouse model of Duchenne's. Targeted expression of oncogenes has been valuable to study mechanisms of neoplasia and to generate immortalized cell lines. For example, expression of the simian virus 40 (SV40) large T antigen under the direction of the insulin promoter induces the formation of islet cell tumors.

Targeted deletion mutagenesis, commonly known as *gene knockout*, is performed using a targeting vector that carries a mutant version of the gene. After homologous recombination in embryonic stem (ES) cells, chimeric animals are produced by injection of ES cells into blastocysts. Subsequently, animals are bred to be heterozygous or homozygous for the mutation. In addition to their use in examining gene function, gene knockouts provide valuable animal models for many loss-of-function mutations. A variation of this strategy is to use cre recombinase to induce genetic recombination in vivo. Cre recombinase will delete genes that have been flanked by its recognition sequences, called *loxP* sites. One advantage to this approach is that transgenic expression of cre in specific tissues can be used to delete a gene in a tissue-specific or developmentally staged manner. This is particularly useful for genes that would be lethal if deleted universally or during early development.

The list of genes that have been knocked out in mice is already very large. Many of these knockouts do not have an apparent phenotype, either because of redundant functions of the other genes or because the phenotype is subtle. For example, deletion of the hypoxanthine phosphoribosyltransferase (HPRT) gene (*Hprt*) does not cause characteristic features of Lesch-Nyhan syndrome in mice because of their reliance on adenine phosphoribosyltransferase (APRT) in the purine salvage pathway. The administration of an APRT inhibitor to HPRT-deficient mice, however, results in the typical self-injurious behavior seen in patients with Lesch-Nyhan syndrome. The phenotypes of some knockouts are quite different from their human disease counterparts. For example, deletion of the retinoblastoma (*Rb*) gene does not lead to retinoblastoma or other tumors that characterize the human syndrome. These examples underscore the fact that the functions of genes, and their interactions with genetic background and the environment, cannot be assumed to be identical in mice and humans. On the other hand, the deletion of many genes provides a remarkably faithful model of human disorders (Table 65-3). In addition to clarifying pathophysiology, these models facilitate the development of therapies, both genetic and pharmaceutical.

In addition to transgenic animal models, naturally occurring mutations in mice and other species continue to provide fundamental insights into human disease. A compendium of natural and transgenic animal models is provided in continuously evolving databases (Online Mendelian Inheritance in Animals OMIA:http://www.angis.su.oz.au/Databases/BIRX/omia/; The Jackson Library:http://www.jax.org/).

Human pluripotential *stem cells* have recently been developed, and, consistent with their potential for self-renewal, these cell lines express high levels of telomerase, an enzyme that is essential for allowing repeated replication of the ends of eukaryotic chromosomes. Although much remains to be learned about the properties of pluripotential stem cells, they may prove useful for transplantation, drug testing, or for other purposes.

THE HUMAN GENOME PROJECT

TheHGP was initiated in the mid-1980s as an ambitious effort to characterize the human genome, culminating in a complete DNA sequence. In the United States, the National Institutes of Health (NIH) and the Department of Energy (DOE) officially launched the genome project in 1990; the project has evolved as an international effort and has also included important contributions from the private sector. The main goals include: (1) creation of genetic maps, (2) development of physical maps, and (3) determination of the complete human DNA sequence.

Some analogies help in appreciating the scope of theHGP. The 23 pairs of human chromosomes are thought to encode approximately 100,000 genes. The total length of DNA is about 3 billion bp, which is nearly 1000-fold greater than the *E. coli* genome. If the human DNA sequence were printed out, it would correspond to about 120 volumes of *Harrison's Principles of Internal Medicine*.

THE GENETIC MAP

Given the size and complexity of the human genome, genetic maps have been developed to provide orientation and to delimit where a gene of interest may be located. A *genetic map* describes the order of genes and defines the position of a gene relative to other loci on the same chromosome. It is constructed by assessing how frequently two markers are inherited together by linkage studies. Distances of the genetic map are expressed in recombination units, or centimorgans (cM). One cM corresponds to a recombination frequency of 1% between two polymorphic markers; 1 cM corresponds to approximately 1 Mb of DNA (Fig. 65-3). Any polymorphic sequence variation can be useful for mapping purposes. Examples of polymorphic markers include variable number of tandem repeats (VNTRs),RFLPs, microsatellite repeats, and single nucleotide polymorphisms (SNPs); the latter two methods are now used predominantly because of the high density of markers and because they are amenable to automated procedures.

The current genetic map exists at about 1 cM resolution. A goal for the near term is to add about 100,000SNPs to these maps. This would provide 1 SNP approximately every 100,000 bp. The addition of SNPs will facilitate automation using DNA chips and will enhance the ability to perform linkage studies of complex genetic diseases.

THE PHYSICAL MAP

Cytogenetics and chromosomal banding techniques provide a relatively low-resolution microscopic view of genetic loci. Physical maps indicate the position of a locus or gene in absolute values. Sequence-tagged sites (STSs) are used as a standard unit for physical mapping and serve as sequence-specific landmarks for arranging overlapping cloned fragments in the same order as they occur in the genome. These overlapping clones, usually inYACs orBACs, allow the characterization of contiguous DNA sequences, commonly referred to as *contigs* (Fig. 65-4). The STSs consist of 200 to 500 bp, which can be retrieved from computer databases; >50,000 STSs have been mapped. The goal of achieving a high-resolution physical map of the human genome has essentially been achieved as all of the genome has been cloned into overlapping fragments. The highest resolution physical map will provide the complete DNA sequence of each chromosome in the human genome.

STATUS OF DNA SEQUENCING

The primary focus of the genome project is to obtain DNA sequence for the entire human genome as well as model organisms. The sequences of *E. coli* and many other bacteria, *S. cerevisiae*, *C. elegans*, and *D. melanogaster* have already been completed. Sequencing of the laboratory mouse genome is in progress. Although the prospect of determining the complete sequence of the human genome was a daunting prospect several years ago, technical advances in DNA sequencing and bioinformatics have led to the completion of a draft human sequence in June 2000, well in advance of the original goal of the year 2003. The current standard is to achieve 99.99% (1 error in 10,000 bp) accuracy. This level of accuracy is important for many reasons, including efforts to determine the degree of DNA sequence variation in the population. Comparisons of the DNA sequence from multiple individuals or populations will allow assessments of genetic variance in the human population. Another goal is to develop a complete set of full-length human cDNAs and to define their locations on the physical map.

ETHICAL ISSUES

Implicit in theHGP is the idea and hope that identifying disease-causing genes can lead to improvements in diagnosis, prognosis, and treatment. It is estimated that most individuals harbor several serious recessive genes. However, completion of the human genome sequence, determination of the association of genetic defects with disease, and studies of genetic variation raise many new issues with implications for the individual and mankind. The controversies concerning the cloning of mammals and the establishment of human embryonic stem cells underscore the relevance of these questions. Moreover, the information gleaned from genotypic results can have quite different impacts, depending on the availability of strategies to modify the course of disease. For example, the identification of mutations that cause multiple endocrine neoplasia (MEN) type 2 or hemochromatosis allows specific interventions for affected family members. On the other hand, at present the identification of an Alzheimer or Huntington disease gene does not alter therapy. Genetic test results can generate anxiety in affected individuals and family members, and there is the possibility of discrimination on the basis of the test results. Most genetic disorders are likely to fall into an intermediate category where the opportunity for prevention or treatment is significant but limited (Chap. 68). For these reasons, the scientific components of the HGP have been paralleled by efforts to examine ethical and legal implications as new issues arise.

Many issues raised by the genome project are familiar, in principle, to medical practitioners. For example, an asymptomatic patient with increased low-density lipoprotein (LDL) cholesterol, high blood pressure, or a strong family history of early myocardial infarction, is known to be at increased risk of coronary heart disease. In such cases, it is clear that the identification of risk factors and an appropriate intervention are beneficial. Likewise, patients with phenylketonuria, cystic fibrosis, or sickle cell anemia are often identified as having a genetic disease early in life. These precedents can be helpful for adapting policies that relate to genetic information. We can anticipate similar efforts, whether based on genotypes or other markers of genetic predisposition, to be applied to many disorders. One confounding aspect of the rapid expansion of information is that our ability to make clinical predictions often lags behind genetic advances. For example, when genes that predispose to breast cancer, such as *BRCA1*, are described, they generate tremendous public interest in the potential to predict disease, but many years of clinical research are still required to rigorously establish genotype and phenotype correlations.

Whether related to informed consent, participation in research, or the management of a genetic disorder that affects an individual or their families, there is a great need for more information about fundamental principles of genetics. The pervasive nature of the role of genetics in medicine makes it imperative for physicians and other health care professionals to become more informed about genetics and to provide advice and counseling in conjunction with trained genetic counselors (Chap. 68). The application of screening and prevention strategies will therefore require intensive patient and physician education, changes in health care financing, and legislation to protect patient's rights.

TRANSMISSION OF GENETIC DISEASE

ORIGINS AND TYPES OF MUTATIONS

A *mutation* can be defined as any change in the primary nucleotide sequence of DNA regardless of its functional consequences. Some mutations may be lethal, others are less deleterious, and some may confer an evolutionary advantage. Mutations can occur in the germline (sperm or oocytes); these can be transmitted to progeny. Alternatively, mutations can occur during embryogenesis or in somatic tissues. Mutations that occur during development lead to *mosaicism*, a situation in which tissues are composed of cells with different genetic constitutions. If the germline is mosaic, a mutation can be transmitted to some progeny but not others, which sometimes leads to confusion in assessing the pattern of inheritance. Somatic mutations that do not affect cell survival can sometimes be detected because of variable phenotypic effects in tissues (e.g., pigmented lesions in McCune-Albright syndrome). Other somatic mutations are associated with neoplasia because they confer a growth advantage to cells. Epigenetic events such as altered DNA methylation may also influence gene expression. With the exception of triplet nucleotide repeats, which can expand (see below), mutations are usually stable.

Mutations are structurally diverse -- they can involve the entire genome, as in triploidy (one extra set of chromosomes), or gross numerical or structural alterations in chromosomes or individual genes (Chap. 66). Large deletions may affect a portion of a gene or an entire gene, or, if several genes are involved, they may lead to a *contiguous gene syndrome*. Unequal crossing-over between homologous genes can result in fusion gene mutations, as illustrated by color blindness (Chap. 28). Mutations involving single nucleotides are referred to as *point mutations*. Substitutions are called *transitions* if a purine is replaced by another purine base (A« G) or if a pyrimidine is replaced by another pyrimidine (C « T). Changes from a purine to a pyrimidine, or vice versa, are referred to as *transversions*. If the DNA sequence change occurs in a coding region and alters an amino acid, it is called a *missense mutation*. Depending on the functional consequences of such a missense mutation, amino acid substitutions in different regions of the protein can lead to distinct phenotypes. *Polymorphisms* are sequence variations that have a frequency of at least 1%. Usually, they do not result in a perceptible phenotype. Often they consist of single base-pair substitutions that do not alter the protein coding sequence because of the degenerate nature of the genetic code, although it is possible that some might alter mRNA stability, translation, or the amino acid sequence. These types of silent base substitutions andSNPs are encountered frequently during genetic testing and must be distinguished from true mutations that alter protein expression or function. Small nucleotide deletions or insertions cause a shift of the codon reading frame. Most commonly, reading frame alterations result in an abnormal protein segment of variable length before termination of translation occurs at a stop codon (*nonsense mutation*). Mutations in intronic sequences or in exon junctions may destroy or create splice donor or splice acceptor sites. Mutations may also be found in the regulatory sequences of genes, resulting in reduced gene transcription.

Mutation Rates As noted before, mutations represent an important cause of genetic diversity as well as disease. Mutation rates are difficult to determine in humans because many mutations are silent and because testing is often not adequate to detect the phenotypic consequences. Mutation rates vary in different genes but are estimated to occur at a rate of about 10^{-10}/bp per cell division. Germline mutation rates (as opposed to somatic mutations) are relevant in the transmission of genetic disease. Because the population of oocytes is established very early in development, only about 20 cell divisions are required for completed oogenesis, whereas spermatogenesis involves about 30 divisions by the time of puberty and 20 cell divisions each year thereafter. Consequently, the probability of acquiring new point mutations is much greater in the male germline than the female germline, in which rates of aneuploidy are increased (Chap. 66). Thus, the incidence of new point mutations in spermatogonia increases with paternal age (e.g., achondrodysplasia, Marfan syndrome, neurofibromatosis). It is estimated that about 1 in 10 sperm carries a new deleterious mutation. The rates for new mutations are calculated most readily for autosomal dominant and X-linked disorders and are ~$10^{-5}$to 10^{-6}/locus per generation. Because most monogenic diseases are relatively rare, new mutations account for a significant fraction of cases. This is important in the context of genetic counseling, as a new mutation can be transmitted to the affected individual but does not necessarily imply that the parents are at risk to transmit the disease to other children. An exception to this is when the new mutation occurs early in germline development, leading to *gonadal mosaicism*.

Unequal Crossing-Over Normally, DNA recombination in germ cells occurs with remarkable fidelity to maintain the precise junction sites for the exchanged DNA sequences (Fig. 65-2). However, mispairing of homologous sequences leads to unequal crossover, with gene duplication on one of the chromosomes and gene deletion on the other chromosome. A significant fraction of growth hormone (*GH*) gene deletions, for example, involve unequal crossing-over (Chap. 328). The *GH* gene is a member of a large gene cluster that includes a growth hormone variant gene as well as several structurally related chorionic somatomammotropin genes and pseudogenes (highly homologous but functionally inactive relatives of a normal gene). Because such gene clusters contain multiple homologous DNA sequences arranged in tandem, they are particularly prone to undergo recombination and, consequently, gene duplication or deletion. On the other hand, duplication of the *PMP22* gene as a result of unequal crossing-over results in increased gene dosage and type IA Charcot-Marie-Tooth disease (Chap. 379). Unequal crossing-over resulting in deletion of *PMP22* results in a distinct neuropathy called *hereditary liability to pressure palsy* (Chap. 379).

Glucocorticoid-remediable aldosteronism (GRA) is caused by a rearrangement involving the genes that encode aldosterone synthase (*CYP11B2*) and steroid 11b-hydroxylase (*CYP11B1*), normally arranged in tandem on chromosome 8q. These two genes are 95% identical, predisposing to gene duplication and deletion by unequal crossing-over. The rearranged gene product contains the regulatory regions of 11b-hydroxylase fused to the coding sequence of aldosterone synthetase. Consequently, the latter enzyme is expressed in the adrenocorticotropic hormone (ACTH)-dependent zone of the adrenal gland, resulting in overproduction of mineralocorticoids and hypertension (Chap. 331).

Gene conversion refers to a nonreciprocal exchange of homologous genetic information; it is probably more common than generally recognized. In human genetics, gene conversion has been used to explain how an internal portion of a gene is replaced by a homologous segment copied from another allele or locus; these genetic alterations may range from a few nucleotides to a few thousand nucleotides. As a result of gene conversion, it is possible for short DNA segments of two chromosomes to be identical, even though these sequences are distinct in the parents. A practical consequence of this phenomenon is that nucleotide substitutions can occur during gene conversion between related genes, often altering the function of the gene. In disease states, gene conversion often involves intergenic exchange of DNA between a gene and a related pseudogene. For example, the 21-hydroxylase gene (*CYP21A*) is adjacent to a nonfunctional pseudogene. Many of the nucleotide substitutions that are found in the *CYP21A* gene in patients with congenital adrenal hyperplasia correspond to sequences that are present in the pseudogene, suggesting gene conversion as a mechanism of mutagenesis. In addition, mitotic gene conversion has been suggested as a mechanism to explain revertant mosaicism in which an inherited mutation is "corrected" in certain cells. For example, patients with autosomal recessive generalized atrophic benign epidermolysis bullosa have acquired reverse mutations in one of the two mutated *COL17A1* alleles, leading to clinically unaffected patches of skin.

Insertions and Deletions Though many instances of insertions and deletions occur as a consequence of unequal crossing-over, there is also evidence for internal duplication, inversion, or deletion of DNA sequences. The fact that certain deletions or insertions appear to occur repeatedly as independent events suggests that specific regions within the DNA sequence predispose to these errors. For example, certain regions of the *DMD* gene appear to be hot spots for deletions.

Errors in DNA Repair Because mutations caused by defects in DNA repair accumulate as somatic cells divide, these types of mutations are particularly important in the context of neoplastic disorders (Chap. 82). Several genetic disorders involving DNA repair enzymes underscore their importance. Patients with xeroderma pigmentosum have defects in DNA damage recognition or in the nucleotide excision and repair pathway (Chap. 86). Exposed skin is dry and pigmented and is extraordinarily sensitive to the mutagenic effects of ultraviolet irradiation. More than 10 different genes have been shown to cause the different forms of xeroderma pigmentosum. This finding is consistent with the earlier classification of this disease into different complementation groups (Table 65-4) in which normal function is rescued by the fusion of cells derived from two different forms of xeroderma pigmentosum.

Ataxia telangiectasia causes large telangiectatic lesions of the face, cerebellar ataxia, immunologic defects, and hypersensitivity to ionizing radiation (Chap. 364). The discovery of the ataxia telangiectasia mutated (*ATM*) gene reveals that it is homologous to genes involved in DNA repair and control of cell cycle checkpoints. Mutations in the *ATM* gene give rise to defects in meiosis as well as increasing susceptibility to damage from ionizing radiation. Fanconi's anemia is also associated with an increased risk of multiple acquired genetic abnormalities. It is characterized by diverse congenital anomalies and a strong predisposition to develop aplastic anemia and acute myelogenous leukemia (Chap. 111). Cells from these patients are susceptible to chromosomal breaks caused by a defect in genetic recombination. At least eight different complementation groups have been identified, and several loci and genes associated with Fanconi's anemia have been mapped or cloned (Table 65-4).

HNPCCis caused by mutations in one of several different mismatch repair (MMR) genes including MutS homologue 2 (*MSH2*) and MutL homologue 1 (*MLH1*) (Chap. 90). These enzymes are involved in the detection of nucleotide mismatches and in the recognition of slipped-strand trinucleotide repeats. Germline mutations in these genes lead to microsatellite instability and a high mutation rate in colon cancer. This syndrome is characterized by autosomal dominant transmission of colon cancer, young age (<50 years) of presentation, predisposition to lesions in the proximal large bowel, and associated malignancies such as uterine cancer and ovarian cancer. Genetic screening tests for this disorder are now being used for families considered to be at risk (Chap. 68). Recognition of HNPCC allows early screening with colonoscopy and the implementation of prevention strategies using nonsteroidal anti-inflammatory drugs.

CpG and Dipyrimidine Sequences Certain DNA sequences are particularly susceptible to mutagenesis. Successive pyrimidine residues (e.g., T-T or C-C) are subject to the formation of ultraviolet light-induced photoadducts. If these pyrimidine dimers are not repaired by the nucleotide excision repair pathway, mutations will be introduced after DNA synthesis. The dinucleotide C-G, or CpG, is also a hot spot for a specific type of mutation. In this case, methylation of the cytosine is associated with an enhanced rate of deamination to uracil, which is then replaced with thymine. This C ® T transition (or G ® A on the opposite strand) accounts for at least one-third of point mutations associated with polymorphisms and mutations. Many of the *MSH2* mutations inHNPCC, for example, involve CpG sequences.

Certain types of mutations (C ® T or G® A) are relatively common. Moreover, the redundant nature of the genetic code results in overrepresentation of certain amino acid substitutions. For example, arginine codons are most likely to be converted to cysteine, tryptophan, or a stop codon when a C® T transition occurs, and to histidine or glutamine when a G ® A transition occurs.

Unstable DNA Sequences *Trinucleotide repeats* may be unstable and expand beyond a critical number. Mechanistically, the expansion is thought to be caused by unequal recombination and slipped mispairing. A premutation represents a small increase in trinucleotide copy number. In subsequent generations, the expanded repeat may increase further in length and result in an increasingly severe phenotype, a process called *dynamic mutation* (see below for discussion of anticipation). Trinucleotide expansion was first recognized as a cause of the fragile X syndrome, one of the most common causes of mental retardation (Chap. 359). Other disorders arising from a similar mechanism include Huntington disease (Chap. 362), X-linked spinobulbar muscular atrophy (Chap. 365), and myotonic dystrophy (Chap. 383) (Tables 65-5 and65-6). Malignant cells are also characterized by genetic instability, indicating a breakdown in mechanisms that regulate DNA repair and the cell cycle.

FUNCTIONAL CONSEQUENCES OF MUTATIONS

Functionally, mutations can be broadly classified as gain-of-function and loss-of-function mutations. Gain-of-function mutations are typically dominant; that is, they result in phenotypic alterations when a single allele is affected. Inactivating mutations are usually recessive, and an affected individual is homozygous or compound heterozygous (i.e., carrying two different mutant alleles) for the disease-causing mutations. Alternatively, mutation in a single allele can result in *haploinsufficiency*, a situation in which one normal allele is not sufficient for a normal phenotype. This phenomenon applies, for example, to expression of rate-limiting enzymes in heme synthesis that cause porphyrias (Chap. 346). An increase in dosage of a gene product may also result in disease, as illustrated by the duplication of the *DAX1* gene in dosage-sensitive sex-reversal (Chap. 338). Mutation in a single allele can also result in loss of function due to a dominant-negative effect. In this case, the mutated allele interferes with the function of the normal gene product by one of several different mechanisms: (1) a mutant protein may interfere with the function of a multimeric protein complex, as illustrated by mutations in type 1 collagen (*COL1A1*, *COL1A2*) genes in osteogenesis imperfecta (Chap. 351); (2) a mutant protein may occupy binding sites on proteins or promoter response elements, as illustrated by thyroid hormone resistance, a disorder in which inactivated thyroid hormone receptor binds to target genes and functions as an antagonist of normal receptors (Chap. 330); or (3) a mutant protein can be cytotoxic as in $a_1$antitrypsin deficiency (Chap. 258) or autosomal dominant neurohypophyseal diabetes insipidus (Chap. 329), in which the abnormally folded proteins are trapped within the endoplasmic reticulum and ultimately cause cellular damage.

GENOTYPE AND PHENOTYPE

Alleles, Genotypes, and Haplotypes An observed trait is referred to as a *phenotype*; the genetic information defining the phenotype is called the *genotype*. Alternative forms of a gene or a genetic marker are referred to as alleles. Alleles may be polymorphic variants of nucleic acids that have no apparent effect on gene expression or function. In other instances, these variants may have subtle effects on gene expression, thereby conferring the adaptive advantages associated with genetic diversity. On the other hand, allelic variants may reflect mutations in a gene that clearly alter its function. The common Glu ® Val sickle cell mutation (E6V) in the*b-globin* gene and theDF508 deletion of phenylalanine (F) in the *CFTR* gene are examples of allelic variants of these genes. Because each individual has two copies of each chromosome (one inherited from the mother and one inherited from the father), he or she can only have two alleles at a given locus. However, there can be many different alleles in the population. The normal or common allele is usually referred to as *wild type*. When alleles at a given locus are identical, the individual is *homozygous*. Inheriting such identical copies of a mutant allele occurs in many autosomal recessive disorders, particularly in circumstances of consanguinity. If the alleles are different, the individual is *heterozygous* at this locus. If two different mutant alleles are inherited at a given locus, the individual is said to be a *compound heterozygote*. *Hemizygous* is used to describe males with a mutation in an X chromosomal gene, or a female with a loss of one X chromosomal locus.

Genotypes describe the specific alleles at a particular locus. For example, there are three common alleles (E2, E3, E4) of the apolipoprotein E (*APOE*) gene. The genotype of an individual can therefore be described as *APOE3/4* or *APOE4/4* or any other variant. These designations indicate which alleles are present on the two chromosomes in the *APOE* gene at locus 19q13.2. In other cases, the genotype might be assigned arbitrary numbers (e.g., 1/2) or letters (e.g., B/b) to distinguish different alleles.

A *haplotype* refers to a group of alleles that are closely linked together at a genomic locus. Haplotypes are useful for tracking the transmission of genomic segments within families and for detecting evidence of genetic recombination, if the crossover event occurs between the alleles (Fig. 65-3). As an example, various alleles at the histocompatibility locus antigen (HLA) on chromosome 6p are used to establish haplotypes associated with certain disease states. For example, 21-hydroxylase deficiency, complement deficiency, and hemochromatosis are each associated with specific HLA haplotypes. It is now recognized that these genes lie in close vicinity to the HLA locus, which explains why HLA associations were identified even before the disease genes were cloned and localized. In other cases, specific HLA associations with diseases such as ankylosing spondylitis (HLA-B27) or type 1 diabetes mellitus (HLA-DR4) reflect the role of specific HLA allelic variants in susceptibility to these autoimmune diseases.

Allelic Heterogeneity *Allelic heterogeneity* refers to the fact that different mutations in the same genetic locus can cause an identical or similar phenotype. For example, many different mutations of the b-globin locus can cause b-thalassemia (Fig. 65-7). In essence, allelic heterogeneity reflects the fact that many different mutations are capable of altering protein structure and function. For this reason, maps of inactivating mutations in genes usually show a near-random distribution. Exceptions include: (1) a founder effect, in which a particular mutation that does not affect reproductive capacity can be traced to a single individual; (2) "hot spots" for mutations, in which the nature of the DNA sequence predisposes to a recurring mutation; and (3) localization of mutations to certain domains that are particularly critical for protein function. Allelic heterogeneity creates a practical problem for genetic testing because one must often examine the entire genetic locus for mutations, as these can differ in each patient.

Phenotypic Heterogeneity *Phenotypic heterogeneity* occurs when more than one phenotype is caused by allelic mutations (e.g., different mutations in the same gene). For example, mutations in the *myosin VIIIA* gene can result in four distinct clinical disorders: (1) autosomal recessive deafness DFNB2, (2) autosomal dominant nonsyndromic deafness DFNA11, (3) Usher 1B syndrome [congenital deafness, retinitis pigmentosa (Plate IV-14)], and (4) an atypical variant of Usher's syndrome. Similarly, identical mutations in the *FGFR2* gene can result in very distinct phenotypes: Crouzon syndrome (craniofacial synostosis), or Pfeiffer syndrome (acrocephalopolysyndactyly).

Locus or Nonallelic Heterogeneity and Phenocopies *Nonallelic or locus heterogeneity* refers to the situation in which a similar disease phenotype results from mutations at different genetic loci (Table 65-4). This often occurs when more than one gene product produces different subunits of an interacting complex or when different genes are involved in the same genetic cascade or physiologic pathway. For example, osteogenesis imperfecta can arise from mutations in two different procollagen genes (*COL1A1* or *COL1A2*) that are located on different chromosomes (Chap. 351). The effects of inactivating mutations in these two genes are similar because the protein products comprise different subunits of the helical collagen fiber. Similarly, muscular dystrophy syndromes can be caused by mutations in various genes, consistent with the fact that it can be transmitted in an X-linked (Duchenne or Becker), autosomal dominant (limb-girdle muscular dystrophy type 1), or autosomal recessive (limb-girdle muscular dystrophy type 2) manner (Chap. 383). Mutations in the X-linked *DMD* gene, which encodes dystrophin, are the most common cause of muscular dystrophy. This feature reflects the large size of the gene as well as the fact that the phenotype is expressed in hemizygous males because they only have a single copy of the X chromosome. Dystrophin is associated with a large group of additional proteins that form the membrane-associated cytoskeleton in muscle. Mutations in several components of this protein complex can also cause muscular dystrophy syndromes. Although the phenotypic features of some of these disorders are distinct, the phenotypic spectrum caused by mutations in different genes overlaps, thereby leading to nonallelic heterogeneity. It should be noted that mutations in dystrophin also cause allelic heterogeneity. For example, mutations in the *DMD* gene can cause either Duchenne or the less severe Becker muscular dystrophy, depending on the severity of the protein defect.

Recognition of nonallelic heterogeneity is important for several reasons: (1) the ability to identify disease loci in linkage studies is reduced by including patients with similar phenotypes but different genetic disorders; (2) genetic testing is more complex because several different genes need to be considered along with the possibility of different mutations in each of the candidate genes; and (3) novel information is gained about how genes or proteins interact, providing unique insights into molecular physiology.

Phenocopies refer to circumstances in which nongenetic conditions mimic a genetic disorder. For example, features of toxin- or drug-induced neurologic syndromes can resemble those seen in Huntington disease, and vascular causes of dementia share phenotypic features with familial forms of Alzheimer dementia (Chap. 362). Children born with activating mutations of the thyroid-stimulating hormone receptor (TSH-R) exhibit goiter and thyrotoxicosis similar to that seen in neonatal Graves' disease, which is caused by the transfer of maternal autoantibodies to the fetus (Chap. 330). As in nonallelic heterogeneity, the presence of phenocopies has the potential to confound linkage studies and genetic testing. Patient history and subtle differences in phenotype can often provide clues that distinguish these disorders from related genetic conditions.

Variable Expressivity and Incomplete Penetrance It is not uncommon for the same genetic mutation to cause a phenotypic spectrum illustrating the phenomenon of *variable expressivity*. This may include different manifestations of a complex disorder (e.g.,MEN), the severity of the disorder (e.g., sickle cell anemia), or the age of disease onset (e.g., Alzheimer dementia). MEN-1 illustrates several of these features. Families with this autosomal dominant disorder develop tumors of the parathyroid gland, endocrine pancreas, and the pituitary gland (Chap. 339). However, the pattern of tumors in the different glands, the age at which tumors develop, and the types of hormones produced vary among affected individuals, even within a given family. In this example, the phenotypic variability arises, in part, because of the requirement for a second mutation in the normal copy of the *MEN1* gene, as well as the large array of different cell types that are susceptible to the effects of *MEN1* gene mutations. In part, variable expression reflects the influence of other genes, or genetic background, on the effects of a particular mutation. Even in identical twins, in whom the genetic constitution is the same, one can occasionally see variable expression of a genetic disease.

Interactions with the environment can also influence the course of a disease. For example, the manifestations and severity of hemochromatosis can be influenced by iron intake (Chap. 345), and the course of phenylketonuria is affected by exposure to phenylalanine in the diet (Chap. 352). Other metabolic disorders, such as hyperlipidemias and porphyria, also fall into this category. Many mechanisms, including genetic effects and environmental influences, can therefore lead to variable expressivity. In genetic counseling, it is particularly important to recognize this variability, as one cannot always predict the course of disease, even when the mutation is known.

Penetrance is the probability of expressing the phenotype given a defined genotype; it can be complete or incomplete. For example, hypertrophic obstructive cardiomyopathy (HOCM) caused by mutations in the *myosin heavy chain b*gene is a dominant disorder with clinical features in only a subset of patients who carry the mutation (Chap. 238). Patients who have the mutation but no evidence of the disease can still transmit the disorder to subsequent generations. In this situation, the disorder is said to be *nonpenetrant* or *incompletely penetrant*. This classification depends to some degree on the criteria and techniques used for diagnosis. For disorders such as Huntington disease or familial amyotrophic lateral sclerosis, which present late in life, the rate of penetrance is influenced by the age at which the clinical assessment is performed. *Imprinting* can also modify the penetrance of a disease (see below). For example, in patients with Albright hereditary osteodystrophy, mutations in the Gsa subunit (*GNAS1* gene) are expressed clinically only in individuals who inherit the mutation from their mother (Chap. 343).

Sex-Influenced Phenotypes Certain mutations affect males and females quite differently. In some instances, this is because the gene resides on the X or Y sex chromosomes (X-linked disorders and Y-linked disorders). As a result, the phenotype of mutated X-linked genes will be expressed fully in males but variably in heterozygous females, depending on the degree of X-inactivation and the function of the gene. For example, most heterozygous female carriers of factor VIII deficiency (hemophilia A) are asymptomatic because sufficient factor VIII is produced to prevent a defect in coagulation (Chap. 117). On the other hand, some females heterozygous for the X-linked lipid storage defect caused bya-galactosidase A deficiency (Fabry disease) experience mild manifestations of painful neuropathy, as well as other features of the disease (Chap. 349). Because only males have a Y chromosome, mutations in genes such as *SRY* (which causes male-to-female sex-reversal) or *DAZ* (which causes abnormalities of spermatogenesis) are unique to males (Chap. 338).

Other diseases are expressed in a sex-limited manner because of the differential function of the gene product in males and females. Activating mutations in the luteinizing hormone receptor cause dominant male-limited precocious puberty in boys (Chap. 335). The phenotype is unique to males because activation of the receptor induces testosterone production in the testis, whereas it is functionally silent in the immature ovary. Homozygous inactivating mutations of the follicle-stimulating hormone (FSH) receptor cause primary ovarian failure in females because the follicles do not develop in the absence of FSH action. In contrast, affected males have a more subtle phenotype, because testosterone production is preserved (allowing sexual maturation) and spermatogenesis is only partially impaired (Chap. 335). In congenital adrenal hyperplasia, most commonly caused by 21-hydroxylase deficiency, cortisol production is impaired and ACTHstimulation of the adrenal gland leads to increased production of androgenic precursors (Chap. 331). In females, the increased androgen level causes ambiguous genitalia, which can be recognized at the time of birth. In males, the diagnosis may be made on the basis of adrenal insufficiency at birth, because the increased adrenal androgen level does not alter sexual differentiation, or later in childhood, because of the development of precocious puberty. Hemochromatosis is more common in males than in females, presumably because of differences in dietary iron intake and losses associated with menstruation and pregnancy in females (Chap. 345).

GENETIC LINKAGE

Genetic linkage refers to the fact that genes are physically connected, or linked, to one another along the chromosomes. Two fundamental principles are essential for understanding the concept of a genetic linkage: (1) When two genes are close together on a chromosome, they are usually transmitted together, unless a recombination event separates them (Fig. 65-3); and (2) the odds of a crossover, or recombination event, between two linked genes is proportional to the distance that separates them. Thus, genes that are further apart are more likely to undergo a recombination event than genes that are very close together. Linkage is used in genetic counseling to predict the odds of disease gene transmission.

Polymorphisms are essential for linkage studies because they provide a means to distinguish the maternal and paternal chromosomes in an individual. On average, 1 out of every 1000 bp varies from one person to the next. Although this degree of variation seems low (99.9% identical), it means that >3 million sequence differences exist between any two unrelated individuals. This sequence variation usually has no significant functional consequence and provides much of the basis for variation in genetic traits. Although many of these sequence variations areSNPs, other variants includeVNTRs or short tandem repeats (STRs). In VNTRs and STRs, the number of times a sequence is repeated is highly variable in the population. Consequently, the probability that sequences will differ on the two homologous chromosomes is high (often>70 to 90%). Most STRs, also called *polymorphic microsatellite markers*, consist of di-, tri-, or tetranucleotide repeats that can be measured readily usingPCR and primers that reside on either side of the repeat sequences (Fig. 65-8). Many other methods for analyzing polymorphic variation are also available. Historically,RFLPswere used to detect sequence variations that caused changes in the recognition sites for restriction enzymes. This procedure has been largely replaced by the use of STRs. Analyses of SNPs, using DNA chips, provide a promising means for rapid analysis of genetic variation and linkage.

In order to identify a chromosomal locus that segregates with a disease, it is necessary to determine the genotype or haplotype of DNA samples from one or several pedigrees. One can then assess whether certain marker alleles cosegregate with the disease. Markers that are closest to the disease gene are less likely to undergo recombination events and therefore receive a higher linkage score. Linkage is expressed as a lod (logarithm of odds) score -- the ratio of the probability that the disease and marker loci are linked rather than unlinked. Lod scores of +3 (1000:1) are generally accepted as supporting linkage, whereas a score of -2 is consistent with the absence of linkage.

An example of the use of linkage analysis is shown in Fig. 65-8. In this case, the gene for the autosomal dominant disorder,MEN-1, is known to be located on chromosome 11q13. Using positional cloning, the *MEN1* gene was identified and shown to encode menin, the function of which is poorly understood. However, the transmission of the disorder suggests that menin acts like a tumor-suppressor gene. Affected individuals inherit a mutant form of the *MEN1* gene, predisposing them to certain types of tumors (parathyroid, pituitary, pancreatic islet) (Chap. 339). In the tissues that develop a tumor, a "second hit" occurs in the normal copy of the *MEN1* gene. This somatic mutation may be a point mutation, a microdeletion, or loss of a chromosomal fragment (detected as loss of heterozygosity, LOH). Within a given family, linkage to the *MEN1* gene locus can be assessed without necessarily knowing the specific mutation in the *MEN1* gene. Using polymorphicSTRsthat are close to the *MEN1* gene, one can assess transmission of the different *MEN1* alleles and compare this pattern to development of the disorder to determine which allele is associated with risk of MEN-1. In the pedigree shown, the affected grandfather in generation I carries alleles 3 and 4 on the chromosome with the mutated *MEN1* gene and alleles 2 and 2 on his other chromosome 11. Consistent with linkage of the 3/4 genotype to the *MEN1* locus, his son in generation II is affected, whereas his daughter (who inherits the 2/2 genotype from her father) is unaffected. In the third generation, transmission of the 3/4 genotype indicates risk of developing MEN-1, assuming that no genetic recombination between the 3/4 alleles and the *MEN1* gene has occurred. After a specific mutation in the *MEN1* gene is identified within a family, it is possible to track transmission of the mutation itself, thereby eliminating uncertainty caused by recombination.

CHROMOSOMAL DISORDERS

Chromosomal or cytogenetic disorders are caused by numerical or structural aberrations in chromosomes. Deviations in chromosome number are common causes of abortions, developmental disorders, and malformations.**For discussion of disorders of chromosome number and structure, see Chap. 66.*

Contiguous Gene Syndromes Large deletions or duplications may affect a portion of a gene, an entire gene, or, if several genes are involved, cause a *contiguous gene syndrome*. Syndromes associated with chromosomal deletions or duplications have a wide phenotypic spectrum that is dependent on the number of involved gene loci. For example, the cri-du-chat syndrome, one of the most common deletion disorders, is associated with deletions on the short arm of chromosome 5 that vary in size from extremely small deletions within 5p15.2 to the loss of the entire short arm. Because of the variable size of the involved deletions, the phenotype encompasses a spectrum that ranges from severe mental retardation and microcephaly to an isolated catlike cry without morphologic or mental abnormalities.

Contiguous gene syndromes have been useful for identifying the location of new disease-causing genes. Because of the variable size of gene deletions in different patients, a systemic comparison of phenotypes and locations of deletion breakpoints allows positions of particular genes to be mapped within the critical genomic region.

MONOGENIC MENDELIAN DISORDERS

Monogenic human diseases are frequently referred to as *Mendelian disorders* because they obey the principles of genetic transmission originally set forth in Gregor Mendel's classic work. The mode of inheritance for a given phenotypic trait or disease is determined by pedigree analysis. All affected and unaffected individuals in the family are recorded in a pedigree using standard symbols (Fig. 65-9). The principles of allelic segregation, and the transmission of alleles from parents to children, are illustrated inFig. 65-10. One dominant (A) allele and one recessive (a) allele can display three Mendelian modes of inheritance: autosomal dominant, autosomal recessive, and X-chromosomal. About 65% of human monogenic disorders are autosomal dominant, 25% are autosomal recessive, and 5% are X-linked (Table 65-5). Genetic testing is now available for many of these disorders and plays an increasingly important role in clinical medicine.

Autosomal Dominant Disorders Autosomal dominant disorders assume particular relevance because mutations in a single allele are sufficient to cause the disease. In contrast to recessive disorders, in which disease pathogenesis is relatively straightforward because there is loss of gene function, in dominant disorders there are various disease mechanisms, many of which are unique to the function of the genetic pathway involved.

In autosomal dominant disorders, individuals are affected in successive generations; the disease does not occur in the offspring of unaffected individuals. Males and females are affected with equal frequency because the defective gene resides on one of the 22 autosomes (Fig. 65-11*A*). Autosomal dominant mutations alter one of the two alleles at a given locus. Because the alleles segregate randomly at meiosis, the probability that an offspring will be affected is 50%. Unless there is a new germline mutation, an affected individual has an affected parent. Children with a normal genotype do not transmit the disorder. Due to differences in penetrance or expressivity (see above), the clinical manifestations of autosomal dominant disorders may be variable. Because of these variations, it is sometimes challenging to determine the pattern of inheritance.

It should be recognized, however, that some individuals acquire a mutated gene from an unaffected parent. De novo germline mutations occur more frequently during later cell divisions in gametogenesis, explaining why siblings are rarely affected. As noted before, new germline mutations occur more frequently in fathers of advanced age. For example, the average age of fathers with new germline mutations that cause Marfan's syndrome is approximately 37 years, whereas fathers who transmit the disease by inheritance have an average age of about 30 years.

Autosomal Recessive Disorders The clinical expression of autosomal recessive disorders is more uniform than in autosomal dominant disorders. Most mutated alleles lead to a complete or partial loss of function. They frequently involve enzymes in metabolic pathways, receptors, or proteins in signaling cascades. Though most recessive disorders are rare, the relatively high frequency of certain recessive disorders, such as sickle cell anemia, cystic fibrosis, and thalassemia, is partially explained by a selective biologic advantage for the heterozygous state (see below).

In an autosomal recessive disease, the affected individual, who can be of either sex, is a homozygote or compound heterozygote for a single-gene defect. With a few important exceptions, autosomal recessive diseases are rare and often occur in the context of parental consanguinity. Though heterozygous carriers of a defective allele are usually clinically normal, they may display subtle differences in phenotype that only become apparent with more precise testing or in the context of certain environmental influences. In sickle cell anemia, for example, heterozygotes are normally asymptomatic. However, in situations of dehydration or diminished oxygen pressure, sickle cell crises can also occur in heterozygotes (Chap. 106).

In most instances, an affected individual is the offspring of heterozygous parents. In this situation, there is a 25% chance that the offspring will have a normal genotype, a 50% probability of a heterozygous state, and a 25% risk of homozygosity for the recessive alleles (Fig. 65-11*B*). In the case of one unaffected heterozygous and one affected homozygous parent, the probability of disease increases to 50% for each child. In this instance, the pedigree analysis mimics an autosomal dominant mode of inheritance (*pseudodominance*). In contrast to autosomal dominant disorders, new mutations in recessive alleles are rarely manifest because they usually result in an asymptomatic carrier state.

X-Linked Disorders Males have only one X chromosome; consequently, a daughter always inherits her father's X chromosome in addition to one of her mother's two X chromosomes. A son inherits the Y chromosome from his father and one maternal X chromosome. Thus, the characteristic features of X-linked inheritance are (1) the absence of father-to-son transmission, and (2) the fact that all daughters of an affected male are obligate carriers of the mutant allele (Fig. 65-11*C*). The risk of developing disease due to a mutant X-chromosomal gene differs in the two sexes. Because males have only one X chromosome, they are hemizygous for the mutant allele; thus, they are more likely to develop the mutant phenotype, regardless of whether the mutation is dominant or recessive. A female may be either heterozygous or homozygous for the mutant allele, which may be dominant or recessive. The terms *X-linked dominant* or *X-linked recessive* are therefore only applicable to expression of the mutant phenotype in women. In addition, the expression of X-chromosomal genes is influenced by X chromosome inactivation (see below).

Y-Linked Disorders Only a few genes are known on the Y chromosome. One such gene, the sex-region determining Y factor (*SRY*), or testis-determining factor (*TDF*), is crucial for normal male development. Normally there is infrequent exchange of sequences on the Y chromosome with the X chromosome. Because the *SRY* region is closely adjacent to the pseudoautosomal region, a chromosomal segment on the X and Y chromosomes with a high degree of homology, a crossing-over occasionally involves the *SRY* region. Translocations can result in XY females with the Y chromosome lacking the *SRY* gene or XX males harboring the *SRY* gene on one of the X chromosomes (Chap. 338). Point mutations in the *SRY* gene may also result in individuals with an XY genotype and an incomplete female phenotype. Most of these mutations occur de novo. Men with oligospermia/azoospermia frequently have microdeletions on the long arm of the Y chromosome that involve one or more of the azoospermia factor (*AZF*) genes.

EXCEPTIONS TO SIMPLE MENDELIAN INHERITANCE PATTERNS

Mitochondrial Disorders Each mitochondrion contains several copies of a circular chromosome. Mitochondrial DNA predominantly encodes transfer RNAs and proteins that are components of the respiratory chain involved in oxidative phosphorylation and ATP generation. The mitochondrial genome is inherited through the maternal line because sperm does not contribute significant cytoplasmic components to the zygote. All children from an affected mother will inherit the disease, but it will not be transmitted from an affected father to his children. During cell replication, the proportion of wild-type and mutant mitochondria can drift; differences in the fraction of defective mitochondria are referred to as *heteroplasmia* and explain, in part, the phenotypic variability that is common in mitochondrial diseases.**For detailed discussion of mitochondrial disorders, see Chap. 67.*

Mosaicism Mosaicism refers to the presence of two or more genetically distinct cell lines in the tissues of an individual. It results from a mutation that occurs during embryonic, fetal, or extrauterine development. The developmental stage at which the mutation arises will determine whether germ cells and/or somatic cells are involved. Chromosomal mosaicism results from non-disjunction at an early embryonic mitotic division, leading to the persistence of more than one cell line, as exemplified by some patients with Turner syndrome (Chap. 338). Somatic mosaicism is characterized by a patchy distribution of genetically altered somatic cells. The McCune-Albright syndrome, for example, is caused by activating mutations in the stimulatory G protein a(G_sa) that occur early in development (Chap. 343). The clinical phenotype varies depending on the tissue distribution of the mutation; manifestations include ovarian cysts that secrete sex steroids and cause precocious puberty, polyostotic fibrous dysplasia, cafe-au-lait skin pigmentation, growth hormone-secreting pituitary adenomas, and hypersecreting autonomous thyroid nodules (Chap. 336).

X-Inactivation, Imprinting, and Uniparental Disomy According to traditional Mendelian principles, the parental origin of a mutant gene is irrelevant for the expression of the phenotype. Nonetheless, there are important exceptions to this rule. X-inactivation prevents the expression of most genes on one of the two X-chromosomes in every cell of a female. Gene inactivation also occurs on selected chromosomal regions of autosomes. This phenomenon, referred to as *genomic imprinting*, leads to preferential expression of an allele depending on its parental origin. It is of pathophysiologic importance in disorders where the transmission of disease is dependent on the sex of the transmitting parent and, thus, plays an important role in the expression of certain genetic disorders. Two classic examples are the Prader-Willi syndrome and Angelman syndrome (Chap. 66). Prader-Willi syndrome is characterized by diminished fetal activity, obesity, hypotonia, mental retardation, short stature, and hypogonadotropic hypogonadism. Deletions in the Prader-Willi syndrome occur exclusively on the paternal chromosome 15. In contrast, patients with Angelman syndrome, characterized by mental retardation, seizures, ataxia, and hypotonia, have deletions at the same site of chromosome 15; however, they are located on the maternal chromosome 15. These two syndromes may also result from *uniparental disomy*. In this case, the syndromes are not caused by deletions on chromosome 15 but by the inheritance of either two paternal chromosomes (Prader-Willi syndrome), or two maternal chromosomes (Angelman syndrome).

Imprinting and the related phenomenon of allelic exclusion may be more common than currently documented, as it is difficult to examine levels of mRNA expression from the maternal and paternal alleles in specific tissues or in individual cells. Genomic imprinting, or uniparental disomy, is involved in the pathogenesis of several other disorders and malignancies (Chap. 66). Hydatidiform mole contains a normal number of diploid chromosomes, but they are all of paternal origin. The opposite situation occurs in ovarian teratomata, with 46 chromosomes of maternal origin. Expression of the imprinted gene for insulin-like growth factor II (IGF-II) is involved in the pathogenesis of the cancer-predisposing Beckwith-Wiedemann syndrome (BWS) (Chap. 81). These children show somatic overgrowth with organomegalies and hemihypertrophy, and they have an increased risk of embryonal malignancies such as Wilm's tumor. Normally only the paternally derived copy of the *IGF-II* gene is active and the maternal copy is inactive. Imprinting of the *IGF-II* gene is regulated by *H19*, which encodes an RNA transcript that is not translated into protein. Disruption or lack of *H19* methylation leads to a relaxation of *IGF-II* imprinting and expression of both alleles. Heritable changes in gene expression not associated with DNA sequence alterations are referred to as *epigenetic effects*; these changes are increasingly recognized to play a role in human diseases and possibly in aging as well (Chap. 9).

Somatic Mutations In many cancer syndromes, there is an inherited predisposition to tumor formation. However, the neoplastic process requires the acquisition of additional somatic mutations (Chap. 81). In retinoblastoma, the tumor develops when both copies of the retinoblastoma (*RB*) gene are inactivated through two somatic events (sporadic retinoblastoma) or through a somatic loss of the normal allele in an individual with a hereditary defect in the other allele (hereditary retinoblastoma). This "two-hit" model applies to other inherited cancer syndromes such asMEN-1 (Chap. 339) and neurofibromatosis type 2 (Chap. 370). The defective allele is transmitted in a dominant pattern, though tumorigenesis results from a recessive loss of the tumor suppressor gene in an affected tissue. In other instances, the development of cancer typically requires somatic defects in multiple genes, a process termed *multistep carcinogenesis* (Chap. 82).

Nucleotide Repeat Expansion Disorders Several diseases are associated with an increase in the number of nucleotide repeats above a certain threshold (Table 65-6). The repeats are sometimes located within the coding region of the genes, as in Huntington disease or the X-linked form of spinal and bulbar muscular atrophy (SBMA, Kennedy syndrome). In other instances, the repeats probably alter gene regulatory sequences. If an expansion is present, the DNA fragment is unstable and tends to expand further during cell division. The length of the nucleotide repeat often correlates with the severity of the disease. When repeat length increases from one generation to the next, disease manifestations may worsen or be observed at an earlier age; this phenomenon is referred to as *anticipation*. In Huntington disease, for example, there is a correlation between age of onset and length of the triplet codon expansion (Chap. 362). Anticipation has also been documented in other diseases caused by dynamic mutations in trinucleotide repeats (Table 65-6). The repeat number may also vary in a tissue-specific manner. In myotonic dystrophy, the CTG repeat may be tenfold greater in muscle tissue than in lymphocytes (Chap. 383).

POPULATION GENETICS AND ASSOCIATION STUDIES

Overview of Population Genetics In population genetics, the focus changes from alterations in an individual's genome to the distribution pattern of different genotypes of alleles in the population. In a case where there are only two alleles, A and a, the frequency of the genotypes will be p_2+ 2*pq* +q_2= 1, with $p_2$corresponding to the frequency of AA, 2*pq* to the frequency of Aa, and $q_2$to aa. When the frequency of an allele is known, the frequency of the genotype can be calculated. Alternatively, one can determine an allele frequency, if the genotype frequency has been determined.

Allele frequencies vary among ethnic groups and geographical regions. For example, heterozygous mutations in the *CFTR* gene are relatively common in populations of European origin but are rare in the African population. Allele frequencies may vary because certain allelic variants confer a selective advantage. For example, heterozygotes for the sickle cell mutation, which is particularly common in West Africa, are more resistant to malarial infection because the erythrocytes of heterozygotes provide a less favorable environment for *Plasmodium* parasites. Though homozygosity for the sickle cell gene is associated with severe anemia and sickle crises (Chap. 106), heterozygotes have a higher probability of survival because of the reduced morbidity and mortality from malaria; this phenomenon has led to an increased frequency of the mutant allele. Recessive conditions are more prevalent in geographically isolated populations because of the more restricted gene pool.

Allelic Association and Linkage Disequilibrium There are two primary strategies for mapping genes that cause or increase susceptibility to human disease: (1) classic linkage can be performed based on a known genetic model (see above) or, when the model is unknown, by studying pairs of affected relatives; or (2) disease genes can be mapped using allelic association studies (Table 65-7). *Allelic association* refers to a situation in which the frequency of an allele is significantly increased or decreased in a particular disease. Linkage and association differ in several aspects. Genetic linkage is demonstrable in families or sibships. Association studies, on the other hand, compare a population of affected individuals with a control population. Association studies can be performed as case-control studies that include unrelated affected individuals and matched controls, or as family-based studies that compare the frequencies of alleles transmitted or not transmitted to affected children.

Allelic association studies are particularly useful for identifying susceptibility genes in complex diseases. When alleles at two loci occur more frequently in combination than would be predicted (based on known allele frequencies and recombination fractions), they are said to be in *linkage disequilibrium*. In Fig. 65-12, a mutation, Z, has occurred at a susceptibility locus where the normal allele is Y. The mutation is in close proximity to a genetic polymorphism with allele A or B. With time, the chromosomes carrying the A and Z alleles accumulate and represent 10% of the chromosomes in the population. The fact that the disease susceptibility gene, Z, is found preferentially, or exclusively, in association with the A allele illustrates linkage disequilibrium. Though not all chromosomes carrying the A allele carry the disease gene, the A allele is associated with an increased risk because of its possible association with the Z allele. This model implies that it may be possible in the future to identify Z directly to provide a more accurate prediction of disease susceptibility. Evidence for linkage disequilibrium can be helpful in mapping disease genes because it suggests that the two loci, in this case A and Z, are tightly linked.

POLYGENIC DISEASE AND COMPLEX GENETIC TRAITS

Approach to Polygenic and Multifactorial Disease The expression of many common diseases such as cardiovascular disease, hypertension, diabetes, asthma, psychiatric disorders, and certain cancers is determined by genetic background, environmental factors, and lifestyle (Table 65-8). A trait is called *polygenic* if multiple genes are thought to contribute to the phenotype or *multifactorial* if multiple genes are assumed to interact with environmental factors. Genetic models for complex traits need to account for genetic heterogeneity and interactions with other genes and the environment. Complex genetic traits may be influenced by modifying genes that are not linked to the main gene involved in the pathogenesis of the trait. This type of gene-gene interaction, or *epistasis*, plays an important role in polygenic traits that require the simultaneous presence of variations in multiple genes in order to result in a pathologic phenotype.
Gene-environment interactions are relevant for many monogenic and polygenic disorders. In phenylketonuria, the phenotypic expression of the disease depends not only on the presence of the mutation in the phenylalanine hydroxylase gene but also on the exposure to the amino acid phenylalanine (Chap. 352). Another example is type 2 diabetes mellitus, in which genetic, nutritional, and lifestyle factors are intimately interrelated in disease pathogenesis (Chap. 333). The identification of genetic variations and environmental factors that either predispose or protect against disease is essential for predicting disease risk, designing preventive strategies, and developing novel therapeutic approaches (Chap. 68). The study of rare monogenic diseases may provide insights into genetic and molecular mechanisms that are subsequently of importance for the understanding of complex diseases. For example, the identification of the insulin promoter factor 1 in maturity-onset of diabetes type 4 was followed by the observation that it also plays a role in the pathogenesis of diabetes mellitus type 2 (Tables 65-1 and65-8).

Approach to the Patient

Identifying the Disease-Causing Gene *Genomic medicine* aims to enhance the quality of medical care through the use genotypic analysis (DNA testing) to identify genetic predisposition to disease, to select more specific pharmacotherapy, and to design individualized medical care based on genotype. Genotype can be deduced by analysis of protein (e.g., hemoglobin, apoprotein E), mRNA, or DNA. However, technological advances have made DNA analysis particularly useful because it can be readily applied to all but the largest genes (Fig. 65-13).

DNA testing is performed by mutational analysis or linkage studies in individuals at risk for a genetic disorder known to be present in a family. Mass screening programs require tests of high sensitivity and specificity to be cost-effective. Prerequisites for the success of genetic screening programs include the following: that the disorder is potentially serious; that it can be influenced at a presymptomatic stage by changes in behavior, diet, and/or pharmaceutical manipulations; and that the screening does not result in any harm or discrimination. Screening in Jewish populations for the autosomal recessive neurodegenerative storage disease Tay-Sachs has reduced the number of affected individuals. In contrast, screening for sickle cell trait/disease in African Americans has led to unanticipated problems of discrimination by health insurers and employers. Mass screening programs harbor additional potential problems. For example, screening for the most common genetic alteration in cystic fibrosis, the DF508 mutation with a frequency of ~70% in northern Europe, is feasible and seems to be effective. One has to keep in mind, however, that there is pronounced allelic heterogeneity and that the disease can be caused by >600 other mutations. The search for these less common mutations would substantially increase costs but not the effectiveness of the screening program as a whole. Occupational screening programs aim to detect individuals with increased risk for certain professional activities (e.g.,$a_1$antitrypsin deficiency and smoke or dust exposure).

MUTATIONAL ANALYSES DNA sequence analysis is increasingly used as a diagnostic tool and significantly enhanced diagnostic accuracy. It is used for determining carrier status and for prenatal testing in monogenic disorders (Table 65-5). Certain cancer susceptibility genes, such as *BRCA1* and *BRCA2*, may identify individuals with an increased risk for the development of malignancies. The detection of mutations is an important diagnostic and prognostic tool in leukemias and lymphomas. The demonstration of the presence or absence of mutations is also relevant for the rapidly evolving field of pharmacogenetics, including the identification of differences in drug treatment response or metabolism as a function of genetic background.

A general algorithm for the approach to mutational analysis is outlined inFig. 65-13. The importance of a detailed clinical phenotype cannot be overemphasized. This is the step where one should also consider the possibility of genetic heterogeneity and phenocopies. If obvious candidate genes are suggested by the phenotype, they can be analyzed directly. After identification of a mutation, it is essential to demonstrate that it segregates with the phenotype. The functional characterization of novel mutations is labor-intensive and may require analyses in vitro or in transgenic models in order to document the relevance of the genetic alteration.

Numerous techniques are available for the detection of mutations (Table 65-9). In a very broad sense, one can distinguish between techniques that allow for screening the absence or presence of known mutations (screening mode) or techniques that definitively characterize mutations. Analyses of large alterations in the genome are possible using cytogenetics, fluorescent in situ hybridization (FISH), and Southern blotting (Chap. 66).

More discrete sequence alterations rely heavily on the use of thePCR, which allows rapid gene amplification and analysis. Moreover, PCR makes it possible to perform genetic testing and mutational analysis with small amounts of DNA extracted from leukocytes or even from single cells, buccal cells, or hair roots. Screening for point mutations can be performed by numerous methods (Table 65-9); most are based on the recognition of mismatches between nucleic acid duplexes, electrophoretic separation of single- or double-stranded DNA, or sequencing of DNA fragments amplified by PCR. DNA sequencing can be performed directly on PCR products or on fragments cloned into plasmid vectors amplified in bacterial host cells.

RT-PCRmay be useful to detect absent or reduced levels of mRNA expression due to a mutated allele. Protein truncation tests (PTT) can be used to detect the broad array of mutations that result in premature termination of a polypeptide during its synthesis. The isolated cDNA is transcribed and translated in vitro, and the proteins are analyzed by gel electrophoresis. Comparison of electrophoretic mobility with the wild-type protein allows detection of truncated mutants.

The majority of traditional diagnostic methods are gel-based. Novel technologies for the analysis of mutations, genetic mapping, and mRNA expression profiles are in rapid development. DNA chip technologies allow hybridization of DNA or RNA to hundreds of thousands of probes simultaneously. Microarrays are being used clinically for mutational analysis of several human disease genes, as well as for the identification of viral sequence variations. Together with the knowledge gained from theHGP, these technologies provide the foundation to expand from a focus on single genes to analyses at the scale of the genome.

ACKNOWLEDGEMENT

This chapter reflects the cumulative contributions of many past contributors to Harrison's Principles of Internal Medicine. Most recently, this includes Dr. Joseph L. Goldstein, Dr. Michael S. Brown, Dr. Andrea Ballabio, and Dr. Arthur L. Beaudet.

(Bibliography omitted in Palm version)

Back to Table of Contents

66. CHROMOSOME DISORDERS - *Terry Hassold*, *Stuart Schwartz*

In humans, the normal diploid number of chromosomes is 46, consisting of 22 pairs of autosomal chromosomes (numbered 1 to 22 in decreasing size) and one pair of sex chromosomes (XX in females and XY in males). The genome is estimated to contain between 80,000 and 100,000 genes, with the smallest autosome housing between 500 and 1000 genes. Not surprisingly, duplications or deletions of even small chromosome segments have profound consequences on normal gene expression.

Deviations in the number or structure of the 46 human chromosomes are astonishingly common, despite severe deleterious consequences. Chromosomal disorders occur in an estimated 10 to 25% of all pregnancies. They are the leading cause of fetal loss and, among pregnancies surviving to term, the leading known cause of birth defects and mental retardation.

In recent years, the practice of cytogenetics has shifted from conventional cytogenetic methodology to a union of cytogenetic and molecular techniques. Formerly the province of research laboratories, *fluorescence in situ hybridization* (FISH) and related molecular cytogenetic technologies have been incorporated into everyday practice in clinical laboratories. As a result, there is an increased appreciation of the importance of "subtle" constitutional cytogenetic abnormalities, such as microdeletions and imprinting disorders, as well as previously recognized translocations and disorders of chromosome number.

VISUALIZING CHROMOSOMES

CONVENTIONAL CYTOGENETIC ANALYSIS

In theory, chromosome preparations can be obtained from any actively dividing tissue by causing the cells to arrest in metaphase, the stage of the cell cycle at which chromosomes are maximally condensed. In practice, only a small number of tissues are used for routine chromosome analysis: amniocytes or chorionic villi for prenatal testing; and blood, bone marrow, or skin fibroblasts for postnatal studies. Samples of blood, bone marrow, and chorionic villi can be processed using short-term culture techniques that yield results in 1 to 3 days. Analysis of other tissue types typically involves long-term tissue culture, requiring 1 to 3 weeks of processing before cytogenetic analysis is possible.

Regardless of the culturing technique, cells are processed to recover chromosomes at metaphase or prometaphase and treated chemically or enzymatically to reveal chromosome "bands" (Fig. 66-1). Analysis of the number of chromosomes in the cell, and the distribution of bands on individual chromosomes, allows the identification of numerical or structural abnormalities. This strategy is useful for characterizing the normal chromosome complement and determining the incidence and types of major chromosome abnormalities.

Chromosomes are complex structures, consisting of the DNA double helix and chromosome-associated proteins. As for virtually all organisms, each human chromosome contains two specialized structures: a centromere and two telomeres. The *centromere*, or primary constriction, divides the chromosome into short (p) and long (q) arms and is responsible for the segregation of chromosomes during cell division. The *telomeres*, or chromosome ends, "cap" the p and q arms and are important for allowing DNA replication at the ends of the chromosomes. Prior to DNA replication, each chromosome consists of a single chromatid copy of the DNA double helix. After DNA replication and continuing until the time of cell division (including metaphase, when chromosomes are typically visualized), each chromosome consists of two identical sister chromatids (Fig. 66-1).

MOLECULAR CYTOGENETICS

The introduction ofFISHmethodologies in the late 1980s revolutionized the field of cytogenetics. In principle, FISH is similar to other DNA-DNA hybridization methodologies. The labeled probe DNA and the target DNA (usually metaphase chromosomes) are denatured to become single-stranded and are hybridized together. The probe is labeled with a hapten, such as biotin or digoxigenin, to allow detection with a fluorophore (e.g., FITC or rhodamine). Alternatively, many probes are already labeled with fluorophores and thus can be detected directly. After the hybridization step, the specimen is counter-stained and the preparations are visualized with a fluorescence microscope.

Types of FISH Probes A variety of probes are available for use with FISH, including chromosome-specific paints (chromosome libraries), repetitive probes, and single-copy probes. Chromosome libraries were developed initially from flow-sorted individual chromosomes and more recently from monochromosomal human-rodent hybrids. These probes hybridize to sequences that span the entirety of the chromosome from which they are derived and, as a result, they can be used to "paint" individual chromosomes (Fig. 66-2).

Repetitive probes recognize amplified DNA sequences present in chromosomes. The most common are a-satellite DNA probes that are complementary to DNA sequences found at the centromeric regions of all human chromosomes. There are also a-satellite probes that hybridize to the centromeric regions of specific chromosomes (Fig. 66-2).

A vast number of *single-copy probes* are now available, both commercially and as a result of the human genome project. These probes can be as small as 1 kb, though normally they are packaged in cosmids (40 kb), bacterial artificial chromosomes (BACs) or P1 clones (100 to 200 kb), or yeast artificial chromosomes (YACs) (1 to 2 Mb). With the advent of the National Cancer Institute BAC initiative, these large DNA fragments will be placed at 1-Mb intervals on every chromosome, each of which can be used forFISHhybridization. Probes for a variety of microdeletion syndromes and for subtelomeric regions of individual chromosomes are commercially available (Fig. 66-2).

Applications of FISH The majority of FISH applications involve hybridization of one or two probes of interest as an adjunctive procedure to conventional chromosomal banding techniques. In this regard, FISH can be utilized to identify specific chromosomes, characterize de novo duplications or deletions, and identify and clarify subtle chromosomal rearrangements. Its greatest utilization, however, is in the detection of microdeletions (see below), including those associated with Prader-Willi syndrome

(PWS), Angelman syndrome (AS), William syndrome, velocardiofacial (VCF) and DiGeorge syndromes, Smith-Magenis syndrome, and Miller-Dieker syndrome (MDS) (see below). Though conventional cytogenetic studies can detect some of these microdeletions, initial detection and/or confirmation with FISH is essential. In fact, since appropriate FISH probes have become available, detection of the aforementioned syndromes has increased significantly.

In addition to metaphaseFISH, cells can be analyzed at a variety of stages. *Interphase analysis*, for example, can be used to make a rapid diagnosis in instances where metaphase chromosome preparations are not yet available (e.g., amniotic fluid interphase analysis). Interphase analysis also increases the number of cells available for examination, allows for investigation of nuclear organization, and provides results when cells do not progress to metaphase. One specialized type of interphase analysis involves the application of FISH to paraffin-embedded sections, thereby preserving the architecture of the tissue.

The use of interphaseFISH has increased recently, especially for analyses of amniocentesis samples. These studies are performed on uncultured amniotic fluid, typically using DNA probes specific for the chromosomes most commonly identified in trisomies (chromosomes 13, 18, 21, and the X and Y). These studies can be performed rapidly (24 to 72 h) and will ascertain about 60% of the abnormalities detected prenatally. Nevertheless, guidelines from the American College of Medical Genetics suggest that standard cytogenetic analysis be conducted on all specimens following FISH analysis.

Another area in which interphase analysis is routinely utilized is cancer cytogenetics (Chap. 81). Many site-specific translocations are associated with specific types of malignancies. For example, there are probes available for both the Abelson (Abl) oncogene and breakpoint cluster region (bcr) involved in chronic myelogenous leukemia (CML). These probes are labeled with rhodamine and FITC, respectively; the fusion of these genes in CML combines the fluorescent colors and appears as a yellow hybridization signal.

In addition to standard metaphase and interphaseFISHanalyses, a number of enhanced techniques have been developed for specific types of analysis, including multicolor FISH techniques, reverse painting, comparative genomic hybridization, and fiber FISH. *Spectral karyotyping* (SKY) and*multicolor FISH* (m-FISH) techniques use combinatorially labeled probes that create a unique color for individual chromosomes. In this manner, all of the chromosomes are studied simultaneously, and computer software is used to generate "pseudo-colors" for the individual chromosomes. This technology is useful in the identification of unknown chromosome material (such as markers of duplications) but is most commonly used with the complex rearrangements seen in cancer specimens.

Reverse painting is accomplished by either flow-sorting a chromosome of interest or scraping the chromosome off a slide. The DNA from this chromosome (or portion of a chromosome) is extracted, amplified, labeled, and used as aFISHprobe. This probe is then hybridized to a normal metaphase chromosome to identify the origin of the DNA of interest. It is also utilized to identify marker chromosomes or chromosome duplications of unknown origin.

Comparative genomic hybridization (CGH) is a method that can be used when only DNA is available from a specimen of interest. The entire DNA specimen from the sample of interest is labeled in one color (e.g., green), and the normal control DNA specimen in another color (e.g., red). These are mixed in equal amounts and hybridized to normal metaphase chromosomes. The red-to-green ratio is analyzed by a computer program, which determines where the DNA of interest may have gains or loss of material. This technique is useful in the analysis of tumors, particularly in those cases where cytogenetic analysis is not possible.

Fiber FISH is a technique in which chromosomes are mechanically stretched, using one of a variety of different methods. FiberFISHprovides a higher resolution of analysis than conventional FISH and more precise information on the chromosomal localization of a specific probe.

CYTOGENETIC TESTING IN PRENATAL DIAGNOSIS (See also Chap. 68)

The vast majority of prenatal diagnostic studies are performed to rule out a chromosomal abnormality, but cells may also be propagated for biochemical studies or molecular analyses of DNA. Three procedures are used to obtain samples for prenatal diagnosis: amniocentesis, chorionic villus sampling (CVS), and fetal blood sampling. *Amniocentesis* is the most commonly used procedure and is routinely performed at 15 to 17 weeks of gestation. On some occasions, early amniocentesis at 12 to 14 weeks is done to expedite results, though less fluid is obtained at this time. Early amniocentesis carries a greater risk of spontaneous abortion or fetal injury but provides results at an earlier stage of pregnancy.

The vast majority of amniocentesis are performed in the context of advanced maternal age, the best-known correlate of trisomy (see below). Additional reasons for referral for amniocentesis include an abnormal "triple-marker assay" and/or detection of ultrasound abnormalities. In the triple-marker assay, levels of human chorionic gonadotropin, a fetoprotein, and unconjugated estriol in the maternal serum are quantified and used to adjust the maternal age-predicted risk of a trisomy 21 or trisomy 18 fetus. Specific ultrasound abnormalities, when detected at mid-trimester, can also be associated with chromosomal defects. When a nonspecific ultrasound abnormality is present, the estimated risk of a chromosomal defect is approximately 16%. Associations of chromosomal abnormalities and specific types of abnormal ultrasound findings are listed in Table 66-1.

Chorionic villus sampling is the second most common procedure for genetic prenatal diagnosis. Because this procedure is routinely performed at about 8 to 10 weeks of gestation, it allows for an earlier detection of abnormalities and a safer pregnancy termination, if desired.CVS is a relatively safe procedure (spontaneous abortions <0.5 to 1%). Because there is an increased association of limb defects when the procedure is performed later (³11 weeks of gestation), CVS is applicable during a very narrow window of time of gestation. CVS involves the use of a catheter inserted transvaginally; approximately 25 mg of villi are aspirated from the chorion frondosum (the fetal portion of the placenta). Care must be taken not to obtain villi from the maternal portion from the placenta to avoid compromising the analysis. The majority of the sample (the mesenchymal cells from the CVS sample) is enzymatically digested and cultured in a fashion similar to amniotic fluid cells. However, cells in the outer layer of the villi -- the cytotrophoblasts -- are actively dividing and can be analyzed directly. Therefore, by adding colchicine directly to these cells, a result can be obtained within 24 to 48 h. Findings from these procedures should be confirmed by analyses of cultured mesenchymal cells, as they are more reliably derived from the fetus.

Percutaneous umbilical blood sampling (PUBS) is a method for obtaining fetal blood during the second and third trimesters of pregnancy. It allows for acquisition of a blood sample, which can be used for cytogenetic studies; results can be obtained within 48 h of sampling. PUBS is carried out under ultrasound guidance. It is usually performed when ultrasound abnormalities are detected late in the second trimester. PUBS is also used when cytogenetic results from amniocentesis need clarification, such as the detection of mosaicism.

CHROMOSOME ABNORMALITIES

CHROMOSOMES IN CELL DIVISION

To understand the etiology of chromosome abnormalities, it is important to review the movement of chromosomes during cell division. In somatic tissues, chromosomes are replicated during the S-phase of the cell cycle, so that each replicated chromosome consists of two identical sister chromatids (Fig. 66-1). When the cell enters mitosis, each of the 46 chromosomes align on the metaphase plate, with the centomeres cooriented toward opposite spindle poles (Fig. 66-3). At anaphase the sister chromatids separate, with each of the daughter cells receiving one sister chromatid from each of the 46 chromosomes.

Chromosome segregation is more complicated in germ cell division, since the number of chromosomes must be reduced from 46 to 23 in the mature sperm and eggs. This is accomplished by two rounds of division -- meiosis I and meiosis II (Fig. 66-3). In meiosis I, homologous chromosomes become paired and exchange genetic material, then align on the metaphase plate, and finally separate from one another. Thus, by the end of meiosis I, only 23 of the original 46 chromosomes are represented in each of the two daughter cells. Meiosis II quickly follows meiosis I and is essentially a "haploid mitosis," involving separation of the sister chromatids in each of the 23 chromosomes.

Although the fundamentals of meiosis are the same in males and females, there are important distinctions, particularly in the timing of the meiotic divisions. In males, meiosis begins with puberty and continues throughout the individual's lifetime. In females, meiosis begins prenatally, with oocytes proceeding through the first stages of meiosis I but arresting at mid-prophase. At the time of birth, the first meiotic division is suspended in oocytes. Only after ovulation many years later do oocytes complete meiosis I and proceed to the metaphase stage of meiosis II; if fertilized, the oocyte then completes the second meiotic division. Thus, in females, the first meiotic division takes at least 10 to 15 years, and possibly as many as 40 to 45 years, to complete. Maternal age-related increases in the incidence of trisomy are likely the consequence of this protracted process of cell division.

INCIDENCE AND TYPES OF CHROMOSOME ABNORMALITIES

Errors in meiosis, or in early cleavage divisions, occur with extraordinary frequency. At least 10 to 25% of all pregnancies, for example, involve chromosomally abnormal conceptions. A large proportion of these terminate in the earliest stages of pregnancy. Nevertheless, even among clinically recognized pregnancies, nearly 10% of fetuses are chromosomally unbalanced. The occurrence of different classes of chromosome abnormalities are summarized inTable 66-2 for the three types of clinically recognized pregnancies: spontaneous abortions, stillbirths, and livebirths. The commonest abnormalities are numerical, involving fetuses with additional (*trisomy*) or missing (*monosomy*) chromosomes or those with one (*triploidy*) or two (*tetraploidy*) additional sets of chromosomes. Structural chromosome abnormalities are much less common, although several of the most important clinical chromosomal disorders involve structural rearrangements (see below).

By far the most common abnormality is trisomy, which is identified in approximately 25% of spontaneous abortions and 0.3% of newborns. Trisomies for all chromosomes have now been identified in embryos or fetuses, but there is considerable variation in frequency for various chromosomes. For example, trisomy 16 is extraordinarily common, accounting for about one-third of all trisomies in spontaneous abortions, whereas trisomies 1, 5, 11, and 19 have been identified less often. Available evidence suggests two reasons for this variation: (1) some chromosomes (e.g., chromosome 16) are more likely to segregate abnormally or undergo nondisjunction during meiosis than are others; and (2) the potential for development varies widely among different trisomic conditions, with some being eliminated very early in gestation, others surviving to the time of clinical pregnancy recognition, and some (e.g., trisomies 13, 18, and 21 and sex chromosome trisomies) being compatible with survival to term.

CHROMOSOMAL SYNDROMES

While most chromosomally abnormal conceptions perish in utero (Table 66-2), several conditions are compatible with survival to term. The best-characterized of these are numerical abnormalities involving loss or gain of individual chromosomes and abnormalities resulting from unbalanced translocations.FISH and other molecular studies have led to the identification of two "new" types of chromosome abnormalities, commonly referred to as *microdeletion syndromes* and *imprinting syndromes*.

Numerical Abnormalities Virtually all types of numerical abnormalities are eliminated prenatally, so that only those involving small, gene-poor autosomes or the sex chromosomes are identified with any frequency among live-borns. Clinically, the most important of these is *trisomy 21*, the most frequent cause of Down syndrome. Depending on the maternal age structure of the population and the utilization of prenatal testing, the incidence of trisomy 21 ranges from 1/600 to 1/1000 live births, making it the most common chromosome abnormality in live-born individuals. Like most trisomies, the incidence of trisomy 21 is highly correlated with maternal age, increasing from about 1/1500 live births for women 20 years of age to 1/30 for women 45 years of age and older.

In addition to trisomy 21, only two other autosomal trisomies, 13 and 18, occur with any frequency in livebirths. Incidence rates for trisomies 13 and 18 in live births are 1/20,000 and 1/10,000 respectively. Unlike trisomy 21, which is associated with near-normal life expectancy, both trisomies 13 and 18 are associated with death in infancy, typically occurring during the first year of life.

Three sex chromosome trisomies -- the 47,XXX, 47,XXY (*Klinefelter syndrome*), and 47,XYY conditions -- are quite common, with each occurring in about 1/2000 newborns. Of all the trisomic conditions, these three have the fewest phenotypic complications. In fact, with the exception of infertility in Klinefelter syndrome (Chap. 335), it is likely that most individuals with such trisomic conditions would go undetected. The additional chromosome in the 47,XYY condition is small and contains only a few genes. Most Y-linked genes are involved in testicular development or spermatogenesis. Thus, dosage imbalance of Y-linked genes has relatively little effect on other developmental processes. The 47,XYY genotype is associated with increased height. Its role in antisocial behavior, postulated initially because of an increased prevalence among some penalized populations, is unclear.

For the 47,XXX and 47,XXY conditions, the situation is different -- the X chromosome contains over 1000 genes, many of them essential for normal development. How, then, are 47,XXX and 47,XXY individuals spared from the catastrophic consequences of dosage imbalance? The answer lies in the biology of X chromosome gene expression. In normal females, one of the X chromosomes is inactivated in somatic cells. The inactivation of the paternal or maternal X chromosome occurs randomly in each somatic cell and thereby serves as a mechanism of dosage compensation, ensuring that males and females have equal expression of most X-linked genes. The inactivation process occurs at the blastocyst stage of development; prior to this time both X chromosomes are active. In addition, the rules for inactivation are different for germ cells than for somatic cells: in female germ cells both X chromosomes remain active, whereas in male germ cells the X chromosome is inactivated. In addition, not all X-linked genes are inactivated. Some genes on the X chromosome "escape" the inactivating mechanism and are expressed from both X chromosomes. In disorders such as Klinefelter syndrome, some genes may be expressed from both X chromosomes, resulting in phenotypic abnormalities. Individuals with Klinefelter syndrome have small testes, hyalinized seminiferous tubules, and azoospermia or severe oligospermia (Chap. 335). Testosterone levels are variably reduced, and often there is gynecomastia and eunuchoidal body proportions. Antisocial behavior and mild mental deficiency are seen in some individuals. Females with the 47,XXX genotype are more likely to have mild mental deficiency and may be subfertile. Despite these features, sex chromosome trisomies impart relatively minor phenotypic complications in comparison to aneuploidies that involve autosomal chromosomes.

As a rule, monosomic conditions are incompatible with fetal development and, consequently, autosomal monosomies are only rarely identified in spontaneous abortions and are not found among live-born individuals. In fact, the only monosomy compatible with live birth is the 45,X condition, which causes *Turner syndrome*. The 45,X chromosome constitution occurs with surprisingly high frequency, being present in at least 1 to 2% of all pregnancies. More than 99% of all 45,X conceptions are spontaneously aborted. Thus, live-born individuals with a 45,X chromosome constitution represent a rare group of survivors. The 45,X phenotype is mild, presumably because the second copy of many X chromosomal genes is normally inactivated. Nonetheless, Turner syndrome causes gonadal dysgenesis, resulting in infertility and failure to undergo secondary sexual development. Other prominent features are more variable and include short stature, webbing of the neck, and shield-shaped chest; lymphedema; increased carrying angle at the elbow; cardiovascular and renal abnormalities; and a propensity to hypertension, glucose intolerance, and autoimmune thyroid disease (Chap. 336). Several other structural abnormalities of the X chromosome such as partial deletions, isochromosome X, or ring chromosomes can cause Turner syndrome. Mosaicism, including 45,X/45,XX, 45X/45,XXX, 45,X/45,XY, and others, also occurs (see below) and contributes to the phenotypic spectrum seen in Turner syndrome.

Because numerical abnormalities originate in meiosis (Table 66-3), affected individuals have missing or extra chromosomes in all cells. In a small proportion of cases, though, a mitotic nondisjunctional event occurs at an early stage in an individual with an initially normal chromosome constitution. Alternatively, a "normalizing" mitotic nondisjunctional event may result in a normal chromosome complement in some cells of an embryo. In either case, the embryo is a *mosaic*, with some cells bearing a normal chromosome constitution and others an aneuploid number of chromosomes. The phenotypic consequences are difficult to predict because they depend on the timing of nondisjunction and the distribution of normal and abnormal cells in different tissues. Nevertheless, mosaicism may lead to clinical abnormalities indistinguishable from those of nonmosaic individuals; for example, nearly 5% of all cases of Down syndrome involve individuals with mosaic trisomy 21, and about 15% of individuals with Turner syndrome are mosaic for various sex chromosomal constitutions as described above.

The Origin and Etiology of Numerical Abnormalities Over the past decade, a number of studies have used DNA polymorphisms to investigate the origin of different types of chromosome abnormalities (Fig. 66-4). The most thoroughly investigated types have been numerical abnormalities (Table 66-3). Sex chromosome monosomy usually results from loss of the paternal sex chromosome. This is the case regardless of whether the conception is live-born or spontaneously aborted, indicating that the parental origin of the abnormality does not affect its likelihood of surviving to term.

Trisomies show remarkable variation in parental origin. For example, paternal nondisjunction is responsible for nearly 50% of 47,XXY but only 5 to 10% of cases of trisomies 13, 14, 15, 21, and 22; it is rarely, if ever, the source of the additional chromosome in trisomy 16. Similarly, there is considerable variability in the meiotic stage of origin. For example, among maternally derived trisomies, all cases of trisomy 16 may be due to meiosis I errors, whereas for trisomy 21, one-third of cases are associated with meiosis II errors, and for trisomy 18, the majority of cases are apparently due to meiosis II nondisjunction. In spite of this variation in parental and meiotic origin, nondisjunction at maternal meiosis I appears to be the most common source of trisomy.

Molecular studies have also begun to shed light on the molecular mechanisms underlying nondisjunction, the source of trisomy and monosomy. Most, if not all, trisomies are associated with alterations in genetic recombination. This is the process by which chromosomes exchange genetic material during the first of the two meiotic divisions. In other organisms, the physical connections, or chiasmata, associated with recombination are known to hold chromosomes together at meiosis I. This mechanism is now known to be true for humans as well. Nondisjunction at meiosis I is linked to a reduced extent of crossing-over, with some cases involving outright failure of recombination between the homologous chromosomes, and others associated with distally placed exchanges. Unexpectedly -- since recombination occurs at meiosis I -- maternal meiosis II errors involving chromosome 21 may also be linked to altered recombination. In this instance, though, the effect involves increased -- not decreased -- recombination, especially in proximal 21q. Presumably, this indicates that errors scored as arising at meiosis II are, in fact, precipitated by events occurring at meiosis I.

Maternal Age and Trisomy The association between increasing maternal age and trisomy is arguably the most important etiologic factor in congenital chromosomal disorders. Among women under the age of 25, approximately 2% of all clinically recognized pregnancies are trisomic; by the age of 36, however, this figure increases to 10% and by the age of 42, to >33% (Fig. 66-5). This association between maternal age and trisomy is exerted without respect to race, geography, or socioeconomic factors and likely affects segregation of all chromosomes.

Despite the importance of increasing age, almost nothing is known about the mechanism by which aging leads to abnormal chromosomal segregation. As noted above, it is thought to originate in maternal meiosis I owing to the protracted time to completion (often³40 years) in females. As noted above, alterations in genetic recombination may explain age-related trisomy. In trisomy 21, for example, recombination patterns appear to be similarly altered in younger and older mothers of trisomic conceptions. With this in mind, it has been suggested that two distinct steps, or "hits," may be involved in maternal age-related nondisjunction. The first hit, which is age-independent, involves the establishment of a "vulnerable" recombination configuration in the fetal oocyte; the second hit, which is age-dependent, involves abnormal processing of the vulnerable bivalent structure at metaphase I. If this model is correct, it means that the nondisjunctional process is the same in younger and older women, but it occurs more frequently with aging, possibly because of age-dependent degradation of cell cycle proteins or meiotic proteins responsible for maintaining sister chromatid cohesion.

Structural Chromosome Abnormalities Structural rearrangements involve breakage and reunion of chromosomes. Although less common than numerical abnormalities, they present additional challenges from a genetic counseling standpoint. This is because structural abnormalities, unlike numerical abnormalities, can be present in "balanced" form in clinically normal individuals but transmitted in "unbalanced" form to progeny, thereby resulting in a hereditary form of chromosome abnormality.

Rearrangements may involve exchanges of material between different chromosomes (*translocations*) or loss, gain, or rearrangements of individual chromosomes (e.g., *deletions*, *duplications*, *inversions*, *rings* or *isochromosomes*). Of particular clinical importance are translocations, which involve two basic types: Robertsonian and reciprocal. *Robertsonian rearrangements* are a special class of translocation, in which the long arms of two acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22) join together, generating a fusion chromosome that contains virtually all of the genetic material of the original two chromosomes. If the Robertsonian translocation is present in unbalanced form, a monosomic or trisomic conception ensues. For example, approximately 3% of cases of Down syndrome are attributable to unbalanced Robertsonian translocations, most often involving chromosomes 14 and 21. In this instance, the affected individual has 46 chromosomes, including one structurally normal chromosome 14, two structurally normal chromosomes 21, and one fusion 14/21 chromosome. This effect leads to a normal diploid dosage for chromosome 14 and to a triplication of chromosome 21, thus resulting in Down syndrome. Similarly, a small proportion of individuals with trisomy 13 syndrome are clinically affected because of an unbalanced Robertsonian translocation.

Reciprocal translocations involve exchanges between any two chromosomes. In this circumstance, the phenotypic consequences associated with unbalanced translocations depend on the location of the breakpoints, which dictate the amount of material that has been "exchanged" between the two chromosomes. Because most reciprocal translocations involve unique sets of breakpoints, it is difficult to predict the phenotypic consequences in any one situation. In general, severity is determined by the amount of excess or missing chromosome material in individuals with unbalanced translocations.

In addition to rearrangements between chromosomes, there are several examples of intrachromosome structural abnormalities. The most common and deleterious of these involve loss of chromosome material due to deletions. The two best-characterized deletion syndromes, *Wolf-Hirschhorn syndrome* and *cri-du-chat syndrome*, result from loss of relatively small chromosomal segments on chromosomes 4p and 5p, respectively. Nonetheless, each is associated with multiple congenital anomalies, developmental delays, profound retardation, and reduced lifespan.

Microdeletion Syndromes The term *contiguous gene syndromes* refers to genetic disorders that mimic a combination of single gene disorders. They result from the deletion of a small number of tightly clustered genes. Because they are usually too small to be detected cytogenetically, they are termed *microdeletions*. The application of molecular techniques has led to the identification of at least 18 of these microdeletion syndromes (Table 66-4). Some of the more common ones include the Wilms' tumor-aniridia complex (WAGR),MDS, andVCFsyndrome. *WAGR* is characterized by mental retardation and involvement of multiple organs, including kidney (Wilm's tumor), eye (aniridia), and the genitourinary system. The cytogenetic abnormality involves a deletion of part of the short arm of chromosome 11 (11p13), which typically is detectable on well-banded chromosome preparations. In *MDS*, a disorder characterized by mental retardation, dysmorphic faces, and lissencephaly, the deletion involves chromosome 17 (17p13). UsingFISH, 17p deletions have been detected in >90% of MDS patients as well as in 20% of cases of isolated lissencephaly.

Deletions involving the long arm of chromosome 22 (22q11) are the most common microdeletions identified to date, present in approximately 1/3000 newborns.VCFsyndrome, the most commonly associated syndrome, consists of learning disabilities or mild mental retardation, palatal defects, a hypoplastic aloe nasi and long nose, and congenital heart defects (conotruncal defect). Some individuals with 22q11 deletion are more severely affected and present with *DiGeorge syndrome*, which involves abnormalities in the development of the third and fourth branchial arches leading to thymic hypoplasia, parathyroid hypoplasia, and conotruncal heart defects. In approximately 30% of these cases, a deletion at 22q11 can be detected with high-resolution banding; by combing conventional cytogenetics,FISH, and molecular detection techniques (i.e., Southern blotting or polymerase chain reaction analyses), these rates improve to >90%. Additional studies have demonstrated a surprisingly high frequency of 22q11 deletions in individuals with nonsyndromic conotruncal defects. Approximately 10% of individuals with a 22q11 deletion inherited it from a parent with a similar deletion.

Smith-Magenis syndrome involves a microdeletion localized to the short arm of chromosome 17 (17p11.2). Affected individuals have mental retardation, dysmorphic facial features, delayed speech, peripheral neuropathy, and behavior abnormalities. Most of these deletions can be detected with cytogenetic analysis, althoughFISH is available to confirm these findings. In contrast, *William syndrome*, a chromosome 7 (7q11.23) microdeletion, cannot be diagnosed with standard or high-resolution analysis; it is only detectable utilizing FISH or other molecular methods. William syndrome involves a deletion of the elastin gene and is characterized by mental retardation, dysmorphic features, a gregarious personality, premature aging, and congenital heart disease (usually supravalvular aortic stenosis).

In addition to microdeletion syndromes, there is now at least one well-described microduplication syndrome, Charcot-Marie-Tooth type 1A (CMT1A). This is a nerve conduction disease previously though to be transmitted as a simple autosomal dominant disorder. Recent molecular studies have demonstrated that affected individuals are heterozygous for duplication of a small region of chromosome 17 (17p11.2-12). Although it is not yet clear why increased gene dosage would result in CMT1A, the inheritance pattern is explained by the fact that one-half of the offspring of affected individuals inherit the duplication-carrying chromosome.

Imprinting Disorders Two other microdeletion syndromes,PWS andAS, exhibit parent-of-origin, or "imprinting," effects. For many years, it has been known that cytogenetically detectable deletions of chromosome 15 occur in a proportion of patients with PWS, as well as in those with AS. This seemed curious, as the clinical manifestations of the two syndromes are very dissimilar. PWS is characterized by obesity, hypogonadism, and mild to moderate mental retardation, whereas AS is associated with microcephaly, ataxic gait, seizures, inappropriate laughter, and severe mental retardation. New insight into the pathogenesis of these disorders has been provided by the recognition that parental origin of the deletion determines which phenotype ensues: if the deletion is paternal, the result is PWS, whereas if the deletion is maternal, the result is AS (Fig. 66-2).

This scenario is complicated further by the recognition that not all individuals withPWS orAS carry the chromosome 15 deletion. For such individuals, it turns out that the parental origin of the chromosome 15 region is again the important determinant. In PWS, for example, nondeletion patients invariably have two maternal and no paternal chromosomes 15 [*maternal uniparental disomy* (UPD)], whereas for some nondeletion AS patients the reverse is true (*paternal UPD*). This indicates that at least some genes on chromosome 15 are differently expressed, depending on which parent contributed the chromosome. Additionally, this means that normal fetal development requires the presence of one maternal and one paternal copy of chromosome 15.

Approximately 70% ofPWScases are due to paternal deletions of 15q11-q13, whereas 25% are due to maternal uniparental disomy, and about 5% are caused by mutations in a chromosome 15 imprinting center. Though 75% of theAS cases are due to maternal deletions, only 2% are due to paternal uniparental disomy. The rest of the cases are presumed to be caused by imprinting mutations (5%) or mutations in the UBE3A gene, which is one of the genes associated with AS. TheUPDcases are mostly caused by meiotic nondisjunction resulting in trisomy 15, subsequently followed by a normalizing mitotic nondisjunction event ("trisomy rescue") resulting in two normal chromosomes 15, both from the same parent. *UBE3A* is the only maternally imprinted gene known in the critical region of chromosome 15. However, several paternally imprinted genes, or expressed-sequence tags (ESTs), have been identified, including *ZNF127*, *IPW*, *SNRPN*, *SNURF*, *PAR1*, and *PAR5*.

Chromosomal regions that behave in the manner observed inPWS andAS are said to be *imprinted*. This phenomenon is involved in differential expression of certain genes on different chromosomes. Chromosome 11 must be one of these with an imprinted region, since it is known that a small proportion of individuals with the *Beckwith-Wiedemann overgrowth syndrome* have two paternal but no maternal copies of this chromosome.

ACQUIRED CHROMOSOME ABNORMALITIES IN CANCER

In addition to the constitutional cytogenetic chromosomal abnormalities that are present at birth, somatic chromosomal changes can be acquired later in life and are often associated with malignant conditions. As with constitutional abnormalities, somatic changes can include the net loss of chromosomal material (due to a deletion or loss of a chromosome), net gain of material (duplication or gain of a chromosome), and relocation of DNA sequences (translocation). These chromosomal changes are intertwined with the three major categories of cancer genes: (1) *tumor-suppressor genes* that inhibit cell proliferation may be deleted; (2) *oncogenes* that activate cell proliferation may be activated by duplication, amplification, or translocation; and (3) *DNA repair genes* may also be deleted from somatic cells, thereby predisposing to the accumulation of additional DNA damage. Cytogenetic changes have been particularly well studied in (1) leukemias, e.g., Philadelphia chromosome translocation inCML[t(9;22)(q34.1;q11.2)]); and (2) lymphomas, e.g., translocations of MYC in Burkitt's [t(8;14)(q24;q32)]. These and other translocations are useful for diagnosis, classification, and prognosis. Analyses of cytogenetic changes are also proving useful in certain solid tumors. For example, a complex karyotype with Wilms' tumor, diploidy in medulloblastoma, and Her-2/neu amplification in breast cancer are poor prognostic signs. The genetic basis of malignancy, including the role of germline and somatic chromosomal alterations, is discussed further inChap. 81.

(Bibliography omitted in Palm version)

Back to Table of Contents

67. DISEASES CAUSED BY GENETIC DEFECTS OF MITOCHONDRIA - *Donald R. Johns*

Mitochondrial defects play a role in several metabolic and neurodegenerative diseases as well as aging. Mitochondrial disorders have protean clinical manifestations, reflecting the fact that nearly all organ systems utilize oxidative metabolism. Clinical features often involve tissues with high energy requirements such as the central and peripheral nervous systems, the eye, muscle, kidney, and endocrine organs. The age and mode of onset and clinical course of mitochondrial diseases range widely as a result of the unusual mechanisms of mitochondrial DNA (mtDNA) replication, which is distinct from that of the nuclear genome. A maternal mode of inheritance is characteristic of many mitochondrial diseases because mtDNA is transmitted by the oocyte. Hundreds of different mtDNA mutations have been described since the first mutation was described in 1988.

STRUCTURE AND FUNCTION OF MITOCHONDRIAL DNA

Most cells contain several hundred mitochondria, though the number varies depending on the energy requirements and function of a tissue. Mitochondria are the only cellular organelles that contain their own extrachromosomal DNA. HumanmtDNA is a small (16,569 bp) double-stranded, circular molecule that encodes 13 protein subunits of 4 different oxidative phosphorylation biochemical complexes. mtDNA also encodes the 24 structural RNAs (2 ribosomal RNAs and 22 transfer RNAs) required for the intramitochondrial translation of these proteins. The noncoding D (displacement)-loop is a regulatory region that controls transcription and replication. mtDNA mutations are found in each type of mitochondrial gene (Fig. 67-1).

Mitochondria probably evolved from independent organisms that became endosymbiotically incorporated into the cell. As a result, mitochondria replicate, transcribe, and translate their DNA independently of nuclear DNA. However, cellular and mitochondrial function are interdependent. Nuclear DNA-encoded proteins are also involved in oxidative phosphorylation, and the myriad macromolecular compounds required for mitochondrial structure and function (e.g.,mtDNAreplication, transcription, and translation) are imported from the cytoplasm into the mitochondria.

Oxidative phosphorylation and the generation of adenosine triphosphate (ATP) for energy-requiring cellular processes are the central functions of mitochondria. Alterations in mitochondrial function lead to disease pathogenesis by three main mechanisms: (1) reduction of ATP supply when mutations impair oxidative phosphorylation; (2) generation of reactive oxygen species such as $H_2O_2$and OH× free radicals that can damage DNA, proteins, or lipids; and (3) execution of the apoptosis pathway when mitochondria release cell death-promoting factors including caspases, cytochrome c, and apoptosis-inducing factor.

Several unique features ofmtDNArender it vulnerable to mutations and contribute to its role in disease. For example, mtDNA has no introns (a random mutation will therefore usually strike a coding DNA sequence) or protective histones, and it has an imperfect DNA repair system and is exposed to oxygen free radicals generated by oxidative phosphorylation. mtDNA is estimated to mutate 10 times more rapidly than nuclear

DNA. Importantly, mtDNA is strictly *maternally inherited* and does not recombine, and mtDNA mutations sequentially accumulate along maternal lineages. These properties have made mtDNA sequence variation an invaluable tool for evolutionary biologists and forensic scientists.

Each mitochondrion contains 2 to 10mtDNAmolecules, and each cell contains multiple mitochondria (*polyplasmy*). Population genetic principles, as opposed to Mendelian genetics, govern mitochondrial genetics. When a new mtDNA mutation arises, cells initially harbor copies of both normal and mutant DNA sequences -- a condition known as *heteroplasmy* -- which allows an otherwise lethal mutation to persist and cause disease. The presence of either completely normal or completely mutant mtDNA is known as *homoplasmy*. During cell division, mitchondria are unevenly partitioned to the daughter cells through the process of *replicative segregation*; consequently, the proportion of mutant and normal mtDNA molecules can drift. Selection pressures apply at the molecular, cellular, and organismal levels. The critical proportion of mutant mtDNA required for deleterious phenotypic expression is known as the *threshold effect*. It varies among individuals, among organ systems, and within a given tissue, depending on the delicate balance of oxidative supply and demand. These features of mtDNA segregation, combined with the uneven transmission of mitochondria to daughter cells during cell division, form much of the basis for the phenotypic diversity seen in mitochondrial diseases.

MITOCHONDRIAL DNA MUTATIONS

mtDNAmutations that cause a severe, lethal impairment of oxidative phosphorylation -- such as gross structural defects or point mutations in structural RNAs -- are only viable if heteroplasmic. In contrast, the majority of the milder, missense mtDNA mutations in protein-coding genes are homoplasmic. mtDNA point mutations have been found in each type of mtDNA gene, but tRNA mutations predominate in severe, multisystemic mitochondrial encephalomyopathy phenotypes; protein-coding gene mutations predominate in Leber's hereditary optic neuropathy (LHON). A point mutation in the 12S ribosomal RNA gene is associated with both spontaneous and aminoglycoside-associated sensorineural deafness (Fig. 67-1).

A definitive cause-and-effect relationship between amtDNAmutation and a clinical phenotype can be difficult to establish for several reasons: (1) mtDNA is highly polymorphic, (2) different mutations can be associated with the same phenotype or the same mutation can be associated with different phenotypes, and (3) epigenetic factors can affect clinical manifestations.

CLASSIC MITOCHONDRIAL ENCEPHALOMYOPATHY PHENOTYPES

Many diseases were provisionally classified as mitochondrial disorders on the basis of abnormal mitochondrial morphology, biochemistry, or a pattern of maternal inheritance. Disease-associatedmtDNAmutations are now an important diagnostic criterion. Though each classic mitochondrial encephalomyopathy phenotype has distinctive clinical features (Table 67-1), each also shares many clinical and laboratory features (Tables 67-2 and67-3).

Mitochondrial Myopathy Mitochondrial myopathy is characterized by fixed proximal weakness with marked exercise intolerance. Fatigability and poor stamina are prominent clinical features, but a mitochondrial etiology is often considered in the context of other neurologic, somatic, or laboratory features. Frank rhabdomyolysis is rare. Electromyography documents a nonirritative myopathy, and serum creatine kinase is usually normal or slightly elevated. Skeletal muscle biopsy shows abnormal proliferating mitochondria and "ragged red fibers," a histologic hallmark of the severe biochemical defects in oxidative phosphorylation. LargemtDNAdeletions and a variety of mtDNA point mutations occur in mitochondrial myopathy.

Chronic Progressive External Ophthalmoplegia (CPEO) Ptosis, ophthalmoplegia, and limb myopathy characterize CPEO. Additional clinical features (Table 67-2) may also occur along with the laboratory abnormalities characteristic of mitochondrial disorders (Table 67-3). Patients with CPEO have abnormal skeletal muscle biopsies, with ragged red fibers and ultrastructural changes. Additional somatic and central nervous system findings in conjunction with CPEO are known as the "CPEO-plus" syndromes. Kearns-Sayre syndrome is a subset of CPEO-plus that begins before age 20 and is characterized by CPEO and atypical pigmentary retinopathy; ancillary features include elevated cerebrospinal fluid protein, ataxia, or heart block.

Most patients withCPEOhave large, single deletions inmtDNAthat can be reliably detected by molecular genetic methods. However, skeletal muscle is required as the source of DNA because almost all single mtDNA deletions occur sporadically. The presence of pigmentary retinopathy is strongly predictive of a deletion. The mechanism of DNA deletion is unknown, though recombination or slippage during replication is plausible. The junctions of most deletions contain directly repeated sequences, including a 13-nucleotide direct repeat "hot spot" that accounts for about 25% of all deletions. Approximately half of all deletions are bound by other direct repeats; one-quarter have no apparent direct repeat. A few patients have partially duplicated mtDNA molecules. A point mutation at nucleotide position 3243 has been found in many patients who lack a mtDNA deletion or duplication.

Autosomally Transmitted Multiple Mitochondrial DNA Deletions Several families with clinical variants ofCPEOhave been found to harbor multiplemtDNAdeletions in skeletal muscle. The autosomal inheritance of multiple mtDNA deletions implies a primary defect in a nuclear DNA gene that has secondary qualitative effects on mtDNA. Multiple mtDNA deletions can be transmitted in either an autosomal dominant or autosomal recessive mode or can occur as somatic mutations.

Thymidine phosphorylase was the first nuclear-encoded gene shown to influence the regulation and function of normalmtDNA in a trans-acting manner. Mutations in thymidine phosphorylase cause an autosomal recessive disease known as *myoneurogastrointestinal encephalomyopathy* (MNGIE). Several different nuclear loci have been linked to the autosomal dominant forms ofCPEO. Tissue-specific, autosomally transmitted depletion of mtDNA represents a quantitative mtDNA defect caused by an intergenomic communication error.

Mitochondrial Encephalomyopathy, Lactic Acidosis, and Strokelike Episodes (MELAS) This syndrome is characterized by strokelike events that cause subacute brain dysfunction, cerebral structural changes, seizures, and several other common clinical and laboratory features (Tables 67-2,67-3). Maternal inheritance of the MELAS syndrome may be obscured because of mild clinical features in relatives. A point mutation at nucleotide 3243 in the tRNA$_{Leu(UUR)}$gene accounts for 80% of MELAS cases. However, the clinical features of the 3243 mtDNA mutation are pleiomorphic; it is also associated with nondeletionCPEO, myopathy, deafness, diabetes, and dystonia.

Myoclonic Epilepsy with Ragged Red Fibers (MERRF) Syndrome The MERRF syndrome consists of myoclonus, seizures, cerebellar ataxia, and mitochondrial myopathy. Pathogenetic mutations have been demonstrated at nucleotide positions 8344 and 8356 in the tRNA$_{Lys}$gene. Neurologic and laboratory features common to other mitochondrial encephalomyopathies are seen. Maternal relatives may be asymptomatic or may have partial clinical syndromes, including lipomas in a characteristic "horse-collar" distribution, and hypertension.

Neuropathy, Ataxia, and Retinitis Pigmentosa (NARP)/Maternally Inherited Leigh's Disease NARP is characterized by proximal weakness, sensory neuropathy, developmental delay, ataxia, seizures, dementia, and retinal pigmentary degeneration. This maternally inherited disorder is associated with two different heteroplasmic missense mutations at nucleotide position 8993 in the ATPase 6 gene. High proportions of the same mutations are also seen in maternally inherited Leigh's disease. Autosomal recessive Leigh's disease is associated with cytochrome c oxidase deficiency and is caused by a deficiency of the nuclear-encoded protein, SURF1, which is required for biogenesis of the cytochrome c oxidase complex (complex IV). ThemtDNApoint mutations associated with NARP,MELAS,MERRF, and other mitochondrial disorders are readily detected by molecular genetic analysis of mtDNA extracted from muscle or blood.

Leber's Hereditary Optic Neuropathy LHONtypically presents with painless, subacute, bilateral visual loss with central scotomas and dyschromatopsia. The mean age of onset is 23 years, and males are affected three to four times more commonly than females. LHON bears little clinical resemblance to the other mitochondrial disease phenotypes. It was first classified in this group on the basis of the maternal inheritance pattern. The pathophysiology of visual loss appears to involve both genetic and epigenetic (tobacco and alcohol) factors. ThemtDNAmutations exhibit a high degree of genetic heterogeneity. Primary LHON-associated mtDNA mutations predominantly affect complex I genes [at nucleotide positions 11778 (ND-4 gene), 3460 (ND-1 gene), and 14484 (ND-6 gene)] (Fig. 67-1). Several other mtDNA mutations may have secondary pathogenetic roles in LHON, including a mutation at nucleotide position 13708 (ND-5 gene) of Caucasian haplogroup J mtDNA.

Genotype-phenotype correlations are beginning to emerge for the primaryLHON-associatedmtDNAmutations. The prognosis for visual recovery, for example, varies nearly ten-fold depending on the mutation. mtDNA mutations that cause LHON plus dystonia have also been described.

ORGAN SYSTEM MANIFESTATIONS OF MITOCHONDRIAL DISEASE

Because virtually all tissues of the body depend, to some extent, on oxidative metabolism, patients with mitochondrial disease can present to many specialists in medicine. The somatic manifestations listed in Table 67-2, first noted in association with the classic mitochondrial diseases, may be the dominant or initial clinical symptom or may be important comorbid features.

The ophthalmologic manifestations ofmtDNAmutations are prominent, with involvement of virtually the entire visual axis from the lids, cornea, and extraocular muscles to the occipital cortex. The cardinal eye findings include ophthalmoplegia, optic neuropathy, and pigmentary retinopathy. Cardiovascular manifestations include dilated and hypertrophic cardiomyopathy, conduction disease and heart block, Wolff-Parkinson-White syndrome, and hypertension. The prevalence of diabetes mellitus is higher than expected in patients with mitochondrial encephalomyopathies, and it occurs in association with a variety of mtDNA mutations. Diabetes mellitus has been linked with the 3243 mtDNA point mutation, usually, but not exclusively, in association with sensorineural hearing loss.

ROLE OF MITOCHONDRIAL DNA MUTATIONS IN PREVALENT DISEASES

The role ofmtDNAmutations in common, socioeconomically significant diseases is under active investigation. The genetic basis of many prevalent diseases is complex and does not follow simple, single-gene Mendelian inheritance (Chap. 65). Mitochondrial diseases, such asLHON and aminoglycoside-induced deafness, illustrate the potential for complex pathophysiologic interactions between genetic and epigenetic factors. As a result of these interactions, mtDNA mutations may be involved in subsets of common diseases, such as diabetes mellitus, in which the maternal inheritance pattern is not obvious.

The tissue-specific accumulation of somatic (noninherited)mtDNAmutations is likely relevant to some late-onset degenerative disorders, such as Alzheimer's disease and Parkinson's disease. It has been shown, for example, that as people age, mtDNA mutations accumulate in tissues, including some postmitotic tissues such as the basal ganglia and cerebral cortex. The high mutations rate and poor repair capacity of mtDNA contribute to the buildup of mtDNA mutations in postmitotic tissues or those with a slower turnover rate. Oxidative damage, as occurs with repeated episodes of ischemia and reperfusion, markedly increases the accumulation of mtDNA mutations. Environmental factors may also affect mtDNA. The anti-retroviral drug azidothymidine depletes muscle mtDNA and causes an acquired mitochondrial myopathy. Cumulative, age-dependent mitochondrial dysfunction, mediated to a significant degree by oxidative damage to mtDNA and other mitochondrial macromolecules, may be a major contributor to aging.

The unequivocal establishment of the diagnosis of a mitochondrial disease by molecular genetic methods is a prerequisite for proper genetic counseling and ultimately treatment.

(Bibliography omitted in Palm version)

Back to Table of Contents

68. SCREENING, COUNSELING, AND PREVENTION OF GENETIC DISORDERS - *Susan Miesfeldt*, *J. Larry Jameson*

IMPLICATIONS OF MOLECULAR GENETICS FOR INTERNAL MEDICINE

Approximately 1 in 50 children is born with a serious congenital abnormality or mental handicap. It is known that each of us harbors mutations in several genes that can potentially lead to serious diseases. The field of medical genetics has traditionally focused on chromosomal abnormalities (Chap. 66) and Mendelian disorders (Chap. 65). However, there is genetic susceptibility to many common adult-onset diseases including atherosclerosis, hypertension, autoimmune diseases, diabetes mellitus, Alzheimer disease, psychiatric disorders, and many forms of cancer. Genetic contributions to these common disorders involve more than the ultimate expression of a disease; these genes can also influence the severity of infirmity, response to treatment, and progression of disease.

The primary care clinician is now faced with the role of recognizing and counseling patients at risk for a large number of genetically influenced illnesses. This role reflects the advances in genetic medicine that are changing the way diseases are classified, enhancing our understanding of pathophysiology, providing practical information concerning drug metabolism and therapeutic responses, and promising to allow individualized screening and health care management programs. In view of these changes, the physician must integrate personal medical history, family history, and diagnostic molecular testing into the overall care of individual patients and their families. In addition, the internist has an important role in educating patients about the indications, benefits, risks, and limitations of genetic testing in the management of diverse diseases. This is a formidable task as scientific advances in genetic medicine, and media attention to these advances, have outpaced the translation into standards of clinical care.

PRENATAL AND NEWBORN GENETIC SCREENING AND TESTING

During pregnancy and in the newborn period, genetic screening and testing are both used to detect genetic disorders (Table 68-1). *Genetic screening* refers to the search for a genetic disorder in the entire population or in a high-risk population. Screening techniques must be cost-effective, have a high positive predictive value, and should yield information that leads to disease prevention or a useful therapeutic intervention. Examples of screening tests include maternal serum markers used to detect increased risk of Down syndrome, postnatal tests for phenylketonuria, and cholesterol levels in children used to identify individuals at risk for familial hyperlipidemias. Thus, genetic screening tests do not necessarily imply DNA- or chrozmosome-based tests. Instead, many such tests use a surrogate biochemical marker or phenotypic feature of the underlying genetic disorder. Determination of which genetic disorders should be screened routinely depends on disease frequency, the severity of the disorder, cost of screening, and whether treatment interventions can alter the course of the disease.

Genetic testing, on the other hand, is used in an individual suspected to have a disease based on physical features, family history, or biochemical findings. Genetic testing can include the following: (1) diagnostic testing as for hemochromatosis, (2) predictive testing as for breast cancer predisposition, (3) carrier testing as for muscular dystrophy, and (4) prenatal testing as forb-thalassemia when both parents are carriers.

Depending on the disorder(s) under consideration, several different techniques are currently used for genetic screening during pregnancy. For instance, first-trimester screening for Down syndrome is performed using measurements of maternal serum pregnancy-associated plasma protein A and the free b subunit of human chorionic gonadotropin in combination with ultrasonographic measurement of nuchal translucency (at 10 to 14 weeks of gestation). Second-trimester maternal serum screening includes measurements of human chorionic gonadotropin, unconjugated estriol, a-fetoprotein, and inhibin that, in combination, increase the sensitivity of Down syndrome detection (Chap. 66). Increaseda-fetoprotein levels are also associated with open neural tube defects.

Amniocentesis or *chorionic villus sampling* (CVS) is used to isolate fetal cells for chromosomal analysis or to test for specific genetic abnormalities. Amniocentesis is typically performed during the early second trimester (14 to 16 weeks' gestational age). CVS can be performed earlier (8 to 10 weeks' gestational age) and involves transcervical or transabdominal biopsy of fetal trophoblastic tissue. CVS is associated with a somewhat greater risk of spontaneous abortion (<0.5 to 1%) than amniocentesis (<0.5%) but allows for elective abortion earlier during pregnancy. Ultrasonography is used to analyze the fetus directly at different stages of development. Preimplantation molecular diagnostic testing is now possible by isolating single cells from the 8- to 10-cell embryo. The polymerase chain reaction (PCR) is then used to test for selected single-gene disorders such as Tay-Sachs disease, cystic fibrosis, or sickle cell anemia. This testing strategy requires in vitro fertilization but has the advantage that affected embryos are not implanted.

It should be remembered that prenatal genetic testing focuses on chromosomal abnormalities (Chap. 66), along with specific genetic disorders for which there is increased risk of parental transmission. There is no guarantee that a child will be free of birth defects or other genetic disorders not included among the diagnostic tests.

The genetic counseling process should begin before prenatal testing and should include: (1) a description of the test, (2) the types of disorders that will be screened, (3) the limitations of the screening,z (4) an exploration of what the parents will do with the information, and (5) an indication of when the results will be available. If a genetic disorder is identified, the full repertoire of genetic counseling skills is required (see below). The nature of the genetic disease needs to be reviewed with the parents, often on separate occasions. The counselor should also discuss the kinds of physical and emotional challenges that a genetic disease might pose for the affected individual and the family. Written information should be provided, if available. Ultimately, the parents must reach a decision to continue or terminate the pregnancy. If pregnancy is terminated, counseling should continue after the procedure. If pregnancy is continued, counseling should continue to address the medical needs of the affected fetus or child, as well as to help the family meet any challenges presented by the genetic disorder.

At the time of birth, all newborns undergo a complete physical examination to detect gross developmental abnormalities. Within 24 to 72 h of birth, blood samples from newborns are sent to a state-designated laboratory to screen for selected diseases, such as congenital hypothyroidism and a variety of inherited metabolic disorders (Table 68-1). This program represents a clear example of the benefits of selected genetic screening. The early diagnosis of phenylketonuria, for example, permits parents to introduce a phenylalanine-free diet before the development of severe neurologic sequelae (Chap. 352).

COMMON ADULT-ONSET GENETIC DISORDERS

MULTIFACTORIAL INHERITANCE

The risk for many adult-onset disorders reflects the additive effects of genetic factors at multiple loci that may function independently -- or in combination -- with other genes or environmental factors. Our understanding of the genetic basis of these disorders is incomplete, despite the clear recognition of genetic susceptibility. In type 2 diabetes mellitus, for example, the concordance rate in monozygotic twins ranges between 50 to 90%. Diabetes or impaired glucose tolerance occurs in 40% of siblings and in 30% of the offspring of an affected individual. Despite the fact that diabetes affects 5% of the population and exhibits a high degree of heritability, there are only a few examples of genetic mutations that might account for the familial nature of the disease (most of which are rare). They include certain mitochondrial DNA disorders (Chap. 67), mutations in a cascade of genes that control pancreatic islet cell development and function (*HNF4*a, *HNF1*a, *IPF1*), insulin receptor mutations, and others (Chap. 333). Obesity and other factors that contribute to insulin resistance also represent important risk factors for type 2 diabetes. Current models for the genetic basis of type 2 diabetes propose the involvement of more than a dozen genes: some genes influence pancreatic islet development or function; others likely modulate glucose-sensing; and an important group determine insulin sensitivity, either directly by affecting insulin signaling or indirectly by regulating body weight or composition. Superimposed on this genetic background are environmental influences such as diet, exercise, pregnancy, and medications.

Identifying these susceptibility genes is a formidable task. Nonetheless, a reasonable goal for this type of disease is to identify genes that increase (or decrease) disease risk by a factor of two or more. For common diseases such as diabetes or heart disease, this level of risk has important implications for health. Much the same way that cholesterol is currently used as a biochemical marker of cardiovascular risk, we can anticipate the development of genetic panels with similar predictive power. Genetic tests for a large number of genetic disorders are now available; a web site (http://www.genetests.org) lists various laboratories that perform specific tests. The advent of DNA-sequencing chips represents an important technical advance that promises to make large-scale testing more feasible (Chap. 65). The decision whether or not to perform a genetic test for a particular inherited adult-onset disorder, such as hemochromatosis, multiple endocrine neoplasia (MEN) type 1, prolonged QT syndrome, or Huntington disease, is complex; it depends on the clinical features of the disorder, the desires of the patient and family, and whether the results of genetic testing will alter medical decision-making or treatment (see below).

THE FAMILY HISTORY

Pending additional advances in genetic testing, the key to determining the inherited risk for common adult-onset diseases still rests in the collection and interpretation of a detailed personal and family medical history in conjunction with a directed physical examination. For example, a history of multiple family members with early-onset coronary artery diseases, glucose intolerance, and hypertension should suggest increased risk for genetic, and perhaps environmental, predisposition to insulin resistance (Chap. 333). Individual patients with this family history should be monitored for the possible development of hypertension, diabetes, and hyperlipidemia. They should be counseled about the importance of avoiding additional risk factors such as obesity and cigarette smoking.

Family history, recorded in the form of a pedigree, greatly assists the assessment of risk in the individual patient. At a minimum, pedigrees should convey health-related data on all first-degree relatives and selected second-degree relatives, including grandparents. When pedigrees appear to suggest an inherited disease, they should be extended to include additional family members. The determination of risk for an asymptomatic individual will vary depending on the size of the pedigree, the number of unaffected relatives, and the types of diagnoses within the family. For example, a woman with two first-degree relatives with breast cancer is more at risk if she has a total of three female first-degree relatives than if she has a total of eight female first-degree relatives. Additional variables that factor into the assessment of risk -- and should be documented in the pedigree -- include the age at diagnosis of each affected family member, present age of all family members, and the presence or absence of nonhereditary risk factors among those affected with diseases.

When assessing the personal and family history, the physician should be alert to a younger age of disease onset than is usually seen in the general population. A 30-year-old with acute myocardial infarction should be considered at risk for a hereditary trait, even if there is no family history of premature coronary artery disease (Chap. 241). The absence of the nonhereditary risk factors typically associated with a disease also raises the prospect of genetic risk factors. A personal or family history of deep vein thrombosis, in the absence of known nongenetic risk factors, might suggest a hereditary thrombotic disorder (Chap. 117). The physical examination may also provide important clues concerning the risk for a specific inherited disorder. A patient with xanthomas at a young age should prompt consideration of familial hypercholesterolemia. Some adult-onset disease-causing mutations are more prevalent in certain ethnic groups. For instance,>2% of the Ashkenazi population carry one of three specific mutations in the *BRCA1* or *BRCA2* genes. The prevalence of the factor V Leiden allele ranges from 3 to 7% in Caucasians but is much less common in Africans or Asians.

Recall of family history is often inaccurate. This is especially so when the history is remote and as families become more dispersed. It can be helpful to ask patients to fill out family history forms before or after their visits, as this provides them with an opportunity to contact relatives. Attempts should be made to confirm the illnesses reported in the family history before making management decisions. This process is often labor-intensive and ideally involves interviews of additional family members or reviewing medical records, autopsy reports, and death certificates.

Nongenetic factors associated with disease risk should also be reviewed in full, including occupation, diet, living conditions, and social habits. For example, patients at hereditary risk for heart disease should be questioned about tobacco use, diet, exercise, and lipid levels. Patients should also be asked about their health screening and prevention behaviors. These nonhereditary factors contribute to the assessment of overall risk and represent an important focus for disease prevention.

Although many inherited disorders will be suggested by the clustering of relatives with the same or related conditions, it is important to note that *disease penetrance* is incomplete for most multifactorial genetic disorders. As a result, the pedigree obtained in such families may not exhibit a clear Mendelian pattern of inheritance, as not all family members carrying the disease-associated alleles will manifest disease. Furthermore, genes associated with some of these disorders often exhibit *variable expression* of disease. For example, the breast cancer-associated gene, *BRCA1*, can predispose to several different malignancies in the same family, including cancers of the breast, ovary, and prostate (Chap. 81). For common diseases such as breast cancer, some family members without the disease-causing mutation may also develop breast cancer, representing another confounding variable in the pedigree analysis.

Some of the aforementioned features of the family history are illustrated inFig. 68-1. The proband, a 36-year-old woman, has a strong history of breast and ovarian cancer on the paternal side of her family. The early age of onset, as well as the co-occurrence of breast and ovarian cancer in this family, suggests the possibility of an inherited mutation in *BRCA1* or *BRCA2*. It is unclear though -- without genetic testing -- whether her father inherited such a mutation and transmitted it to her. After appropriate genetic counseling of the proband and her family, one approach to DNA analysis in this family is to test the potentially affected 42-year-old living cousin (IV-4) for the presence of a *BRCA1* or *BRCA2* mutation. If a mutation is found, then it is possible to test for this particular alteration in the proband and other family members, if they so desire. In the example shown, if the proband's father has inherited the *BRCA1* mutation, there is a 50:50 probability that the mutation has been transmitted to her. Genetic testing can be used to establish the absence or presence of this particular risk factor.

GENETIC TESTING FOR ADULT-ONSET DISORDERS

A critical first step before initiating genetic testing is to assure that the correct clinical diagnosis has been made, whether based on family history, characteristic physical findings, or biochemical testing. Careful clinical assessment will prevent unnecessary testing and will direct testing towards the most probable candidate genes. Many disorders exhibit the feature of *locus heterogeneity*, which refers to the fact that mutations in different genes can cause phenotypically similar disorders. For example, osteogenesis imperfecta (Chap. 351), muscular dystrophy (Chap. 383), homocystinuria (Chap. 352), and hereditary predisposition to colon cancer (Chap. 90) or breast cancer (Chap. 89) can each be caused by mutations in distinct genes. The pattern of disease transmission, clinical course, and treatment may differ significantly, depending on which gene is affected. In these cases, the choice of which genes to test is often determined by unique clinical features, the relative prevalence of mutations in various genes, or test availability.

Like all laboratory tests, there are limitations to the accuracy and interpretation of genetic tests. In addition to technical errors, genetic tests are often designed to detect only the most common mutations. In this case, a negative result must be qualified by the possibility that the individual may have a mutation that is not included in the test.

In addition to molecular testing for established disease, presymptomatic testing for susceptibility to chronic disease is being increasingly integrated into the practice of medicine. In most cases, however, the discovery of disease-associated genes has greatly outpaced studies that assess clinical outcomes and the impact of interventions. Until such outcomes-based studies are available, predictive molecular testing must be approached with caution and should be offered only to patients who have been adequately counseled and have provided informed consent (Fig. 68-2). In the majority of cases, presymptomatic testing should be offered only to individuals with a suggestive personal or family medical history or in the context of a clinical trial.

Molecular analysis is generally more informative if testing is initiated in a symptomatic family member, since the identification of a mutation can direct the testing of other at-risk family members (whether they are symptomatic or not). In the absence of additional familial or environmental risk factors, individuals who test negative for the mutation found in the affected family member can be informed that they are at general population risk for that particular disease. Furthermore, they can be reassured that they are not at risk for passing on the mutation to their children. On the other hand, asymptomatic family members who test positive for the known mutation must be informed that they are at increased risk for disease development and for transmitting the mutation to their children. Nevertheless, for most multifactorial genetic disorders, the test results cannot predict with confidence whether, or when, the disease will develop. For example, not everyone with the apolipoprotein E allele (e4) will develop Alzheimer disease, and many individuals without this susceptibility gene can still develop the disorder (Chap. 362).

A negative test result is interpreted differently when no genetic mutation is found in a symptomatic family member. In this difficult circumstance, the test performed on a given gene may not detect all mutations in that gene (false negative) or the individual may have a mutation in a different disease-associated gene that was not tested.

Clinicians providing pretest counseling and education should assess the patient's emotional ability to cope with test results. Individuals who demonstrate signs and symptoms of psychiatric illness should have their emotional needs addressed before proceeding with molecular testing. Generally, genetic testing should not be offered at a time of personal crisis or acute illness within the family. Patients will derive more benefit from test results if they are emotionally able to comprehend and absorb the information. It is important to assess patients' preconceived notions of their personal likelihood of disease in preparing pretest educational strategies. Often, patients harbor unwarranted fear or denial of their likelihood of genetic risk.

Genetic testing has the potential of affecting the way individual family members relate to one another, both negatively and positively. As a result, patients addressing the option of molecular testing must consider how test results might impact their relationships with relatives, spouses, and friends. In families with a known genetic mutation, those who test positive must address the impact of the disease on their present and future lifestyles; those who test negative may manifest survivor guilt. Family members are likely to differ in their emotional and social responses to the same information. Counseling should also address the potential consequences of test results on relationships with a spouse or child. Parents who are found to have a disease-associated mutation often express considerable anxiety and despair as they address the issue of risk to their children.

When a condition does not manifest until adulthood, clinicians will be faced with the question of whether at-risk children should be offered molecular testing and, if so, at what age. Several professional organizations have cautioned that genetic testing for adult-onset disorders should not be offered to children. Many of these conditions are not preventable and, consequently, such information can pose significant psychosocial risk. In addition, there is concern that testing during childhood violates a child's right to make an informed decision regarding testing upon reaching adulthood. On the other hand, testing should be offered in childhood for disorders that may be manifest early in life, especially when management options are available. For example, children at risk for familial adenomatous polyposis (FAP) may develop polyps as early as their teens, and progression to an invasive cancer can occur by their twenties. Likewise, children at risk forMEN type 2, which is caused by mutations in the *RET* proto-oncogene, may develop medullary thyroid cancer as early as 6 years of age, and the issue of prophylactic thyroidectomy should be addressed with the parents of children with documented mutations (Chap. 339).

INFORMED CONSENT

When the issue of testing is addressed, patients should be strongly encouraged to involve other relatives in the decision-making process, as molecular diagnostics will likely have an impact on the entire family. Informed consent for molecular testing begins with detailed education and counseling. The patient must fully understand the risks, benefits, and limitations of undergoing the analysis. Informed consent should be in the form of a written document, drafted clearly and concisely in a language and format that is comprehensible to the patient, who should be made aware of the disposition of test results. Informed consent should also include a discussion of the mechanics of testing. Most molecular testing for hereditary disease involves DNA-based analysis of peripheral blood. In the majority of circumstances, test results should be given only to the individual, in person, and with a support person in the room.

Because molecular testing of an asymptomatic individual often allows prediction of future risk, the patient should understand any potential long-term medical, psychological, and social implications of this decision. In the United States, legislation affecting this area is still evolving, and it is important to explore with the patient the potential impact that test results may have on employment, future health, and life insurance coverage.

Patients should understand that alternatives to molecular analysis remain available if they decide not to proceed with this option. They should also be notified that testing is available in the future if they are not prepared to undergo analysis immediately. The option of DNA banking should be presented so that samples are readily available for future use by family members, if needed.

FOLLOW-UP CARE AFTER TESTING

Depending on the nature of the genetic disorder, posttest interventions may include (1) cautious surveillance and appropriate health care screening, (2) specific medical interventions, (3) chemoprevention, (4) risk avoidance, and (5) referral to support services. For example, patients with known pathologic mutations in *BRCA1* or *BRCA2* are offered intensive screening as well as the option of prophylactic mastectomy and/or oophorectomy. In addition, such women may be eligible for preventive treatment with agents such as tamoxifen, raloxifene, or retinoids. In contrast, those at known risk for Huntington disease are offered continued follow-up and supportive services, including physical and occupational therapy, and social services and support groups as indicated. Specific interventions will change as translational research enhances our understanding of these genetic diseases and as more is learned about the functions of the genes involved.

Individuals who test negative for a mutation in a disease-associated gene identified in an affected family member must be reminded that they may still be at risk for the disease. This is of particular importance for common diseases such as diabetes mellitus, cancer, and coronary artery disease. For example, a woman who finds that she does not carry the disease-associated mutation in *BRCA2* previously discovered in her family must be reminded that she still requires the same breast cancer screening used in the general population.

GENETIC COUNSELING AND EDUCATION

Genetic counseling should be distinguished from genetic testing and screening, even though genetic counselors are often involved in issues related to testing. Genetic counseling refers to *a communication process that deals with human problems associated with the occurrence or risk of a genetic disorder in a family*. Genetic risk assessment can be complex and often involves elements of uncertainty. Counseling therefore includes genetic education as well as psychosocial counseling. Genetic counselors may be called upon by other health care professionals (or by individual patients and families) to address a broad range of issues directly and indirectly involved with genetic disease (Table 68-2). The genetic counselor will do the following:

- Gather and document a detailed family history

- Educate the patient about general genetic principles related to disease risk, both for themselves and others in the family

- Assess and enhance the patient's ability to cope with the genetic information offered

- Discuss how nongenetic factors may relate to the ultimate expression of disease

- Address medical management issues

- Assist in determining the role of genetic testing for the individual and family

· Ensure that the patient is aware of the risks, benefits, and limitations of the various genetic testing options

· Refer the patient and other at-risk family members for additional medical and support services, if necessary.

The complexity of genetic counseling and the broad scope of genetic diseases are leading to the development of specialized, multidisciplinary clinics designed to provide broad-based support and medical care for those at risk and their family members. Such multidisciplinary teams are often composed of medical geneticists, specialist physicians, genetic counselors, nurses, psychologists, social workers, and biomedical ethicists who work together to consider difficult diagnostic, treatment, and testing decisions. Such a format also provides primary care physicians with invaluable support and assistance as they follow and treat at-risk patients.

The approach to genetic counseling has important ethical, social, and financial implications. Philosophies related to genetic counseling vary widely by country and center. In North American centers, for example, counseling is generally offered in a nondirective manner, wherein patients learn to understand how their values factor into a particular medical decision. Nondirective counseling is particularly appropriate when there are no data demonstrating a clear benefit associated with a particular intervention or when an intervention is considered experimental. For example, nondirective genetic counseling is employed when a person is deciding whether or not to undergo genetic testing for Huntington disease (Chap. 362). At this time, there is no clear benefit (in terms of medical outcome) to an at-risk individual undergoing genetic testing for this disease, as its course cannot be altered by therapeutic interventions. However, testing can have an important impact on such a person's perception of the future and his or her interpersonal relationships and plans for reproduction. Therefore, the decision to pursue testing rests on the individual's belief system and values. On the other hand, a more directive approach is appropriate when a condition can be treated. In a family withFAP(associated with mutations in the *APC* gene), colon cancer screening and prophylactic colectomy should be recommended for known *APC* mutation carriers. The counselor and clinician following this family must ensure that the at-risk individuals have access to the resources necessary to adhere to these recommendations.

Genetic education is central to an individual's ability to make an informed decision regarding testing options and treatment. Although genetic counselors represent one source of genetic education, other health care providers also need to contribute to patient education. Patients at risk for genetic disease should understand fundamental medical genetic principles and terminology relevant to their situation. This includes the concept of genes, how they are transmitted, and how they confer hereditary disease risk. An adequate knowledge of patterns of inheritance will allow patients to understand the probability of disease risk for themselves and other family members. It is also important to impart the concepts of disease penetrance and expression. For complex genetic disorders, asymptomatic patients should be advised that a positive test result does not always translate into future disease development. In addition, the role of nongenetic factors, such as environmental exposures, must be discussed in the context of multifactorial disease risk and disease prevention. Finally, patients should understand the natural history of the disease as well as the potential options for intervention, including screening, prevention, and -- in certain circumstances -- pharmacologic treatment or prophylactic surgery.

THERAPEUTIC INTERVENTIONS BASED ON GENETIC SUSCEPTIBILITY TO DISEASE

Specific treatments are now available for an increasing number of genetic disorders, whether identified through population-based screening or directed testing (Table 68-3). A number of metabolic disorders fall into this group. For example, the complications of phenylketonuria can be mitigated by recognizing the disease early and avoiding foods that contain phenylalanine (Chap. 352). Similar principles apply to maple syrup urine disease (Chap. 352) and galactosemia (Chap. 350). Children with 21-hydroxylase deficiency present with adrenal insufficiency, usually within the first few weeks of life (Chap. 331). Because of the block in cortisol synthesis, the adrenal steroid precursors are shunted into the androgen pathway, causing ambiguous genitalia in females and premature virilization in males. In this disorder, treatment with glucocorticoid and mineralocorticoid not only corrects the hormone deficits but is also required to suppress ACTH overproduction, which otherwise worsens virilization.

Although the strategies for therapeutic interventions are best developed for childhood genetic diseases, these principles are gradually making their way into the diagnosis and management of adult-onset disorders. Hereditary hemochromatosis illustrates many of the issues raised by the potential availability of genetic screening in the adult population. For instance, it is relatively common (approximately 1 in 200 individuals of northern European descent are homozygous), and its complications are potentially preventable through phlebotomy (Chap. 345). The recent identification of the *HFE* gene, mutations of which are associated with this syndrome, has sparked interest in the use of DNA-based testing for presymptomatic diagnosis of the disorder. However, up to one-third of individuals who are homozygous for the *HFE* mutation do not have evidence of iron overload. Consequently, in the absence of a positive family history, current recommendations are phenotypic screening for evidence of iron overload followed by genetic testing. Whether genetic screening for hemochromatosis will someday be coupled to assessment of phenotypic expression awaits further studies. In contrast to the issue of population screening, it is important to test and counsel other family members when the diagnosis of hemochromatosis has been made in a proband. Testing allows the physician to exclude family members who are not at risk. It also permits presymptomatic detection of iron overload and the institution of treatment (phlebotomy) before the development of organ damange.

Preventive measures and therapeutic interventions are not restricted to metabolic disorders. Identification of familial forms of long QT syndrome, associated with ventricular arrythmias, allows early electrocardiographic testing and the use of prophylactic antiarrythmic therapy (Chap. 230). Individuals with familial hypertrophic cardiomyopathy can be screened by ultrasound, treated with beta blockers or other drugs, and counseled about the importance of avoiding strenuous exercise and dehydration (Chap. 238). Likewise, individuals with Marfan syndrome can be treated with beta blockers and monitored for the development of aortic aneurysms (Chap. 247). Individuals witha$_1$antitrypsin deficiency can be strongly counseled to avoid cigarette smoking and exposure to environmental pulmonary and hepatotoxins. Various host genes influence the pathogenesis of certain infectious diseases in humans, including HIV (Chap. 309). The factor V Leiden allele increases risk of thrombosis (Chap. 62). Approximately 3% of the worldwide population is heterozygous for this mutation. Moreover, it is found in up to 25% of patients with recurrent deep venous thrombosis or pulmonary embolism. Women who are heterozygous or homozygous for this allele should therefore avoid the use of oral contraceptives and receive heparin prophylaxis after surgery or trauma.

The field of pharmacogenetics seeks to identify genes that alter drug metabolism or confer susceptibility to toxic drug reactions (Chap. 71). Examples include succinylcholine sensitivity, malignant hyperthermia, the porphyrias, and glucose-6-phosphase dehydrogenase (G6PD) deficiency.

As noted above, the identification of genes that increase the risk of specific types of neoplasia is rapidly changing the management of many cancers. Identifying family members with mutations that predispose toFAP or hereditary nonpolyposis colon cancer (HNPCC) can lead to recommendations of early screening by colonoscopy or prophylactic surgery (Chap. 90). Similar principles apply to familial forms of melanoma, basal cell carcinoma, and breast cancer. It should be recognized, however, that most cancers harbor several distinct genetic abnormalities by the time they acquire invasive or metastatic potential (Chaps. 81 and82). Consequently, the major impact of genetic testing in these cases is to allow more intensive clinical screening, as it remains very challenging to predict disease penetrance or the clinical course of these diseases.

Although genetic diagnosis of these and other disorders is only beginning to be used in the clinical setting, susceptibility testing holds the promise of allowing earlier and more targeted interventions that can reduce the morbidity and mortality associated with these disorders. We can expect the availability of genetic tests to expand rapidly. A critical challenge for physicians and other health care providers is to keep pace with these advances in genetic medicine and to implement testing judiciously. Meeting this goal will enhance patient care through adequate counseling, directed testing, and appropriate interventions, with the ultimate objective being the reduction of morbidity and mortality from genetic diseases.

(Bibliography omitted in Palm version)

Back to Table of Contents

69. GENE THERAPY - *Mark A. Kay*, *David W. Russell*

Gene therapy is generally defined as *the delivery of nucleic acids to alter or prevent a pathologic process*. Although initially considered primarily in the context of inherited monogenic disorders, gene therapy is now recognized as a potential treatment strategy for a wide range of acquired disorders, such as cancer, neurodegenerative diseases, and infections. Gene therapy has been used in several hundred protocols. Despite early hopes that gene therapy might be quickly incorporated into medical practice, it has yielded limited success to date and remains an investigational treatment. The technical challenges associated with gene therapy are formidable, but with steady progress in vector development, definitive therapeutic milestones may soon be realized for selected disorders.

GENERAL APPROACHES TO GENE THERAPY

FORMS OF GENE THERAPY

Gene therapy can be used, in principle, to modify all cells in the body, including the germ line. *Germ-line gene therapy* would allow transmission of the modified genetic information to the next generation and is not currently accepted as an appropriate therapeutic approach. For ethical reasons, opposition to germ-line gene therapy is likely to continue in the foreseeable future. *Somatic gene therapy* refers to modification of the somatic, differentiated cells of the body. It is used in an effort to correct inherited diseases such as cystic fibrosis and acquired disorders such as rheumatoid arthritis or malignancies.

Whole-organ transplantation (e.g., bone marrow, liver, and kidney) has been used for years as a strategy to replace the function of defective genes. The idea of using nonautologous cell transplantation has been revisited in the past few years, and with the advent of human embryonic stem cells, similar transplantation strategies are being considered for a variety of disorders once considered prime targets for gene therapy. Thus, the interface between gene therapy and *cell transplantation* is complementary.

The production of *genetically engineered proteins* is another area closely aligned with gene therapy. The cloning of human genes allows proteins to be produced in unlimited quantities and free of the potential contaminants associated with their purification from natural sources such as plasma. Moreover, recombinant DNA technology allows these proteins to be modified in ways that can enhance their therapeutic benefit. Examples of recombinant proteins are listed in Table 69-1. These include: (1) hormones such as insulin, growth hormone, and gonadotropins; (2) factors used to enhance blood cell production including erythropoietin, granulocyte colony stimulating factor (CSF), granulocyte-macrophage CSF, and thrombopoietin; (3) interferons (IFNs) and interleukins (ILs) used to treat a variety of autoimmune, infectious, and neoplastic diseases; (4) clotting factors VIII and IX; (5) thrombolytic agents such as tissue plasminogen activator or the antithrombotic agent hirudin; (6) recombinant antigens used for hepatitis B vaccines; and (7) humanized monoclonal antibodies used for immunosuppression or to treat specific types of malignancy. Although these recombinant proteins are an indirect form of gene therapy, they represent an important outgrowth of genetic medicine.

Long-Term versus Transient Gene Delivery Strategies for Gene Therapy The goals of gene therapy vary depending on the nature of the disease being treated. For an inherited, monogenic disorder such as hemophilia, the goal is lifelong replacement of the missing gene product. Expression of the missing secreted clotting factor, even at modest levels, might ameliorate the disease or reduce the need for exogenous treatment. For other inherited disorders, such as sickle cell anemia, strategies for gene replacement are more demanding. In this case, there is a requirement for cell-specific and exquisitely regulated expression of the transferred gene, and one is still faced with endogenous expression of the mutant form ofb-globin. Long-term gene delivery strategies typically involve direct modification of host chromosomal sequences, which allows normal inheritance and stability of the delivered gene.

A growing use of gene therapy involves transient gene expression to treat a variety of diseases (Table 69-2). Gene therapy is now being considered or used for (1) killing cancer cells; (2) providing chemoprotection to normal cells; (3) preventing coronary restenosis or enhancing vascularization; (4) providing DNA-based immunization (e.g., viral DNA), in which the injection of DNA results in antigen expression and generation of an immune response; and (5) impairing viral replication. In cancer therapy, for example, the requirements for long-term gene expression and the level of gene expression are not as stringent as for gene replacement. Rather, the challenge of many cancer-based gene therapies is to achieve highly efficient gene transfer into cancer cells without expression in surrounding normal cells.

EX VIVO VERSUS IN VIVO ADMINISTRATION OF GENE THERAPY

Gene therapy has been administered ex vivo for conditions in which cells can be readily harvested, manipulated in tissue culture, and then reintroduced into the patient. The ex vivo approach is potentially applicable to a variety of hematologic or immune deficiency disorders. For example, in adenosine deaminase (ADA) deficiency, the missing ADA gene has been integrated into the patient's T lymphocytes, which are then reintroduced after genetic manipulation. In familial hypercholesterolemia, an analogus approach has been used for ex vivo treatment of hepatocytes. After surgical resection of liver tissue, the low-density lipoprotein (LDL) receptor gene is inserted into hepatocytes in cell culture, and the modified cells are infused back through the portal vein. Another form of ex vivo gene therapy involves the treatment of saphenous veins with oligonucleotides designed to block vascular smooth-muscle cell proliferation before using the tissue for coronary artery bypass graft.

In vivo approaches to gene therapy vary depending on the nature of the disease. In cystic fibrosis, an aerosol has been used to administer viral vectors to the lung. Factor IX expression has been achieved by introduction of the gene into muscle. Tumors have been injected directly with viral vectors expressing cytotoxic genes or immunotherapies. A major advantage of in vivo approaches is the potential to target cells that cannot easily be removed from the patient.

GENE TRANSFER VECTORS

A vector, or vehicle, is required to transfer a gene to an appropriate cell. When the goal is to add a gene (often called a *transgene*) to supply a function not present in the recipient cell, the vector typically contains DNA sequences encoding a therapeutic protein under the control of transcriptional regulatory elements necessary for gene expression (Fig. 69-1). The gene is expressed from an ectopic location as opposed to its normal chromosomal position. An alternative strategy is to correct mutant genes at their normal chromosomal location through *gene targeting*. Although this represents an ideal approach for many genetic diseases, it is far more technically demanding, and therapeutic gene-targeting efficiencies are difficult to achieve currently.

Two major classes of vectors are used for transferring nucleic acids into cells for the purposes of gene therapy: viral and nonviral vectors. *Viral vectors* have been genetically engineered so that the viruses transfer exogenous (therapeutic) nucleic acids into cells through a process called transduction. *Nonviral vectors* consist of nucleic acids that are typically complexed with other chemicals to facilitate gene transfer. Although nonviral vectors offer improved safety by avoiding viral components, their gene transfer efficiencies are generally much lower than those of viral vectors. The major vector systems currently used in clinical trials or under development are summarized inTable 69-3.

Vectors that integrate into host chromosomes are considered ideal for lifelong gene expression, whereas episomal (unintegrated) vectors are preferred for transient gene delivery. The host range of a particular vector, in combination with the mode of delivery, will determine which cell types can be genetically modified. Also, it may be possible to alter the natural tropism of viral capsids or envelopes, as well as of nonviral vector complexes, to limit transduction to particular cell types. Transcriptional control elements, such as promoters and enhancers, can be used to regulate expression of the transgene. In some cases, the inclusion of special regulatory elements may allow gene expression to be controlled by the administration of pharmacologic agents that specifically switch the regulatory elements on or off. Many types of vectors, each exhibiting a unique set of properties, will ultimately be needed for safe and effective gene therapy of different diseases.

RETROVIRAL VECTORS

Retroviral vectors based on murine leukemia viruses (MLVs) were the first vectors used in clinical gene transfer protocols and illustrate many of the principles of viral-mediated gene therapy. Typically, the vector genome consists of a transgene cassette placed between the two *cis*-acting long terminal repeats (LTRs) of the viral genome. Viral coding regions are removed from the vector genome to allow the insertion of the therapeutic gene and to prevent viral replication and toxicity to the host. The packaging capacity of MLV vectors is approximately 8 kb (including the LTRs), enough to accommodate most therapeutic cDNAs. The viral gene products needed for vector production and packaging are supplied by helper virus expression constructs.

The envelope protein of the virus interacts with specific cellular receptors to allow cellular entry of the core proteins and vector genome. After transduction, the vector gene integrates at random locations in host chromosomes and replicates with the host chromosome during cell proliferation, offering the potential for lifelong transgene expression. *Chromosomal integration* may also cause insertional mutagenesis, but no clinically relevant consequences of such events have been observed to date with replication-incompetent retroviral vectors.

AlthoughMLVvectors are widely used, primarily because of their ability to integrate and simple production requirements, they still have significant disadvantages. Complement-mediated vector particle inactivation has largely limited their use to ex vivo applications. MLV vectors also require cell division for transduction because the vector genome can only enter the cell nucleus during mitosis, when the nuclear membrane breaks down. Since the majority of cells present in the human body are nondividing, this severely limits the potential uses of MLV-based vectors.

Promising new retroviral vectors are based on complex retroviruses (rather than the relatively simple oncoviruses such asMLV). Both lentiviral vectors and spumaviral vectors have broad host ranges, improved transduction of nondividing cells, and may function well in vivo. Although lentiviral vector production systems can be designed to eliminate the potential for replication-competent retrovirus contamination, and these vectors appear safe from a virologic standpoint, the stigma associated with vectors based on human pathogens such as HIV may limit their acceptance.

ADENOVIRAL VECTORS

Adenoviral-mediated gene expression is not lifelong, making it well suited for applications that require transient gene expression. In contrast to retroviruses, which integrate into the host genome, the adenoviral vector genome remains *episomal*, or extrachromosomal. Adenoviral vectors typically are derived from serotypes 2 or 5 and contain double-stranded DNA genomes. Wild-type adenovirus encodes over 50 peptides on overlapping gene fragments from both DNA strands. Nonessential viral genes have been removed to make room for up to 8 kb of exogenous DNA. The vector containing the therapeutic transgene is amplified in a cell line that supplies viral proteins needed for replication and packaging. Recombinant adenoviruses can be generated at high titers in the range of 10_{13} to 10_{14}particles per milliliter, a feature that is important for efficient in vivo gene therapy.

In clinical practice, adenoviral vectors are limited by relatively short durations of expression (usually several weeks) and by the synthesis of remaining cytotoxic or antigenic viral proteins that can cause an acute inflammatory response and/or a robust cellular immune response. These features of the virus appear to have contributed to hepatic failure and death in an individual receiving adenoviral-based gene therapy for ornithine transcarbamylase deficiency. On the other hand, the inflammatory properties of adenoviruses may enhance their efficacy in cancer trials, as discussed below. Several methods that eliminate the remaining viral genes have been devised recently, and these "gutted" vectors appear to exhibit much reduced toxicity and inflammatory properties in comparison to previous generations of adenoviral vectors.

ADENO-ASSOCIATED VIRUS VECTORS

Adeno-associated viruses (AAV) are parvoviruses that normally require a helper virus, such as adenovirus, to mediate a productive infection. A major limitation of the AAV vector is its relatively small packaging capacity, which restricts the size of exogenous

DNA to about 4.5 kb. AAV vectors have been shown to transduce cells both through episomal transgene expression and by chromosomal integration. They can also be used to modify homologous chromosomal sequences through gene targeting, which is an important strategy for the correction of genetic mutations.

Since theAAVvector genome lacks viral coding sequences, the vector itself causes little immune or inflammatory response (except for the generation of neutralizing antibodies that may limit readministration). The vector particle can be delivered to many different organs [e.g., the central nervous system (CNS), liver, lung, and muscle] by in vivo administration, and AAV vectors have been found to transduce nondividing cells efficiently. Clinical trials using AAV for the treatment of cystic fibrosis and hemophilia are underway.

OTHER VIRAL VECTORS

Many other viruses are under development as vector systems, such as herpesviruses, double-stranded RNA viruses, autonomous parvoviruses, and papovaviruses. Hybrid viral vectors that use components from more than one virus are also under development, offering the potential to combine the most desirable properties of different vectors. Ultimately, vectors will be developed that utilize both viral components and synthetic functions, allowing vectors to be custom designed.

NONVIRAL VECTORS

Nonviral vectors usually consist of DNA complexes with lipids, carbohydrates, proteins, and/or other synthetic chemicals to facilitate delivery or increase vector stability. Gene delivery with naked DNA is also possible but is relatively inefficient. Nonviral vectors are desirable because they eliminate the risk of viral contamination and can be produced under more controlled conditions. Their major disadvantage is low gene transfer rates, at least in relation to most viral vector systems. Because the vectors do not integrate and the majority of DNA molecules that enter the cell are rapidly degraded, only transient gene expression is possible. Examples of applications that may be well suited to gene delivery with nonviral vectors include vaccination with specific antigen genes, ex vivo treatment of vessels used for coronary artery bypass graft, and perhaps transient immune modulation for the treatment of cancer.

Nonviral vectors composed of RNA or modified nucleotides may also prove useful, if problems associated with effective delivery can be addressed. For example, antisense oligonucleotides can be used to inhibit gene expression by pairing with complementary mRNA molecules. Antisense oligonucleotides might be used, for example, to block the expression of cell cycle proteins or cytokines. Oligonucleotides with binding sites for specific DNA- or RNA-binding proteins may be designed to function as decoys that inhibit protein function. Chimeric RNA/DNA oligonucleotides have been used to introduce specific genetic modifications through gene targeting. In animal studies, this novel approach has been shown to work efficiently in the liver, and it may be used soon for clinical trials of uridine diphosphate-glucuronosyltransferase deficiency.

GENE THERAPY FOR SELECTED DISEASE CATEGORIES

LIVER AND GASTROINTESTINAL TRACT

The liver has been studied intensively as a target organ for gene therapy, in part because of the many genetic diseases amenable to gene replacement in hepatocytes. In addition, because of the regenerative capacity of the hepatocyte, integration of a vector offers the possibility of lifelong gene expression. The first hepatic gene therapy attempted in humans involved the ex vivo transplant of autologous hepatocytes transduced in culture by retroviral vectors encoding theLDLreceptor as a treatment for familial hypercholesterolemia. Due to the low levels (nontherapeutic) of gene expression observed and the labor-intensive nature of the ex vivo transduction process, current strategies are geared towards in vivo gene transfer.MLV-based retroviral vectors require cell division for transduction of host cells. The use of growth factors (rather than partial hepatectomy) to stimulate hepatocellular replication can enhance the efficacy of MLV vectors. Alternative retroviral vectors based on lentiviruses may overcome these obstacles.

Adenovirus vectors are very effective at gene transfer into the liver, thus allowing for the transduction of a majority of hepatocytes. However, the early generations of these vectors exhibited dose-dependent toxicity, inflammation, and immunogenicity that limited the persistence of transgene expression. Adenoviral vectors devoid of all viral genes have been shown to transduce hepatocytes efficiently with reduced toxicity or immunogenicity. In rodents, transgene expression appears to be lifelong, but in nonhuman primates, expression declines by 90% over a 2-year period.

AAVvectors stably integrate their proviral DNA into hepatocytes in vivo with no apparent toxicity. This vector has been used in preclinical liver gene therapy studies to cure mice with hemophilia B and to partially correct the defect in dogs with hemophilia B. Hepatocyte gene transfer by AAV vectors is likely to be useful for treating a variety of metabolic diseases, such as urea cycle disorders, aminoacidopathies, disorders of carbohydrate metabolism, and lysosomal storage diseases. This strategy is particularly applicable when expression of the transgene in a relatively small percentage of hepatocytes is of therapeutic benefit.

The intestinal tract offers a large cell mass for gene therapy and is accessible by oral administration of vectors. Although a number of vectors have been tested in the gastrointestinal tract, major limitations still exist, such as the rapid turnover of the intestinal epithelial cells, the inability to target the stem or early progenitor cells, and the effects of digestive secretions on vector delivery. If these obstacles can be overcome, many diseases may be amenable to treatment by intestinal gene transfer.

THE HEMATOPOIETIC SYSTEM

Clinical trials targeting hematopoietic cells have almost exclusively taken an ex vivo approach, with gene transfer occurring outside the body, followed by reinfusion of transduced cells. The most common cellular targets are lymphocytes and hematopoietic stem cells. Because stem cells can reconstitute the entire hematopoietic system, ex vivo genetic manipulation and autologous transplantation of bone marrow or peripheral blood stem cells represents a powerful therapeutic paradigm. Long-term gene transfer to hematopoietic stem cells requires the use of integrating viral vectors to ensure that the transgene replicates with the chromosomes during the extensive cellular proliferation that occurs during hematopoiesis. Thus, most stem cell gene therapy trials have used retroviral vectors. In part, because stem cells are not actively dividing, transduction rates have been too low to be therapeutic. Despite this, recent success in treating X-linked severe combined immunodeficiency with MLV vectors encoding the cytokine receptor gamma common chain demonstrates the feasibility of this approach. The use of different envelope proteins and vectors based on lentiviruses or spumaviruses may yield greater success in stem cell transduction.

Gene transfer to hematopoietic stem cells could be used to treat diseases of erythroid, myeloid, and lymphoid cells as well as platelet disorders (since the stem cell ultimately differentiates into all these cell types). An important aspect of this approach is the level of myeloablation required to allow engraftment of ex vivo-modified stem cells. An optimal therapy would minimize myeloablation and its associated toxicity, perhaps by delivering a gene that confers a selective advantage to transduced stem cells along with the therapeutic transgene. Many of the clinical trials to date have involved genes that confer increased resistance to chemotherapeutic agents used in treating cancer. An alternative strategy is to apply gene therapy to disorders in which expression is not required in all cells to prevent clinical disease. In chronic granulomatous disease, in which there is failure of granulocytes and monocytes to generate the hydrogen peroxide needed for bacterial killing, it may be feasible to achieve a threshold level of normal cells to respond to infections. Another major area of stem cell gene therapy research is the treatment of hemoglobinopathies with vectors expressing globin genes. Although clinical trials have not begun, these efforts have helped define the problems that must be solved for successful stem cell gene therapy, especially with regard to regulating transgene expression levels. Globin gene replacement is a daunting task, since the gene must be expressed at high levels in a select subset of erythroid cells, thus requiring the inclusion of a variety of transcriptional control elements in the vector (many of which have resulted in vector instability and low viral titers).

Lymphocyte gene transfer has the potential to treat genetic immunodeficiencies and to modulate immune functions.ADAdeficiency was one of the first diseases to be treated with gene therapy and employed ex vivo transduction of lymphocytes. Although this early clinical trial represented an important milestone in gene therapy, the assessment of a therapeutic effect has been complicated by concomitant treatment with the ADA protein. Replacement of the ADA gene also confers a selective advantage to ADA-deficient lymphocytes, a property that would not be generally applicable to lymphocyte gene transfer. Still, as lymphocyte transduction protocols improve, many treatments may be developed with this cellular target, especially for acquired diseases. Genetic manipulations with antibody or cytokine genes offer a multitude of possible treatments for infectious and autoimmune diseases, as well as for cancer.

PULMONARY SYSTEM

Gene therapy of the lung has concentrated primarily on treatments for cystic fibrosis. The gene for cystic fibrosis, *CFTR*, was cloned in 1989; by 1993 the first trials using adenovirus vectors were attempted in the nasal epithelium and airway. Though no clinical efficacy was demonstrated, these studies underscored the need to develop defined clinical endpoints for future trials. In addition to the adenovirus trials, clinical trials with nonviral liposome andAAVvectors also failed to demonstrate therapeutic effects. Difficulties encountered in pulmonary gene therapy include poor vector delivery due to respiratory secretions, lack of accessible vector receptors on the exposed cellular surface, transient transgene expression due to turnover of the respiratory epithelium, and vector-induced inflammation and pneumonitis. It also remains unclear which type of lung cells should be targeted with the *CFTR* gene to result in clinical benefit. As efficient and safe vectors for lung gene transfer are developed, treatments for acquired disorders such as chronic obstructive pulmonary disease may also be possible.

NERVOUS SYSTEM AND RETINA

Gene transfer to the nervous system will be important for the treatment of many inherited and acquired neurologic diseases. Depending on the disorder, both glial cells and neurons may be appropriate cellular targets. Several vectors have been shown to transduce cells (including neurons) in animals after in vivo injection into the brain. Gene transfer is typically localized near the site of injection, a feature that is desirable for disorders such as Parkinson's disease, where transduction in the striatum could allow selective expression of genes involved in dopamine synthesis. Vector delivery throughout the nervous system will be more difficult to achieve, making gene therapy of global neurologic disorders problematic. The possibility of delivering neurotrophic factors to theCNS is a potential strategy for the treatment of neurodegenerative diseases, such as amyotrophic lateral sclerosis, and for facilitating recovery after spinal cord injury. Genetically modified stem cells have also been suggested as a strategy for the treatment of neurodegenerative disorders, but studies of these cells are at very early stages.

Retinitis pigmentosa (RP), a common cause of blindness from retinal degeneration, represents one class of disorders that stands to benefit from gene therapy. Though many different gene defects may cause RP, treatment of a form caused by mutation in the rod photoreceptor cGMP phosphodiesterase b-subunit gene has been studied in animal models by intraocular gene addition with adenovirus, lentivirus, andAAVvectors. In a different type of RP, progression of an autosomal dominant form of the disease was significantly slowed in a transgenic rat model using AAV vectors that express *ribozymes* (RNA enzymes) against the mutant mRNA. Ribozymes are catalytically active RNA molecules that specifically anneal to, and cleave, other RNA sequences, resulting in their selective degradation. This strategy may be useful for the treatment of other genetic diseases caused by dominant mutations. Viral vectors that express cytotoxic genes have been used to treatCNStumors (see below).

MUSCULOSKELETAL SYSTEM

Efficient gene transfer to myotubes through intramuscular injection has been demonstrated with adenoviral vectors,AAVvectors, and lentiviral vectors. In the case of AAV vectors, long-term transgene expression in muscles has been observed in several animal species. The treatment of the muscular dystrophies will require efficient delivery to multiple muscle groups throughout the body, and vector delivery to all the necessary locations remains a formidable task. In the case of Duchenne muscular dystrophy, the large size of the dystrophin cDNA also poses technical challenges in vector design. In contrast to primary diseases of muscle, success is more likely in muscle gene therapy trials aimed at the synthesis of secreted proteins, such as clotting factors. The large tissue mass and accessibility of muscle make it particularly attractive for these applications.

CARDIOVASCULAR SYSTEM

The cardiovascular system (including the peripheral vasculature) has become an important target for gene therapy. Vascular wall gene delivery is being studied as a way to inhibit smooth-muscle cell proliferation and prevent restenosis. The transgenes used in this application are designed to interfere with the cell cycle or induce apoptosis in smooth-muscle cells. This approach is especially attractive if the vector can be delivered during angioplasty.

Other gene therapy strategies are used to promote the vascularization of tissues. There is some evidence for a therapeutic response in patients with critical limb ischemia due to poor peripheral vascularization who received intramuscular injection of naked DNA vectors encoding the vascular endothelial growth factor (VEGF) gene. Small amounts of the protein are secreted from the muscle, resulting in collateral vessel development that can reverse the ischemia. Patients who were expected to require limb amputation were spared this procedure after gene therapy. Similar clinical trials with the VEGF gene using nonviral and viral vectors are being studied in the context of myocardial ischemia.

CANCER

Most of the gene therapy clinical trials to date have been aimed at the treatment of cancer. One approach uses gene therapy with cytokine or neoantigen genes to increase tumor immunogenicity. The vector is usually injected directly into the tumor, and there is some evidence that once the immune system is stimulated, nontransduced tumor cells may also be eliminated by the immune system. In melanoma, for example, cells have been genetically altered to express mismatched histocompatibility antigens or cytokines such as tumor necrosis factor,IFN-g, or IL-2 in an effort to stimulate an immune response. Another approach involves the delivery of genes to tumor cells that convert a prodrug into a cytotoxic compound. Although several strategies are being developed, the most common involves the transfer of the herpes thymidine kinase (TK) gene. The TK enzyme converts gancyclovir into a thymidine analogue that interferes with DNA synthesis and causes cell death. The toxic effects also occur in adjacent nontransduced cells due to the uptake of the toxic analogue; this process is referred to as the *bystander effect*. Vectors designed to replace defective tumor-suppressor proteins with normal versions are being considered for cancer gene therapy, but it is difficult to envision success with this approach unless virtually all the tumor cells are genetically modified. Genes that control tumor growth when expressed in nontumor cells may also be effective in cancer gene therapy. The delivery of gene products that interfere with tumor angiogenesis best exemplifies this approach. Finally, lytic viral vectors that selectively replicate and kill malignant cells are being developed. One example is an adenovirus designed to replicate in cells deficient in p53, a tumor-suppressor protein that is mutated in many different cancers.

These and other related strategies have shown efficacy in animal models when tumor cells are transplanted into various anatomic locations. These models do not always emulate the natural processes that result in bonafide cancers, in part due to the tumor cells not being truly autologous in origin. Although there have been encouraging results in some of the clinical trials performed to date, efficacy has not been demonstrated definitively. In many cases the patient populations studied had advanced malignancies, and more informative results might be obtained at earlier disease stages. Still, cancer gene therapy has a promising future, and ongoing research involving tumor-specific antigens, angiogenesis, cell cycle control, and apoptosis are all likely to lead to new gene therapy approaches. Improvements in vector targeting of tumor cells and in the development of tumor-specific gene expression will also enhance future therapeutic approaches.

COAGULOPATHIES

Inherited coagulopathies, especially hemophilia A and B, are promising areas of gene therapy research because even a low level of coagulation factor reconstitution can potentially benefit a patient with a severe phenotype. Although the liver is the major site of synthesis for factors VIII and IX, several other tissues may support factor synthesis and secretion into the bloodstream.AAVvectors, in particular, have shown prolonged therapeutic effects in animal models of hemophilia B when delivered to muscle or liver. Hemophilia A has been more difficult to treat, due to the larger cDNA (which approaches the packaging capacity of AAV vectors) and the requirement that expressing cells deliver the protein directly into the intravascular space. Other coagulopathies, such as factor X deficiency, are also good candidates for gene therapy. Various approaches to the treatment of hemophilia are currently in the early stages of clinical trials.

INFECTIOUS DISEASES

Most infectious pathogens studied as targets for gene therapy are viral (particularly HIV), in part because they replicate inside human cells. One approach involves introducing inhibitory versions of essential viral proteins that disrupt the viral life cycle even in the presence of their normal counterparts (e.g., by disrupting the packaging of virions). Another strategy involves the expression of proteins, peptides, or even RNA transcripts that function as decoys by binding to other proteins required for viral replication and preventing them from acting on their normal viral target sites. The cellular proteins required for a pathogen's life cycle, such as receptor molecules used for viral entry, or specific proteins used for adherence of the pathogen, can also be manipulated through gene therapy; this tactic could also be applied to nonviral pathogens. Ribozymes can be engineered to cleave specific viral transcripts, thereby blocking expression of viral gene products or destroying viral genomes (in the case of RNA viruses). Because ribozymes can, in principle, be engineered to attack any RNA transcript, they are being considered as treatments for many different viral diseases. Clinical trials have begun using ribozymes to block HIV infection. Antisense oligonucleotides can also be used to interfere with viral or cellular nucleic acid sequences. Other gene therapy approaches that may be applied to infectious diseases include vaccination with specific antigens and manipulation of the immune system to enhance the clearance of pathogens.

SUMMARY

The gene can be thought of as a new pharmaceutical agent in the armamentarium used to treat disease. The availability of cloned genes has already yielded a large array of recombinant proteins for clinical use. The limited success of gene therapy to date is due primarily to difficulties inherent in the efficient and safe delivery of genes to their appropriate target cells. The field is still in its infancy, and many of the technical problems are likely to be solved by advances in vector design.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART FOUR -CLINICAL PHARMACOLOGY

70. PRINCIPLES OF DRUG THERAPY - *John A. Oates*, *Dan M. Roden*, *Grant R. Wilkinson*

Safe and effective drug therapy requires that drugs be delivered to their molecular targets in tissues at concentrations within the range that yields efficacy without toxicity. Variability in drug effects between individual patients can thus be attributed, in part, to differences in the processes that determine these concentrations, i.e., drug disposition involving absorption, distribution, and elimination. *Pharmacokinetics* provides a quantitative description of these processes, i.e., what the body does to the drug. In addition, variability in response may also reflect differences in the drug-target interaction that affect the concentration-related effects of the drug on the body, termed *pharmacodynamics*. This chapter reviews the principles of pharmacokinetics and pharmacodynamics and their application to optimizing therapeutic regimens.

CLINICAL PHARMACOKINETICS

PLASMA LEVELS AFTER A SINGLE DOSE

The levels of lidocaine in plasma following intravenous administration decline in two phases, as illustrated in Fig. 70-1; such a biphasic decline is typical for many drugs. Immediately after rapid injection, essentially all of the drug is in the plasma or central compartment, and the high initial plasma level reflects the confinement of the drug to this small volume. The drug is then transferred into an extravascular or peripheral compartment during a period called the *distribution phase*. For lidocaine, the distribution phase is virtually complete within 30 min. A phase of slower decline, the *equilibrium phase*, then occurs. During this phase, the drug levels in the plasma and in tissues change in parallel. Although this disposition profile is common to many drugs given intravenously, the characteristic parameters vary among drugs.

Distribution Phase The pharmacologic effects during the distribution phase depend on whether the concentration of drug at the receptor site is similar to that in the plasma. If that is the case, the pharmacologic effects may be intense during this period because of the high initial levels in the plasma. For example, after a small bolus dose (50 mg) of lidocaine, antiarrhythmic effects may be evident during the early distribution phase but will disappear as levels fall below those that are minimally effective, even before equilibrium between plasma and tissue is reached. Thus, a larger single dose or multiple small doses must be administered to achieve an effect that is sustained into the equilibrium phase. Toxicity resulting from high levels of some drugs during the distribution phase precludes administration of a single intravenous loading dose that will yield therapeutic levels during the equilibrium phase. For example, the administration of the entire loading dose of phenytoin as a single intravenous bolus can cause cardiovascular collapse due to the high levels during the distribution phase. If a loading dose of phenytoin is administered intravenously, it must be given slowly, e.g., at infusion rates of 25 to 50 mg/min. For similar reasons, the intravenous loading dose of many potent drugs that equilibrate rapidly with their receptors is either divided into fractional doses given at intervals or administered by infusion over a similar period.

A dose given orally results in lower plasma levels during the initial period than an intravenous bolus dose that delivers the same amount of drug to the systemic circulation. Because the drug is not absorbed instantly after oral administration and is delivered into the systemic circulation more slowly, much of it has been distributed by the time absorption is complete. Thus, procainamide, which is almost totally absorbed after oral administration, can be given orally as a single 750-mg loading dose with little risk of hypotension; in contrast, loading of this drug by the intravenous route is more safely accomplished by giving the dose in fractions of about 100 mg at 5-min intervals or by slow infusion to avoid hypotension during the distribution phase.

Some drugs are so predictably lethal when infused too rapidly that special precautions should be taken to prevent even the inadvertent occurrence. For example, solutions of potassium for intravenous administration in excess of 20 meq/L should be avoided in all but the most exceptional and carefully monitored circumstances. This minimizes the possibility of cardiac arrests, which can occur as a result of accidental increases in infusion rates of more concentrated solutions.

As these examples illustrate, excessively rapid administration of many drugs can lead to catastrophic consequences that result from high concentrations in the blood during the distribution phase.

In contrast, for some centrally active drugs, the higher concentration of drug during the distribution phase after intravenous administration is used to advantage. The use of midazolam for "IV sedation," for example, depends upon its rapid uptake by the brain during the distribution phase to produce sedation quickly, with subsequent egress from the brain during the redistribution of the drug as equilibrium is achieved.

Some drugs are distributed slowly to their sites of action during the distribution phase, i.e., concentration at the relevant receptor does not parallel that in plasma early after drug administration. For example, the level of digoxin at the receptor site (and the drug's pharmacologic effect) do not reflect plasma levels during the distribution phase. Digoxin is transported (or bound) to its cardiac receptors slowly by a process that proceeds throughout distribution. Thus, over the distribution phase of several hours, plasma levels fall while the level at the site of action and the pharmacologic effect increase. Only at the end of the distribution phase, when the drug has reached equilibrium with the receptor, does the concentration of digoxin in plasma reflect pharmacologic effect. For this reason, there should be a 6- to 8-h wait after administration before plasma levels of digoxin are measured as a guide to therapy.

Equilibrium Phase After the concentration of drug in plasma has reached a dynamic equilibrium with that in the tissues, the levels in plasma and tissues fall in parallel as the drug is eliminated from the body. Thus, the equilibrium phase is also called the *elimination phase*. During this phase, drug concentrations measured in plasma can provide a useful index of drug levels in tissues.

Most drugs are eliminated as a first-order process. This means that the time required for the plasma level of the drug to fall to one-half the original value (the half-life, $t_{1/2}$) is the same regardless of the point on the plasma level curve at which measurement begins. Another characteristic of the first-order process is that a plot of the logarithm of the plasma concentration versus time is linear. From such a plot (Fig. 70-1), it can be seen that the half-life of lidocaine is 108 min. If the half-life is known, the amount of a dose remaining in the body at any time following administration of a single dose can be calculated. Table 70-1 shows how this amount changes over five successive half-lives.

From a clinical standpoint, elimination is essentially complete when it has reached about 90%. Therefore, for practical purposes, *a first-order elimination process reaches completion after three to four half-lives*.

DRUG ACCUMULATION -- LOADING AND MAINTENANCE DOSES

If a drug is given repeatedly at intervals shorter than the time required to eliminate a dose, both the amount of drug in the body and its pharmacologic effect increase with successive doses until they reach a plateau.Figure 70-2 shows the accumulation of digoxin administered in repeated maintenance doses (without a loading dose). Since the half-life of digoxin is about 1.6 days in a patient with normal renal function, 65% of a digoxin dose remains in the body at the end of 1 day. Thus, the second dose will raise the amount of digoxin in the body (and the average plasma level) to 165% of the level produced by the first dose. Each subsequent dose causes a further increase until a *steady state* is achieved. At that point, the drug dosing rate (bioavailable dose/dosage interval) is equal to the rate of elimination, with the fluctuation between peak and trough plasma levels remaining constant. If the rate of drug delivery is then altered, a new steady state will be attained. Continuous infusion of a drug at constant rate also results in progressive accumulation to a predictable steady state (Fig. 70-2). In this case, the steady-state plasma level (C_{ss}) is equivalent to the average between the peak and trough levels ($C_{avg,ss}$) produced by intermittent administration of the same amount of drug over the same period. For *all* drugs with first-order kinetics, the time required to achieve steady-state levels can be predicted from the half-life, because accumulation is a first-order process with a half-life identical to that for elimination. Thus, accumulation reaches 90% of steady-state levels at the end of three to four half-lives. This is true for either intermittent or continuous dosing (Fig. 70-2).

When a therapeutic effect is required urgently, simply administering the maintenance dose of a drug with a long half-life results in an unacceptable delay in reaching steady-state levels of the drug and its intended effect. The time required to achieve the desired pharmacologic effect may be shortened by the administration of a *loading dose* -- the amount of drug that will bring the plasma concentration rapidly to the steady-state level. Therefore, if treatment with lidocaine ($t_{1/2}$~ 108 min) were initiated by infusion at only the maintenance dose level, it would take about 4 to 8 h before the drug's maximal effect was achieved. Because ventricular arrhythmias may be life-threatening, it is not reasonable to wait that long to achieve an effective steady-state drug level. Accordingly, it would be appropriate to administer one or more loading doses at the onset of therapy, together with infusion at the rate of the maintenance dose.

Loading may be accomplished by the administration of a single loading dose or, if that would create a risk of toxicity, by the administration of the loading dose in a series of fractions of that dose over time. The latter approach is particularly appropriate for most drugs that have a low therapeutic index. A divided loading dose strategy or the administration of the loading dose in a slow intravenous infusion are particularly advisable when the drug is given intravenously. Thus, in the case of lidocaine, a common regimen is to administer an initial intravenous bolus of 1 mg/kg followed by up to three additional bolus injections of 0.5 mg/kg every 8 to 10 min as necessary, and a maintenance infusion of 2 mg/min.

Regardless of the size of the loading dose, *after maintenance therapy has been given for three to four half-lives, the amount of drug in the body is determined by the maintenance dose*. The independence of the steady state plasma levels from the loading dose is illustrated in Fig. 70-2, which shows that the elimination of the loading dose would be practically complete after three to four half-lives.

DETERMINANTS OF DRUG DISPOSITION

A number of physiologic and pathophysiologic factors determine the disposition of a drug and hence its pharmacokinetics in an individual patient. The three most important are *clearance*, a measure of the body's ability to eliminate drug; *volume of distribution*, an indication of the extent to which the drug is distributed outside of the blood compartment; and *bioavailability*, the fraction of the administered dose that reaches the systemic circulation. The elimination half-life ($t_{1/2}$), which measures the rate of drug removal from the body, is determined by the relationship between the physiologically determined clearance and volume of distribution.

Clearance The majority of drugs are given over a prolonged period of time according to a multiple dosage regimen, e.g., *x* mg every *y* h. The clinical goal is to maintain the drug's steady-state concentration within the therapeutic range for the individual patient; if the level is too low, reduced efficacy results, whereas if it is too high, the likelihood of adverse effects increases.

Steady state is achieved when the rate of drug elimination equals the rate of drug delivery into the systemic circulation, which, if bioavailability is complete, corresponds to the rate at which the drug dose is administered:

where *Cl* is clearance and $C_{ss}$is the steady-state concentration in plasma. Therefore, *at a given dosing rate the concentration of drug in plasma is completely dependent on its clearance*. If the drug is infused, the concentration remains constant so long as drug delivery continues; however, when the drug is given intermittently, the above relationship is expressed as:

where $C_{avg,ss}$is the *average* value during the dosage interval and is equal to C_{ss}although the *actual* concentrations will be higher or lower at various points during this period. Thus, clearance determines the rate at which a drug should be administered in order to obtain a desired steady-state concentration; stated in a different fashion, steady-state drug levels can be modified either up or down by changing the dosage rate. *The clearance of the vast majority of drugs is constant over the therapeutic range of concentrations*.

Clearance is a measure of the rate at which the organs that eliminate drug from the body remove drug from the blood.

Accordingly, clearance reflects the volume of blood (or plasma) from which the drug would have to be removed per unit time to account for the elimination; it can be related to either total or unbound drug.

Drug elimination generally occurs as a result of metabolism and/or excretion in the liver, kidney, and possibly other organs. Clearance of drug from the body, therefore, reflects the overall contribution of each of these organs, as indicated by their individual rates of elimination normalized to the concentration of drug, and is additive.

Clearance of a drug is usually estimated following administration of an intravenous dose ($dose_{iv}$) and measurement of the resulting total area under the curve (AUC) for the blood or plasma concentration-time curve (AUC_{iv}) from zero to infinite time.

Table 70-2 indicates the marked differences in plasma clearance for some commonly used drugs. Some drugs such as phenobarbital and valproic acid have relatively low values (<10 mL/min), whereas others such as procainamide and lidocaine have much larger clearances (>500 mL/min). Such differences mainly reflect different rate-limiting determinants such as blood flow through the organ(s) of clearance, the extent of binding of the drug to plasma proteins, and the efficiency of the clearance process to remove drug from tissue water by metabolism and/or excretion. The data inTable 70-2 also demonstrate that the relative contributions of the two major routes of elimination, i.e., renal and nonrenal, also vary according to the individual drug. In some cases, such as amikacin, digoxin, gentamicin, lithium, and tobramycin, excretion by the kidneys is predominant. However, with many other drugs (e.g., carbamezipine, lidocaine), nonrenal elimination, which usually reflects metabolism in the liver, is more important.

Volume of Distribution The relationship between the amount of drug in the body and the concentration of drug in the plasma provides a measure of the apparent volume of distribution:

This volume does not, except in a limited number of special cases, reflect an identifiable physiologic volume but corresponds to the virtual volume of fluid that would be required to contain all of the drug in the body at the same concentration as in the plasma. In a typical 70-kg human, plasma volume is about 3 L, blood volume around 5.5 L, and extracellular water outside the vasculature is approximately 42 L. The volume of distribution of drugs that are extensively bound to plasma proteins but are not bound to tissue components approaches plasma volume. However, for most drugs, the volume of distribution is far greater than any physiologic space. For example, the volume of distribution of digoxin is about 700 L, which obviously exceeds total body volume. This simply indicates that digoxin is largely distributed outside the vascular system and hence the proportion of the drug present in the plasma compartment is low.

The volume of distribution may be estimated by back-extrapolation of the plasma concentration-time curve to zero time (C_0) (Fig. 70-1) and dividing this into the dose of drug administered intravenously.

When distribution of the drug is not instantaneous, a more useful volume estimate is based on the area under the plasma-concentration time curve (*AUC*) and the terminal elimination half-life of the drug ($t_{1/2}$).

Extent and Rate of Bioavailability After intravenous administration, all of the administered dose reaches the systemic circulation. In contrast, with all other routes of administration, such as oral, intramuscular, and subcutaneous, there is the potential for only a part of the dose to be absorbed; this fraction (*F*) is termed the drug's *bioavailability*. Lack of complete bioavailability may reflect the inability of the drug to be completely released from the dosage form or vehicle, chemical destruction at the site of administration, incomplete absorption into the vascular system, and metabolism and/or excretion during translocation of the drug from its site of administration to the systemic circulation. In the case of oral dosing, this would include the intestinal epithelium and the liver and lungs. Metabolism and/or excretion by the gastrointestinal tract (F_{GI}) and liver (F_L) are collectively referred to as the *first-pass effect*, since the resulting drug elimination only occurs during drug delivery to the systemic circulation. Loss of drug because of a first-pass effect thus requires that the dosing rate appropriately take into account bioavailability. For example, after oral drug administration:

Bioavailability and the first-pass effect are important with respect to possible differences in drug responsiveness dependent on the route of administration. For example, glyceryl trinitrate has such a large oral first-pass effect (>99%) that systemic concentrations after an oral dose are negligible and no antianginal effect is present. Giving the drug by the sublingual or transdermal routes bypasses the splanchnic organs and allows essentially all of the drug to reach the systemic circulation. For drugs that are efficiently metabolized by the intestinal epithelium and/or liver, i.e., drugs with high extraction ratios in either of these organs, differences in the extent of the first-pass effect between individuals frequently explain variability in drug response. With propranolol, for example, the 15-fold variability in plasma concentrations after the same oral dose results from differences in the individual hepatic extraction ratios reflective of different levels of drug metabolizing activity.

Drug administration by nonintravenous routes involves an absorption process characterized by the plasma level increasing to a maximum value at some time after administration and then declining as the rate of drug elimination exceeds the rate of absorption. Thus, the peak concentration is lower and occurs later than after the same dose given by rapid intravenous injection. The rate of absorption can be an important consideration during the initial period after drug administration, especially for drugs with a narrow *therapeutic index* -- the ratio of the toxic dose to the therapeutic dose. If absorption is too rapid, then the resulting high concentration may cause adverse effects not observed with a more slowly available formulation. At the other extreme, slow absorption is deliberately designed into "slow-release" or "sustained-release" drug formulations in order to maintain plasma concentrations essentially constant during the dosage interval, because the drug's rate of elimination is offset by an equivalent rate of absorption controlled by formulation factors.

Half-Life The organs of elimination can only clear drug from the blood. Thus, the rate at which drug is eliminated from the body is a function of both clearance and the extent to which drug is distributed outside of the vascular compartment. The fraction of total drug in the body that is eliminated in a given time is designated the *fractional elimination constant* (*k*).

For example, if the volume of distribution is 10 L and clearance is 1 L/min, then one-tenth of the drug is eliminated per minute. If *k* is multiplied by the total amount of drug in the body, the actual rate of elimination at any given time can be determined:

This relationship, indicating that the rate of drug elimination is proportional to the drug concentration, describes a first-order, or monoexponential, process. With a few notable exceptions, the elimination of drugs used clinically is first-order.

Half-life ($t_{1/2}$) is the time that it takes for the plasma concentration or amount of drug in the body to decline by 50%. This parameter is related to *k* as follows:

where 0.693 is the natural logarithm of 2(C_0/0.5 C_0).

Because

then

This is an important relationship since it indicates that the rate of drug elimination, reflected by $t_{1/2}$, is dependent on both the efficiency of drug removal (*Cl*) and the drug's volume of distribution (*V*). When *V* remains constant, $t_{1/2}$is a reflection of clearance. Thus, $t_{1/2}$is shortened when rifampin induces the enzymes responsible for a drug's hepatic clearance and is lengthened when a drug's renal clearance is impaired in renal failure. However, when there are concomitant alterations in *V*, as occurs for some drugs in cardiac failure, $t_{1/2}$is not an accurate measure of *Cl* or drug dose.

DESIGNING DOSAGE REGIMENS

Most drugs are administered as part of long-term therapy involving multiple dosing, and it is critical that the dosage regimen be optimized to the individual patient. With some drugs, the desired response, e.g., coagulation or blood pressure, is readily measurable and an individualized dosage regimen can be developed with dosage titration. However, dosage changes should be conservative (<50% for drugs with a low therapeutic index) and not more frequent than every three to four half-lives. Other drugs have little dose-related toxicity so the therapeutic window is large, e.g., penicillins andb-adrenoceptor antagonists. In these situations, effective and prolonged drug effects may be obtained by a "maximal dose" strategy. It is also possible to use this strategy to extend the duration of action of a drug, especially one that is eliminated rapidly from the body. Thus, 75 mg of captopril will result in reduced blood pressure for up to 12 h, even though the elimination half-life of this angiotensin-converting enzyme (ACE) inhibitor is about 2 h; this is because the dose raises the concentration of drug in plasma many times higher than the threshold for its pharmacologic effect.

Determination of the Maintenance Dose The relationship between the maintenance dose and the final steady-state concentration is

Thus, steady-state concentrations can be predictably increased or decreased by appropriate modification of the maintenance dosing rate to achieve a desired target value:

In most cases, this is best achieved by changing the drug dose but not the dosing interval, e.g., by giving 250 mg every 8 h instead of 200 mg every 8 h. However, this approach is acceptable only if the resulting maximum concentration is not toxic and the trough value does not fall below the minimum effective concentration for an undesirable period of time. Alternatively, the steady state may be changed by altering the frequency of intermittent dosing but not the size of each dose. In this case, the magnitude of the fluctuations around the average steady-state level will change -- the shorter the dosing interval, the smaller the difference between peak and trough levels (Fig. 70-3).

The extent of fluctuation is determined by the relationship between the dosing interval and the drug's half-life. For example, if the dosing interval is equal to the drug's half-life, then the fluctuation would be twofold, which is usually a tolerable variation. If a longer dosing interval is used, then the difference between the maximum and minimum plasma levels will be greater (Fig. 70-3). Marked fluctuations increase the likelihood of increased concentration-dependent drug effects early during the dosing interval and possible ineffectiveness at the end of the period, even though the average steady-state drug concentration is the same as that following administration at the same dosing rate but at shorter intervals.

Determination of the Loading Dose The loading dose can be estimated if both the desired plasma level (*C*) and the apparent volume of distribution (*V*) are known:

The loading amount required to achieve steady-state plasma levels can also be determined from the fraction of drug eliminated during the dosing interval and the maintenance dose. For example, if the fraction of digoxin eliminated daily is 35% and the planned maintenance dose is 0.25 mg daily, then the loading dose to achieve steady-state levels would be (100/35) times the maintenance dose, or approximately 0.75 mg. Thus,

NONLINEAR DRUG ELIMINATION

The elimination of some drugs (e.g., phenytoin, salicylate, propafenone, and theophylline) does not follow first-order kinetics because the clearance of these drugs changes as levels in the body fall during elimination or change after alterations in dose. Such elimination is called *concentration-dependent* or *dose-dependent*. Accordingly, the time for the concentration to fall to one-half becomes less as plasma levels fall. (This halving time is not truly a half-life, because the term *half-life* applies to first-order kinetics and is a constant.) When a drug is eliminated by first-order kinetics, the plasma level at steady state is directly related to the amount of the maintenance dose, and a doubling of the dose should lead to doubling of the steady-state plasma level. However, for drugs with dose-dependent kinetics, an increase in the dose may be accompanied by a disproportionate increase in the plasma level. For example, a threefold increase in the dose of propafenone (from 300 mg to 900 mg daily) leads to a tenfold increase in the concentration of propafenone in plasma. Changes in dosage regimens for drugs with dose-dependent kinetics should always be accompanied by surveillance for adverse effects and by measurement of the concentration of the drug in plasma during the time of transition to the new steady-state, if this is feasible.

INDIVIDUALIZATION OF DRUG THERAPY

EFFECTS OF RENAL DISEASE

Whether a drug's dosing rate needs to be modified in patients with renal dysfunction depends on whether the drug is primarily excreted through the kidneys and whether increased drug levels, secondary to impaired renal clearance, will be associated with adverse effects. If both of these factors are present, it is likely that with decreased renal clearance the drug will accumulate to a greater extent than in patients with normal renal function and toxicity will result. This is especially true for drugs with long half-lives and narrow therapeutic indexes (e.g., digoxin). In general, over 60 to 70% of the drug must be renally excreted for dosage modification to be necessary and then only when renal function is less than about 30 to 50% of normal.

The goal of any dosing rate adjustment is to modify the dosing schedule so that the drug's plasma concentration-time profile is as similar to the desired one as possible and that the steady state is reached in about the same time as in a patient with normal renal function. To obtain the desired profile, a modification may be made by decreasing the dose while maintaining the dosage interval, keeping the dose the same but increasing the dosing interval, or a combination of these two approaches.

A drug's renal clearance is proportional to creatinine's clearance (Cl_{CR}), which may be measured directly or estimated from the serum creatinine level (C_{CR}). In men:

For women, the estimate by the above equation should be multiplied by 0.85 to reflect their smaller muscle mass. It should also be noted that this equation is not valid for patients with severe renal insufficiency (C_{CR}< 5 mg/dL) or when renal function is changing rapidly. For simplicity, normal creatinine clearance is conveniently considered to be 100 mL/min. Thus, if the relative contributions of renal and nonrenal elimination to systemic clearance are known, an appropriate modification of the dose in a patient with a given level of insufficiency can be estimated. For example, if the fraction of drug excreted unchanged is 0.9 and creatinine clearance is reduced to 10% of normal, the dosing rate should be reduced to 19% of normal. This modification, which in practice would be rounded to 20%, is based on the fact that nonrenal clearance is unchanged (10% of normal clearance); renal clearance is reduced from 90% to 9% of normal *Cl*; thus systemic clearance is reduced to 10 + 9 = 19% of normal *Cl*.

In clinical practice today, most decisions involving dosing adjustment in patients with renal failure use published tables of recommended dosage reduction or dosing interval lengthening based on the level of renal function indicated by Cl_{CR}, or similar information provided in the drug "label" (Table 70-3). Such modifications are, however, rigorously based on pharmacokinetic principles and are best used when resulting plasma concentration data and clinical observation are used, as necessary, to further optimize therapy for the individual patient.

Often metabolites of the drug are pharmacologically active or cause toxicity, and renal insufficiency may result in their unanticipated accumulation. Meperidine, for example, is extensively metabolized, and renal failure has little effect on its plasma concentration; however, its metabolite, normeperidine, accumulates above its usual level when renal function is impaired. Because normeperidine has greater convulsant activity than meperidine, this accumulation probably accounts for the signs of central nervous system (CNS) excitation, such as irritability, twitching, and seizures, that appear when multiple doses of meperidine are administered to patients with renal disease.

EFFECTS OF LIVER DISEASE

In contrast to the predictable decline in renal clearance of drugs in renal insufficiency, it is not possible to make a general prediction of the effect of liver disease on hepatic biotransformation of drugs (Chap. 292). Rather, the possible effects of hepatitis or cirrhosis range from impaired to increased drug clearance. Even in advanced hepatocellular disease, drug clearance is usually impaired only about two- to fivefold.

The extent of such changes, however, cannot be predicted by the common tests of liver function. Consequently, even when it is suspected that drug elimination is altered in liver disease, there is no quantitative basis on which to adjust the dosage regimen other than assessment of clinical response and the concentration of the drug in plasma.

A drug's oral bioavailability may markedly increase in patients with liver disease. This is particularly the case for those drugs that normally are very well extracted by the liver and thus have a high first-pass effect. In addition, the presence of portacaval shunts may further reduce first-pass elimination and lead to higher drug concentrations reaching the systemic circulation, with the increased risk of adverse effects. For example, the oral availability for high first-pass drugs such as morphine, meperidine, midazolam, and nifedipine is almost doubled in patients with cirrhosis, compared to those with normal liver function. The size of the oral dose of such drugs should, therefore, be reduced in such patients.

EFFECTS OF CIRCULATORY INSUFFICIENCY -- CARDIAC FAILURE AND SHOCK

Under conditions of decreased tissue perfusion, the cardiac output is redistributed to preserve blood flow to the heart and brain at the expense of other tissues (Chap. 38). As a result, the drug may be distributed into a smaller volume of distribution, higher drug concentrations will be present in the plasma, and the tissues that are best perfused will be exposed to these higher concentrations. If either the brain or heart is sensitive to the drug, an alteration in response will occur.

Furthermore, the decreased perfusion of the kidney and liver may impair drug clearance by these organs, directly or indirectly. Thus, in severe congestive heart failure, in hemorrhagic shock, and in cardiogenic shock, the response to the usual dose of drug may be excessive, and dosage modification may be necessary. For example, the clearance of lidocaine is reduced by about 50% in cardiac failure, and therapeutic plasma levels are achieved at infusion rates only about half those usually required. The volume of distribution of lidocaine is also reduced, meaning that the correct loading dose will be smaller than usual. Similar situations are thought to exist for procainamide, theophylline, and possibly quinidine. Unfortunately, predictors of these types of pharmacokinetic alterations are unavailable. Therefore, loading doses should be conservative, and continued therapy should be monitored closely, following clinical indicators of toxicity and plasma levels.

DISEASE-INDUCED CHANGES IN PLASMA BINDING

Many drugs circulate in the plasma partly bound to plasma proteins. Since only the unbound (free) drug can distribute to the site of pharmacologic action, the therapeutic response should be related to the free rather than the total circulating plasma drug concentration. In most cases, the degree of binding is fairly constant across the therapeutic concentration range, so that the total drug levels in plasma can be used as a basis for adjusting dosage without resulting in significant error. However, conditions such as hypoalbuminemia, liver disease, and renal disease can decrease the extent of drug binding, particularly of acidic and neutral drugs so that at any total plasma level there is a greater concentration of free drug than usual and thus a risk of increased response and toxicity. By contrast, conditions that lead to an increased plasma concentration of the acute-phase reactantα_1-acid glycoprotein -- such as myocardial infarction, surgery, neoplastic disease, rheumatoid arthritis, and burns -- cause an increase in drug binding for the basic drugs, e.g., lidocaine and quinidine, that bind to this macromolecule, resulting in an opposite set of effects. The drugs for which changes in binding are important are those that are normally highly bound to plasma proteins (>90%), because a small alteration in the extent of binding produces a large change in the amount of unbound drug.

For many drugs, elimination and distribution are restricted largely to the unbound fraction, and so a decrease in binding leads to an increase in the clearance and distribution of the drug. The relative magnitudes of these changes are such that the net effect is a shortened half-life.

GENETIC DETERMINANTS OF THE RESPONSE TO DRUGS

Knowledge of the enzyme that catalyzes the predominant pathway of metabolism of a drug provides a basis for understanding the therapeutic consequences of variations in the genotype of that enzyme. For a number of the enzymes that metabolize drugs, there are differences (polymorphisms) in catalytic function that are genetically determined (Chap. 65). *A phenotypic trait or its corresponding gene is said to be polymorphic if there is more than one form of the trait or gene in the population.* Polymorphisms in the function of an enzyme are determined by allelic variants in its gene. Increasingly it is possible to individualize treatment based on analysis of the phenotype and/or genotype of the relevant drug-metabolizing enzyme.

Similarly, polymorphism in the receptor for a drug can determine variability in its pharmacologic effect. Genotyping those drug receptors for which polymorphisms influence response may also assist in individualizing drug therapy.

THIOPURINE *S*-METHYLTRANSFERASE (TPMT)

The metabolism of azathioprine provides an example of the importance of genetic polymorphisms of enzymes. Azathioprine exerts its immunosuppressive action via an active metabolite, 6-mercaptopurine. Within target cells, the major pathway of inactivation of 6-mercaptopurine is by TPMT. Genetic polymorphisms in this enzyme lead to differences in inactivation of 6-mercaptopurine, with corresponding vast differences in the sensitivity of patients to the toxic and therapeutic effects of azathioprine. Homozygotes for alleles encoding inactive TPMT (0.3 to 1% of the population) predictably exhibit severe pancytopenia on standard doses of azathioprine. Heterozygotes for alleles encoding enzymes with deficient TPMT activity also experience more bone marrow suppression on "usual" doses. From a therapeutic standpoint it is likely that the bone marrow suppression in the heterozygotes has influenced the empiric determination of the "usual" dose range of azathioprine and that this results in undertreatment of some of the patients homozygous for the allele encoding a TPMT with full catalytic activity. To detect the TPMT deficient phenotype in patients anticipating therapy with azathioprine, the catalytic function of TPMT may be measured in red blood cells (if no blood or red cell transfusions have been given within 2 months). A high concordance of genotype with phenotype (~95%) suggests that analysis of genotype may be used to individualize dosing and therefore improve treatment with azathioprine (and 6-mercaptopurine) in the future.

ACETYLATION

Isoniazid, hydralazine, sulfonamides, procainamide, and a number of other drugs are metabolized by acetylation of a hydrazino or amino group. This reaction is catalyzed by *N*-acetyl transferase-2 (NAT-2), an enzyme in the liver cytosol that transfers an acetyl group from acetyl coenzyme A to the drug. Individuals differ markedly in the rate at which drugs are acetylated, because of polymorphisms in the NAT-2 gene, resulting in a bimodal distribution of the population into "rapid acetylators" and "slow acetylators."

Acetylation phenotype can be determined by measuring the ratio of acetylated to nonacetylated forms of the probe drugs, dapsone, caffeine, or sulfamethazine, in plasma or urine following administration of a test dose of these acetylation substrates. Slow, intermediate, and rapid acetylators may be identified by these methods for phenotyping. It is also possible to identify slow acetylators by genotyping, using genomic DNA obtained from blood leukocytes.

METABOLISM BY CYTOCHROME P450 MONOOXYGENASES

In healthy individuals taking no other medications, the major determinant of the rate of metabolism of drugs by the cytochrome P450 monooxygenases is genetic. Hepatic endoplasmic reticulum contains a family of cytochrome P450 (CYP) isoforms with different substrate specificities. Many drugs undergo oxidative metabolism by more than one isoform, and the steady-state concentrations of such drugs in the plasma is a function of the sum of the activities of these and other metabolizing enzymes. When a drug is metabolized by multiple pathways, the catalytic activities of the participating enzymes are regulated by a number of genes, so that the clearance rates and steady-state concentrations of the drug tend to distribute unimodally within the population. The range of activity may differ markedly (^{3}tenfold) between different individuals, as is the case for chlorpromazine, and there is no way to predict the rate before beginning therapy.

Certain metabolic pathways show a bimodal or trimodal distribution of activity, suggesting control by a single gene, and polymorphisms in these genes have been identified. Most individuals are extensive metabolizers (EM phenotype); a smaller group have a lower ability to metabolize the drug (or no ability at all) and are called poor metabolizers (PM phenotype). Heterozygotes for genes encoding the enzymes lacking catalytic activity may be intermediate metabolizers (IM phenotype). And patients with duplicate or multiple copies of the gene may exhibit ultrarapid metabolism. These polymorphisms are of greatest clinical relevance during administration of substrate drugs for which there are no major alternative routes of elimination. The clinical consequences of the PM phenotype will then depend on the resultant accumulation of the drug or occasionally on the absence of generation of active metabolites.

The cytochrome P450 isoform CYP2D6 is polymorphically distributed in the population, and about 8 to 10% of Caucasians are deficient in this enzyme. CYP2D6 represents the main metabolic pathway for a number of drugs, including antiarrhythmic agents (propafenone, flecainide), b-adrenoceptor blockers (timolol, metoprolol, and alprenolol), tricyclic antidepressants (nortriptyline, desipramine, imipramine, clomipramine), neuroleptic drugs (perphenazine, thioridazine, and possibly haloperidol), selective serotonin reuptake inhibitors (fluoxetine and paroxetine), and certain opiates, such as codeine and dextromethorphan. Thus, codeine has a much lower analgesic effect inPMpatients because of impaired production of the active metabolite, morphine. Conversely, a patient with duplicate or multiple copies of CYP2D6 will exhibit an exaggerated response to codeine. Patients with the PM phenotype experience more pronounced systemic b-adrenoceptor blockade after the administration of timolol ophthalmic solution. The catalytic activity of CYP2D6 in humans may be assessed by using a test drug, debrisoquin, which is eliminated almost entirely via hydroxylation by CYP2D6. Individuals with the PM phenotype can be identified by genotyping for the alleles that encode proteins with loss of catalytic function. There are ethnic variations in the frequency of the PM phenotype, which occurs in 5 to 10% of Caucasians but with a lesser frequency in Asians (1 to 2%).

The isoform CYP2C19 also exhibits polymorphism; it was initially detected with the hydroxylation of mephenytoin, which is used as a probe drug for the function of this P450 isoform. This enzyme catalyzes the major metabolic pathway of omeprazole, proguanil, diazepam, and citalopram. The impact of the polymorphism in CYP2C19 on treatment outcome is clearly illustrated by omeprazole. The efficacy of omeprazole (20 mg in combination with amoxicillin) in eradicating *Helicobacter pylori* is markedly reduced in persons with the homozygousEMgenotype (29% cured) as compared with 100% cure in those with homozygousPMgenotype. This reflects the relative lack of effect of the recommended dose of omeprazole (20 mg) on gastric acid secretion and ulcer healing in patients with the CYP2C19 EM genotype. Certainly knowledge of a patient's CYP2C19 genotype would improve therapy with this proton pump inhibitor. Impaired hydroxylation of mephenytoin is present in only 3 to 5 percent of Caucasians, but the incidence is about 20 percent in individuals of Japanese and Chinese descent.

CYP2C9 catalyzes the major pathways of metabolism of warfarin and phenytoin. There are allelic variants of the gene for this enzyme that encode proteins with loss of catalytic function. These variant alleles are associated with requirement for a very low dose of warfarin, difficulties in initiating warfarin therapy, and an increased risk of bleeding complications. Similarly, high concentrations of phenytoin in plasma and resulting adverse effects of phenytoin occur in patients with loss of function alleles for CYP2C9.

Polymorphisms in drug-metabolizing ability may be associated with large differences in the disposition of a drug among individuals, especially when the involved pathway makes a major contribution to the elimination of the drug. For example, the clearance of mephenytoin given orally differs 100- to 200-fold between individuals of theEM andPMphenotypes. As a result, the peak plasma concentrations and bioavailability after oral administration are much higher, and the rate of drug elimination much lower, in PM than in EM individuals. In PM individuals, the result is excessive drug accumulation and exaggerated pharmacologic responses, including toxicity, when usual drug dosages are administered. Individualization of drug therapy is especially critical for drugs that exhibit polymorphic drug metabolism. The increasing availability of laboratory methods to identify the PM phenotype for NAT-2, CYP2D6, and CYP2C19 by genotyping should be useful for this purpose.

INTERINDIVIDUAL VARIABILITY IN THE MOLECULAR TARGETS WITH WHICH DRUGS INTERACT

The increasing emphasis on identifying molecular mechanisms of disease (Chap. 65) has important consequences for further understanding a genetic basis for individual variability in drug actions. As molecular approaches identify the role of specific gene products in human physiology, polymorphisms that alter expression or function of those gene products are being recognized; it is estimated that such polymorphisms occur in 1 in 1000 bp in the human genome. These genes in turn, are now being recognized as the molecular targets with which available and new drugs interact to produce beneficial and adverse effects.

Genome-wide searches in families with premature Alzheimer's disease identified the *APOE* locus as linked to the disease (Chap. 362). Specifically, the *E4* allele of the *APOE* gene appears associated with a worse prognosis, and this is thought to relate to reduced expression of choline acetyl transferase. Further, a therapeutic response to the choline acetyl transferase inhibitor, tacrine, appears to be more common with the prognostically more benign *APOE2* or *APOE3* alleles. Multiple polymorphisms identified in theb$_2$-adrenergic receptor appear to be linked to specific phenotypes in asthma and congestive heart failure, diseases in whichb$_2$-receptor function might be expected to determine prognosis. It has been suggested that polymorphisms in theb$_2$receptor may be a determinant of response to inhaledb$_2$-receptor agonists.

The development of marked QT prolongation and the polymorphic ventricular tachycardia, *torsade de pointes* (Chap. 230), in response to certain action potential-prolonging drugs such as quinidine used to be characterized as an "idiosyncratic" response. Advances in understanding the molecular basis of normal cardiac repolarization have resulted in identification of genes encoding ion channel proteins, the molecules whose normal function results in physiologic cardiac repolarization. Mutations in these genes cause congenital arrhythmia syndromes, such as the long QT syndrome, and block of ion channels is a common mechanism whereby drugs prolong QT intervals. Patients with mutations in these genes that remain subclinical until challenge with drugs are now recognized. In summary, continuing efforts to unravel the molecular basis of disease are likely also to provide insights into determinants of the response to drug therapy.

DRUG USE IN THE ELDERLY (See alsoChap. 9)

Aging results in changes in organ function, especially of the organs involved in drug disposition, as well as alterations in body size and composition. Not surprisingly, therefore, pharmacokinetics are often different in elderly individuals than in younger adults. Also, elderly patients often have multiple diseases and may therefore be taking a large number of drugs. Consequently, drug interactions, as well as an increased vulnerability to morbidity and mortality, contribute to the higher incidence of adverse drug reactions in elderly patients. Increased sensitivity of target organs and impairment of physiologic control systems, such as those involved in the regulation of the circulation, may also be a factor. Accordingly, optimization of drug therapy in the elderly, particularly in frail patients, is often difficult, as a variety of factors (often poorly defined) accentuate the usual interindividual variability in drug response.

Although many individuals preserve good renal function into old age, elderly patients as a group have an increased likelihood of impaired renal excretion of drugs. Even in the absence of kidney disease, renal clearance is generally reduced by about 35 to 50% in elderly patients. Dosage adjustments analogous to those in patients with renal dysfunction (see above) are therefore necessary for drugs that are eliminated mainly by the kidneys, such as digoxin, aminoglycosides, lithium, and other drugs listed in Table 70-3. In this regard, it is important to recognize that the reduced muscle mass of older individuals results in a reduced rate of creatinine production; thus, a normal serum creatinine concentration can be present even though creatinine clearance is impaired.

Aging also results in a decrease in the size of and blood flow to the liver and possibly in the activity of hepatic drug-metabolizing enzymes; accordingly, the hepatic clearance of some drugs is impaired in the elderly. Unfortunately, no consistent pattern of clinical application appears to be present. Moreover, the changes are often modest relative to other causes for interindividual variability in these patients. However, even a small reduction in hepatic extraction may significantly increase the oral bioavailability of drugs with a high first-pass effect, such as propranolol and labetalol.

Impaired clearance and/or increased distribution may cause the elimination half-life of a drug to increase with aging. Thus, if a dosage modification in an elderly patient is required, it is often possible to accomplish it by decreasing the frequency of drug administration, possibly along with a reduction in dose.

Even if the pharmacokinetics of a drug are not altered, an elderly patient may require a smaller dosage because of an increase in pharmacodynamic sensitivity. Examples include increased analgesic effects of opioids, increased sedation from benzodiazepines and otherCNSdepressants, and increased risk of bleeding while receiving anticoagulant therapy, even when clotting parameters are well controlled. Exaggerated responses to cardiovascular drugs are also common because of the impaired responsiveness of normal homeostatic mechanisms. Such age-related changes require close monitoring of the patient's clinical response and appropriate dosage titration. Accordingly, in the elderly, initial doses should be less than the usual adult dosage and should be increased slowly. The final therapeutic regimen should be as simple as possible, and the number of different drugs used should be kept as low as possible. Also, because interindividual variability in drug responsiveness is greater in geriatric patients than in younger adults, individualization of therapy is even more critical.

INTERACTIONS BETWEEN DRUGS

The effect of some drugs can be altered markedly by the administration of other agents. Such interactions can complicate therapy by adversely increasing or decreasing the action of a drug. Drug interactions must be considered in the differential diagnosis of unexpected responses to drugs, and it should be recognized that patients often come to the physician with a legacy of drugs acquired during previous medical experiences. A meticulous drug history will minimize such unknown elements. It should include examination of the patient's medications and, if necessary, calls to the pharmacist to identify prescriptions. It should also address the use agents not often volunteered on initial questioning, such as over-the-counter drugs, health food supplements, and topical agents such as eye drops.

There are two principal types of interactions between drugs. In *pharmacokinetic interactions*, the delivery of a drug to its site of action is altered, whereas in *pharmacodynamic interactions*, the responsiveness of the target organ or system is modified.

An index of the drug interactions discussed in this chapter is provided inTable 70-4. The table includes interactions that have verified significance in patients, plus a few that are so potentially dangerous that cognizance should be taken of the experimental data or case reports suggesting they occur.

I. PHARMACOKINETIC INTERACTIONS CAUSING DIMINISHED DRUG DELIVERY

A. Impaired Gastrointestinal Absorption Examples include aluminum ions, present in antacids, which form insoluble chelates with the tetracyclines, preventing absorption of these drugs. Ferrous ions similarly block tetracycline absorption. Kaolin-pectin suspensions bind digoxin, and when these substances are administered together, digoxin absorption is reduced by about one-half. However, when kaolin-pectin is administered 2 h after digoxin, digoxin absorption is unaffected.

Ketoconazole is a weak base that dissolves well only at acidic pH. Histamine H_2receptor antagonists, such as ranitidine and cimetidine, reduce gastric acidity and thus impair the dissolution and absorption of ketoconazole. By contrast, the absorption of fluconazole is not impaired by an increase in gastric pH.

B. Induction of Hepatic Drug-Metabolizing Enzymes When a drug is eliminated largely by metabolism, an increase in the rate of its metabolism reduces its availability to sites of action. Most drugs are metabolized largely in the liver because of this organ's large mass, high blood flow, and high concentration of drug-metabolizing enzymes. The first step in the metabolism of many drugs is catalyzed by a group of cytochrome P450 mixed-function oxidases located in the endoplasmic reticulum (see "Metabolism by Cytochrome P450 Monooxygenases," above). These enzyme systems oxidize drug molecules by a variety of reactions, including aromatic hydroxylations, *N*-demethylations, *O*-demethylations, and sulfoxidations. The products of these reactions are usually more polar than the parent compound (and more readily excreted by the kidney).

The expression of some of the mixed-function oxidase (CYP) isoforms is regulated, and their content in the liver can be increased, by a number of drugs. Phenobarbital is the prototype of these inducers, and all the barbiturates in clinical use increase CYP enzyme activity. Induction with phenobarbital can occur with doses of as little as 60 mg daily. Mixed-function oxidases are also induced by rifampin, carbamazepine, phenytoin, and glutethimide and by smoking, exposure to chlorinated insecticides such as DDT, and chronic alcohol ingestion.

Phenobarbital, rifampin, and other inducers lower plasma levels of many drugs, including warfarin, quinidine, mexiletine, verapamil, ketoconazole, itraconazole, cyclosporine, dexamethasone, methylprednisolone, prednisolone (the active metabolite of prednisone), oral contraceptive steroids, methadone, metronidazole, and metyrapone. These interactions all have obvious clinical significance. In the case of the coumarin anticoagulants, the patient is placed at major risk if the appropriate level of anticoagulation is achieved when an inducer is also being administered and the inducer is later discontinued (for example, at discharge from the hospital). The plasma levels of the coumarin anticoagulant will rise as the induction effect wears off, leading to excessive anticoagulation. There is considerable variation among individuals in the extent to which drug metabolism can be induced.

C. Inhibition of Cellular Uptake or Binding The guanidinium antihypertensive agents guanethidine and guanadrel are transported to their site of action in adrenergic neurons by an energy-requiring membrane transport system for biogenic monoamines; the physiologic function of this system is reuptake of the adrenergic neurotransmitter. Inhibitors of norepinephrine uptake prevent the uptake of the guanidinium antihypertensive agents into adrenergic neurons and thereby block their pharmacologic effects. The tricyclic antidepressants are potent inhibitors of norepinephrine uptake. Consequently, concomitant administration of clinical doses of tricyclic antidepressants, including desipramine, protriptyline, nortriptyline, and amitriptyline, almost totally abolishes the antihypertensive effects of guanethidine and guanadrel. Although they are less potent inhibitors of norepinephrine uptake, doxepin and chlorpromazine produce dose-related antagonism of the action of the guanidinium antihypertensives.

The antihypertensive effect of clonidine is partially antagonized by tricyclic antidepressants. Clonidine lowers arterial pressure by reducing sympathetic outflow from the blood pressure-regulating centers in the hindbrain (Chap. 246). This central hypotensive action is antagonized by the tricyclic antidepressants.

II. PHARMACOKINETIC INTERACTIONS CAUSING INCREASED DRUG DELIVERY

A. Inhibition of Drug Metabolism If the active form of a drug is eliminated largely by biotransformation, inhibition of its metabolism leads to reduced clearance, prolonged half-life, and accumulation of the drug during maintenance therapy. Excessive accumulation due to inhibited metabolism can lead to adverse effects.

Cimetidine is a potent inhibitor of the oxidative metabolism of many drugs, including warfarin, quinidine, nifedipine, lidocaine, theophylline, and phenytoin. Adverse reactions, many of them severe, have resulted from the administration of these drugs in conjunction with cimetidine. Cimetidine is a more potent inhibitor of mixed-function oxidases than ranitidine, whereas ranitidine is more potent as a histamine H_2receptor antagonist. Famotidine and nizatidine are not known to produce clinically appreciable inhibition of drug metabolism.

Knowledge of the CYP isoforms that catalyze the main pathway of metabolism of a drug provides a basis for predicting and understanding drug interactions. For example, the CYP3A subfamily of isoforms catalyzes the metabolism of many drugs for which blockage of metabolism results in toxicity. Drugs that depend on CYP3A as a major route of metabolism include cyclosporine, quinidine, lovastatin, simvastatin, atorvastatin, nifedipine, lidocaine, cisapride, erythromycin, methylprednisolone, carbamazepine, midazolam, and triazolam.

The antifungal agents ketoconazole and itraconazole are potent inhibitors of enzymes in the CYP3A family. When fluconazole levels are elevated as a result of higher doses and/or renal insufficiency, this drug can also inhibit CYP3A. The macrolide antibiotics erythromycin and clarithromycin inhibit CYP3A4 to a clinically significant extent, but azithromycin does not inhibit this enzyme. Some of the calcium antagonists, diltiazem, nicardipine, and verapamil can also inhibit CYP3A, as can some of its other substrates, such as cyclosporine.

Cyclosporine can cause serious toxicity when its metabolism is inhibited by erythromycin, ketoconazole, diltiazem, nicardipine, or verapamil. A serious complication of HMG-CoA reductase inhibitors is myopathy. Fortunately, this is infrequent except in the context of interactions of a subset of the HMG-CoA reductase inhbitors with other drugs, particularly those that inhibit CYP3A4. The disposition of lovastatin is reduced markedly by drugs that inhibit CYP3A4, causing increases in plasma levels by more than tenfold. As a consequence, lovastatin has produced severe myopathy with rhabdomyolysis when administered together with erythromycin or cyclosporine. Not all of the HMG-CoA reductase inhibitors are as dependent on CYP3A4 for disposition as is lovastatin. Blocking CYP3A4 causes moderate elevations of plasma levels of simvastatin and atorvastatin (increases of severalfold), whereas elevations of the levels of fluvastatin and cerivastatin are only slight. Pravastatin disposition and plasma levels are not altered by inhibitors of CYP3A4. Cisapride can cause polymorphic ventricular tachycardia (torsade de pointes) when its metabolism is blocked by inhibitors of CYP3A, such as ketoconazole, itraconazole, clarithromycin, and erythromycin.

Whenever an inhibitor of CYP3A4 is administered to a patient, the physician should be alert to the possibility of serious interactions with drugs that are metabolized by CYP3A.

The CYP2D6 isoform that catalyzes the polymorphic metabolism of debrisoquin is markedly inhibited by quinidine and is also blocked by a number of neuroleptic drugs, such as chlorpromazine and haloperidol, and by fluoxetine. The analgesic effect of codeine depends on its metabolism to morphine via CYP2D6 in individuals with theEMphenotype. Thus, quinidine reduces the analgesic efficacy of codeine in EMs. Since desipramine is cleared largely by metabolism via CYP2D6 in EMs, its levels are increased substantially by concurrent administration of quinidine, fluoxetine, or the neuroleptic drugs that inhibit CYP2D6.

Some drugs are inactivated by mechanisms other than the hepatic drug-metabolizing enzymes. Azathioprine is converted in the body to an active metabolite, 6-mercaptopurine, which in turn is oxidized by xanthine oxidase to 6-thiouric acid. When allopurinol, a potent inhibitor of xanthine oxidase, is administered concurrently with standard doses of azathioprine or 6-mercaptopurine, life-threatening toxicity (bone marrow suppression) can result.

Other drugs that inhibit biotransformation of pharmacologic compounds (with examples of drugs whose metabolism is blocked by the inhibitor listed in parenthesis) include:

· Amiodarone (warfarin, quinidine)

· Clofibrate (phenytoin, tolbutamide)

· Excessive ingestion of ethanol (warfarin)

· Isoniazid (phenytoin)

· Metronidazole, cotrimoxazole (warfarin)

· Phenylbutazone (warfarin, phenytoin, tolbutamide)

B. Inhibition of Drug Transport Specific molecules that transport drugs into and out of cells are increasingly recognized, and inhibition of their function can be a major cause of clinically important drug interactions. The best studied to date is P-glycoprotein, originally isolated from tumor cells displaying resistance to multiple, structurally unrelated anticancer agents. The mechanism underlying this "multidrug resistance" phenomenon is P-glycoprotein-mediated pumping of anticancer agents out of cells, thereby inhibiting their anticancer effects. P-glycoprotein is also expressed in normal tissues (the luminal aspect of intestinal and renal tubular cells, the canalicular aspect of hepatocytes, the capillary endothelium of the blood-brain barrier), where it is responsible for efflux of not only antineoplastics but also digoxin and HIV protease inhibitors. Quinidine inhibits P-glycoprotein function in vitro, and it now seems apparent that the widely recognized doubling of plasma digoxin when quinidine is coadministered reflects this action in vivo, particularly since the effects of quinidine (increased digoxin bioavailability and reduced renal and hepatic secretion) occur at the sites of P-glycoprotein expression. Many other drugs also elevate digoxin concentrations (e.g., amiodarone, verapamil, cyclosporine, itraconazole, and erythromcyin), and a similar mechanism seems likely. ReducedCNSpenetration of multiple HIV protease inhibitors (with the attendant risk of facilitating a sanctuary site for the virus) appears attributable to P-glycoprotein-mediated exclusion of the drug from the CNS.

A number of drugs are secreted by the renal tubular transport systems for organic anions. Inhibition of this tubular transport system can cause excessive accumulation of a drug. Phenylbutazone, probenecid, and salicylates competitively inhibit this transport system. Salicylate, for example, reduces the renal clearance of methotrexate, an interaction that may lead to methotrexate toxicity. Renal tubular secretion contributes substantially to the elimination of penicillin, which can be inhibited by probenecid.

Inhibition of the tubular cation transport system by cimetidine impedes the renal clearance of procainamide and its active metabolite *N*-acetylprocainamide.

III. PHARMACODYNAMIC AND OTHER INTERACTIONS BETWEEN DRUGS

Therapeutically useful interactions occur in which the effect of two drugs in combination is greater than the sum of their effects when used individually. Favorable drug combinations are described in specific therapeutic sections in this text, and this section focuses on interactions that create unwanted effects. Two drugs may act on separate components of a common process to yield effects greater than either has alone. For example, although small doses of aspirin (<1 g daily) do not alter the prothrombin time appreciably in patients who are receiving warfarin therapy, aspirin nevertheless increases the risk of bleeding in these patients because it inhibits platelet aggregation. Thus the combination of impaired functions of platelets and the clotting system, while useful for some therapeutic purposes, also increases the potential for hemorrhagic complications in patients receiving warfarin therapy.

Nonsteroidal antiinflammatory drugs (NSAIDs) cause gastric and duodenal ulcers, and, in patients treated with warfarin, the risk of bleeding from a peptic ulcer is increased almost threefold by concomitant use of a NSAID. This clearly is a serious drug interaction.

Indomethacin, piroxicam, and probably otherNSAIDsantagonize the antihypertensive effects ofb-adrenergic receptor blockers, diuretics,ACEinhibitors, and other drugs. The resulting elevation in blood pressure ranges from trivial to severe. Aspirin and sulindac, however, do not elevate the blood pressure in treated hypertensive patients.

Polymorphic ventricular tachycardia (torsade de pointes) during quinidine administration occurs much more frequently in patients receiving diuretics, probably owing to potassium and/or magnesium depletion.

The administration of supplemental potassium leads to more frequent and more severe hyperkalemia when potassium elimination is reduced by concurrent treatment withACEinhibitors, spironolactone, amiloride, or triamterene.

The pharmacologic effects of sildenafil result from inhibition of the phosphodiesterase type 5 isoform that inactivates cyclic GMP in the vasculature. Nitroglycerin and related nitrates produce vasodilation by elevating cyclic GMP. Thus, coadministration of these nitrates with sildenafil will cause profound and potentially catastrophic hypotension.

CONCENTRATION OF DRUGS IN PLASMA AS A GUIDE TO THERAPY

In many cases, the plasma concentration of a drug is measured as a guide in the individualization of therapy. Genetic variation in elimination rates, interactions with other drugs, disease-induced alterations in elimination and distribution, and other factors combine to yield a wide range of plasma levels in patients given the same dose. Furthermore, the problem of noncompliance with prescribed regimens during continuing therapy is an endemic and elusive cause of therapeutic failure (see below). Clinical indicators assist in the titration of some drugs into the desired range, but no chemical determination is a substitute for careful observations of the response to treatment. However, the therapeutic and adverse effects are not precisely quantifiable for all drugs, and, in complex clinical situations, estimates of the action of a drug may be misleading. For example, previously existing neurologic disease may obscure the neurologic consequences of intoxication with phenytoin. Because clearance, half-life, accumulation, and steady-state plasma levels are difficult to predict, the measurement of plasma levels is often useful as a guide to the optimal dose. This is particularly true when there is a narrow range between the plasma levels yielding therapeutic and adverse effects. For drugs having such characteristics -- e.g., digoxin, theophylline, lidocaine, aminoglycosides, cyclosporine, and anticonvulsants -- dose optimization should involve modification of the standard dose on the basis of the pharmacokinetic principles described above. In certain instances, predictive nomograms and algorithms have been developed to facilitate the necessary modifications. However, the most flexible and accurate method for individualizing drug dosage appears to be a feedback approach using a small number of previously obtained plasma levels and Bayesian forecasting. In controlled studies, this type of computer-assisted dosing has been shown to improve patient care. However, the overall cost/benefit ratio of such methods in routine management still remains to be concusively demonstrated.

For drugs with a narrow therapeutic window that exhibit first-order elimination, then, dosage adjustments may be made on the assumption that the average, maximum, and minimum steady-state concentrations are related linearly to the dosing rate. Accordingly, the dose may be adjusted on the basis of the ratio between the desired and measured concentrations:

For drugs that have dose-dependent kinetics (e.g., phenytoin and theophylline), plasma concentrations change disproportionately more than the alteration in the dosing rate. Not only should changes in dose be small to minimize the degree of unpredictability, but plasma concentration monitoring is also critical to ensure appropriate modification.

The variability among individual responses to given plasma levels must be recognized. This is illustrated by a hypothetical population concentration-response curve (Fig. 70-4) and its relationship to the therapeutic range or therapeutic window of desired plasma levels. The defined therapeutic window should include the levels at which the intended pharmacologic effect is achieved in most patients. However, a few persons, who are sensitive to the therapeutic effects, respond to lower levels, whereas others are refractory enough to require levels that may cause adverse effects. For example, a few patients with strong seizure foci require plasma levels of phenytoin exceeding 20 ug/mL to control seizures. Dosages to achieve this effect may be appropriate if tolerated.

As also illustrated inFig. 70-4, some patients are prone to adverse effects at levels that are tolerated by most of the population. Therefore, raising the plasma concentration of a drug to a level that has a high probability of being therapeutically effective may bring on unwanted actions in an occasional patient. Table 70-2 presents for a number of drugs the plasma concentrations that are associated with adverse and therapeutic effects in most patients. Use of this information according to the guidelines discussed should permit more effective and safer therapy for those patients who are not "average."

EFFECTIVE PARTICIPATION OF THE PATIENT IN THERAPY

Measurement of the concentration of a drug in plasma is the most effective way to detect failure to take a drug. Such "noncompliance" is a frequent problem in the long-term treatment of diseases such as hypertension and epilepsy, occurring in 25% or more of patients in therapeutic environments in which no special effort is made to involve patients in the responsibility for their own health. Occasionally, noncompliance can be uncovered by sympathetic, nonincriminating questioning, but more often it is recognized only after determining that the concentration of drug in plasma is nil or is recurrently low. Because other factors can cause plasma levels to be lower than expected, comparison with levels obtained during inpatient treatment may be required to confirm that noncompliance has occurred. Once the physician is certain of noncompliance, a nonaccusatory discussion of the problem with the patient may clarify the reason for the noncompliance and serve as a basis for more effective cooperation on the part of the patient. Many approaches have been tried to help patients exercise more responsibility for their own treatment, most based on better communication regarding the nature of the disease and the chances of success or failure of the treatment. The patient is given a chance to discuss problems associated with treatment. The process may be improved by the involvement of nurses and other paramedical personnel. Minimizing the complexity of the regimen is helpful in terms of both the number of drugs and the frequency of administration. Educating patients to assume the principal role in their own health care requires a blend of the art and science of medicine.

(Bibliography omitted in Palm version)

Back to Table of Contents

71. ADVERSE REACTIONS TO DRUGS - *Alastair J. J. Wood*

The beneficial effects of drugs are coupled with the inescapable risk of untoward effects. The morbidity and mortality from these untoward effects often present diagnostic problems because they can involve every organ and system of the body and are frequently mistaken for signs of underlying disease. Major advances in the investigation, development, and regulation of drugs ensure in most instances that they are uniform, effective, and relatively safe and that their recognized hazards are publicized. However, prior to regulatory approval and marketing, new drugs are tested in relatively few patients who tend to be less sick and to have fewer concomitant diseases than those patients who subsequently receive the drug therapeutically. Because of the relatively small number of patients studied in clinical trials, and the selected nature of these patients, rare adverse effects may not be detected prior to a drug's approval, and physicians therefore need to be cautious in the prescription of new drugs and alert for the appearance of previously unrecognized adverse events.

The large number and variety of drugs available over the counter (OTC), herbal preparations, and by prescription, make it impossible for patient or physician to obtain or retain the knowledge necessary to use all drugs well. It is understandable, therefore, that many OTC drugs are used unwisely by the public and that restricted drugs may be prescribed incorrectly by physicians.

Most physicians use no more than 50 drug products in their practice, gaining familiarity with their effectiveness and safety. Most patients probably use only a limited number ofOTCdrugs. Nevertheless, many patients receive care and drug prescriptions from more than one physician, and in any 30-day period, many patients consume more than three different OTC drug products containing nine or more different chemical agents.

Some 25 to 50% of patients make errors in self-administration of prescribed medicines, and these errors can be responsible for adverse drug effects. Elderly patients are the group most likely to commit such errors, perhaps in part because they consume more medicines. One- third or more of patients also may not take their prescribed medications. Similarly, patients commit errors in takingOTCdrugs by not reading or following the directions on the containers. Physicians must recognize that providing directions with prescriptions does not always guarantee compliance.

Every drug can produce untoward consequences, even when used according to standard or recommended methods of administration. When used incorrectly, the effectiveness may be reduced, and adverse reactions can be expected to occur more frequently. Also, the administration of several drugs concurrently may result in adverse drug interactions (Chap. 70). In the hospital, all drugs a patient is given should be under the control of a physician, and patient compliance is, in general, ensured. Errors may occur nevertheless -- the wrong drug or dose may be given, or the drug may be given to the wrong patient -- although improved drug distribution and administration systems have reduced this problem. On the other hand, there are no easy means for controlling how ambulatory patients take prescription orOTCdrugs.

EPIDEMIOLOGY

Epidemiologic studies of adverse drug reactions have been helpful in evaluating the magnitude of the overall problem, in calculating the rate of reactions to individual drugs, and in characterizing some of the determinants of adverse drug effects.

Patients receive, on average, 10 different drugs during each hospitalization. The sicker the patient, the more drugs are given, and there is a corresponding increase in the likelihood of adverse drug reactions. When fewer than 6 different drugs are given to hospitalized patients, the probability of an adverse reaction is about 5%, but if more than 15 drugs are given, the probability is over 40%. Retrospective analyses of ambulatory patients have revealed adverse drug effects in 20%.

Thus, the magnitude of drug-induced disease is large. Of patients admitted to the medical and pediatric services of general hospitals, 2 to 5% are admitted because of illnesses attributed to drugs. The case/ fatality ratio from drug-induced disease in hospitalized patients varies from 2 to 12%. Furthermore, some fetal or neonatal abnormalities are due to medicines taken by the mother during pregnancy or parturition.

A small group of widely used drugs accounts for a disproportionate number of reactions. Aspirin and other nonsteroidal anti-inflammatory drugs, analgesics, digoxin, anticoagulants, diuretics, antimicrobials, glucocorticoids, antineoplastics, and hypoglycemic agents account for 90% of reactions, although the drugs involved differ between ambulatory and hospitalized patients. Estimates of the cost of drug-related morbidity and mortality in the ambulatory setting range from $30 billion to $130 billion.

ADVERSE DRUG REACTIONS IN THE ELDERLY (See also Chap. 9)

The elderly as a group have a greater burden of disease and receive a greater number of medications than other persons. Thus, it is not surprising that adverse drug reactions occur frequently in elderly patients. The issue of whether an elderly individual is more likely to develop an adverse drug reaction than a young person with a similar number of concurrent diseases and taking the same number of drugs has not been answered unequivocally. However, in population surveys of the noninstitutionalized elderly, as many as 10% report having had at least one adverse drug reaction in the last year. The incidence appears to be even greater in hospitalized elderly patients. Although it is widely believed that the elderly are more sensitive to drugs than the young, that is not true for all drugs. For example, a consistent decrease in sensitivity to drugs acting at theb-adrenergic receptor has been demonstrated in the elderly. The consequences of adverse drug effects may differ in the elderly because of their greater likelihood of other disease. For example, use of long-half-life benzodiazepines is linked to the occurrence of hip fractures in elderly patients, perhaps reflecting both a risk of falls from these drugs and the increased incidence of osteoporosis in elderly patients. Even when a drug impairs function similarly in patients of different age groups, the poorer baseline function in elderly persons may put them at greater risk for an adverse drug reaction. When prescribing for an elderly patient, the possibility that hepatic or renal mechanisms of drug excretion may be impaired should be taken into account. Adverse drug effects in the elderly may be subtle and, as in all populations, the physician must be alert to the possibility that a patient's signs and symptoms reflect an adverse effect of medication.

ETIOLOGY

Most adverse drug reactions are preventable, and recent studies using a systems analysis approach suggest that the most common system failure associated with an adverse drug reaction is the failure to disseminate knowledge about drugs to individuals involved in prescribing and administering them. Most adverse reactions can be classified into two groups. The most frequent ones result from an exaggeration of a predicted pharmacologic action of the drug. Other adverse reactions ensue from toxic effects unrelated to the intended pharmacologic actions. The latter effects are often unpredictable, are frequently severe, and result from recognized as well as undiscovered mechanisms. Some mechanisms unrelated to the drug's primary pharmacologic activity may include direct cytotoxicity, initiation of abnormal immune responses, and perturbation of metabolic processes in individuals with genetic enzymatic defects. Further understanding of interindividual differences in the expression of the enzymes responsible for drug metabolism has contributed to the understanding of adverse drug reactions that previously were thought to be idiosyncratic (see below). Prior consideration of the factors known to modify drug action often make it possible to prevent adverse reactions of this type.

Genetic Variations in Drug Oxidation by Cytochromes There is considerable interindividual variability in drug metabolism, resulting in variability in drug concentrations (Chap. 70). The majority of drugs are oxidized by cytochrome P450s (CYP) in the liver and gut. Some of these enzymes exhibit genetic polymorphisms resulting in enzymes with absent or reduced drug metabolizing activity, which may result in concentration-dependent toxicity. Conversely, where toxicity or pharmacologic effect is produced by a metabolite, individuals with low enzyme activity may have reduced drug effect whereas those with genetically determined increased enzyme activity will have increased drug effect. Examples of such polymorphically distributed oxidation enzymes include CYP2D6, CYP2C9, and CYP2C19.

The clinical consequences of the poor metabolizer phenotype are now becoming clearer and depend on the specific consequences of excessive drug concentrations. For example, the more potent (S) isomer of warfarin is metabolized by the polymorphically distributed enzymeCYP2C9, resulting in lower (S) warfarin clearance and higher concentrations in both heterozygotes and homozygotes for the allelic variants associated with reduced enzyme activity. Recently, the clinical consequences of this polymorphism have been demonstrated in patients followed in an anticoagulant clinic. Patients who were stabilized on warfarin doses of 1.5 mg/d or less had an increased frequency of the genotypes associated with low warfarin metabolism compared to either community controls or patients requiring higher doses of warfarin (Table 71-1). In addition, the group stabilized on low-dose warfarin had a greater incidence of initial over-anticoagulation and hemorrhage than the group on the higher dose. This serves as an example of how genetic variations of a cytochrome P450 enzyme alter the response to a drug.

The oral hypoglycemic glipizide is also metabolized byCYP2C9, and excessively low blood glucose concentrations occur after usual doses in genetic CYP2C9 poor metabolizers. Another oral hypoglycemic, phenformin, produced lactic acidosis in some patients. It is now recognized that phenformin is metabolized by another polymorphically distributed oxidative enzyme, CYP2D6. Patients who have genetically determined low activity of CYP2D6 may be at particular risk from phenformin-induced lactic acidosis.

Genotypically determined variability in drug toxicity may also occur with drugs metabolized by enzymes other than cytochrome P450s. Such toxicity can be severe with the clinical use of the antimetabolites azathioprine and 6-mercaptopurine (to which it is converted in vivo). The cytotoxic thioguanine nucleotides produced in vivo are detoxified by further metabolism by xanthine oxidase and thiopurine methyltransferase (TPMT). The latter enzyme shows a trimodal distribution (Fig. 71-1). In children receiving mercaptopurine for treatment of leukemia, low activity of this enzyme is associated with excessive myelosuppression, whereas children with high TPMT levels have a poor antileukemic response. Azathioprine is currently used as a disease-modifying agent in the treatment of rheumatoid arthritis. Individuals with mutant TPMT alleles that result in impaired metabolism developed toxicity rapidly after beginning azathioprine and were uniformly unable to take the drug chronically. Thus the genotypic basis for drug toxicity is beginning to emerge.

Pharmacokinetic Bases for Adverse Reactions *An abnormally high drug concentration at the receptor site* (site of action) owing to pharmacokinetic variability is the usual cause of these reactions (Chap. 70). For example, a reduction in the volume of distribution, in the rate of metabolism, or in the rate of excretion all result in higher than expected concentration of drug at the receptor site, with a consequent increase in the pharmacologic effect.

Alteration in the dose-response curve due to increased receptor sensitivity results in an increase in drug effect at a given drug concentration. An example is the excessive response of elderly persons to the anticoagulant warfarin at normal or lower than normal blood levels. Such alterations in the dose-response curve may reflect altered drug sensitivity due to receptor polymorphisms, which are now being recognized. One such example is the prolonged QT syndrome, which has both a genetic basis, in individuals with abnormal potassium channels involved in cardiac repolarization, and a pharmacologic basis in individuals who receive drugs known to prolong the QT interval. Such individuals may develop torsade de pointes (Chap. 230). A large number of drugs have now been identified that can produce this potentially lethal effect (Table 71-2; alsohttp://www.dml.georgetown.edu/depts/pharmacology/torsades.html).

The shape of the dose-response curve also determines the likelihood of adverse drug reactions. Drugs with a steep dose-response curve or a narrow therapeutic index (Chap. 70) are more likely to cause dose-related toxicity because a small increase in dose produces a large change in pharmacologic effect. An increase in the dose of drugs that exhibit nonlinear kinetics, such as phenytoin (Chap. 70), may produce a proportionately greater increase in the blood level, resulting in toxicity.

Concurrent administration of other drugs may affect pharmacokinetics or pharmacodynamics. Pharmacokinetics may be affected by alterations in bioavailability, protein binding, or the rate of metabolism or excretion. Pharmacodynamics may be altered by another drug that competes for the same receptor sites, that prevents the drug from reaching its site of action, or that antagonizes or enhances the drug's pharmacologic effect. Inhibition of the metabolism of one drug by another may occur when both drugs bind to the sameCYP. Therefore, as the specific CYPs responsible for the metabolism of individual drugs become known, prediction of drug interactions is put on a more rational scientific basis. An important example of such a mechanism is the inhibition of terfenadine's metabolism by inhibitors of CYP3A, such as erythromycin and systemic antimycotics. Such inhibition has resulted in torsade de pointes and lethal cardiac arrhythmia (Chap. 70).

TOXICITY UNRELATED TO A DRUG'S PRIMARY PHARMACOLOGIC ACTIVITY

Cytotoxic Reactions The understanding of so-called idiosyncratic reactions has greatly improved with the recognition that many of them are due to irreversible binding of a drug or its metabolites to tissue macromolecules by covalent bonds. Some chemical carcinogens, such as the alkylating agents, combine directly with DNA. Usually, it is only after metabolic activation of a drug to reactive metabolites that covalent binding occurs. This activation usually occurs in the microsomal mixed-function oxidase system, the hepatic enzyme system responsible for the metabolism of many drugs (Chap. 70). During the course of drug metabolism, reactive metabolites may covalently bind to tissue macromolecules, causing tissue damage. Because of the reactive nature of these metabolites, covalent binding often occurs close to the site of production. Typically that is the liver, but the mixed-function oxidase system is found in other tissues as well.

An example of this type of adverse drug reaction is the hepatotoxicity associated with isoniazid. This drug is metabolized principally by acetylation to acetylisoniazid, which is then hydrolyzed to acetylhydrazine. The further metabolism of acetylhydrazine by the mixed- function oxidase system liberates reactive metabolites that covalently bind to hepatic macromolecules, causing hepatic necrosis. The administration of drugs known to increase the activity of the mixed-function oxidase system, such as phenobarbital or rifampin, together with isoniazid results in the production of increased amounts of reactive metabolites, increased covalent binding, and a greater risk of hepatic damage.

The hepatic necrosis produced by overdosage of acetaminophen is also caused by reactive metabolites. Normally these metabolites are detoxified by combining with hepatic glutathione. When glutathione becomes exhausted, the metabolites bind instead to hepatic protein, with resultant hepatocyte damage. The hepatic necrosis produced by the ingestion of acetaminophen can be prevented, or at least attenuated, by the administration of substances such as *N*-acetylcysteine that reduce the binding of electrophilic metabolites to hepatic proteins. The risk of hepatic necrosis is increased in patients receiving drugs such as phenobarbital that increase the rate of drug metabolism and the rate of production of toxic metabolite(s).

It is likely, though as yet not proved, that other idiosyncratic reactions are caused by the covalent binding of reactive metabolites to tissue macromolecules, resulting either in direct cytotoxicity or in the initiation of an immune response.

Immunologic Mechanisms Most pharmacologic agents are poor immunogens because they are small molecules with molecular weights of less than 2000. Stimulation of antibody synthesis or sensitization of lymphocytes by a drug or one of its metabolites usually requires in vivo activation and covalent linkage to protein, carbohydrate, or nucleic acid.

Drug stimulation of antibody production may mediate tissue injury by one of several mechanisms. The antibody may attack the drug when the drug is covalently attached to a cell, and thereby destroy the cell. This mechanism occurs in penicillin-induced hemolytic anemia. Antibody-drug-antigen complexes may be passively adsorbed by a bystander cell, which is then destroyed by activation of complement; this occurs in quinine- and quinidine-induced thrombocytopenia. Drugs or their reactive metabolites may alter a host tissue, rendering it antigenic and eliciting autoantibodies. For example, hydralazine and procainamide can chemically alter nuclear material, stimulating the formation of anti-nuclear antibodies and occasionally causing lupus erythematosus. Autoantibodies can be elicited by drugs that neither interact with the host antigen nor have any chemical similarity to the host tissue; for example, a-methyldopa frequently stimulates the formation of antibodies to host erythrocytes, yet the drug neither attaches to the erythrocyte nor shares any chemical similarities with the antigenic determinants on the erythrocyte.

Drug-induced pure red cell aplasia (Chap. 109) is due to an immune-based drug reaction. Red cell formation in bone marrow cultures can be inhibited by phenytoin and purified IgG obtained from a patient with pure red cell aplasia associated with phenytoin.

Serum sickness (Chap. 310) results from the deposition of circulating drug-antibody complexes on endothelial surfaces. Complement activation occurs, chemotactic factors are generated locally, and an inflammatory response develops at the site of complex entrapment. Arthralgias, urticaria, lymphadenopathy, glomerulonephritis, or cerebritis may result. Penicillin is the most common cause of serum sickness today. Many drugs, particularly antimicrobial agents, induce production of IgE, which binds to mast cell membranes. Contact with a drug antigen initiates a series of biochemical events in the mast cell and results in the release of mediators that can produce the urticaria, wheezing, flushing, rhinorrhea, and (occasionally) hypotension characteristic of anaphylaxis.

Drugs may also excite cell-mediated immune responses. Topically administered substances may interact with sulfhydryl or amino groups in the skin and react with sensitized lymphocytes to produce the rash characteristic of contact dermatitis. Other types of rashes may also result from the interaction of serum factors, drugs, and sensitized lymphocytes. The role of drug-activated lymphocytes in the immune mechanisms governing destruction of visceral tissue is unknown.

Toxicity Associated with Genetically Determined Enzymatic Defects In the porphyrias, drugs that increase the activity of enzymes proximal to the deficient enzyme in the biosynthetic pathway of porphyrins can increase the quantity of porphyrin precursors that accumulate proximal to the deficient enzyme (Chap. 346). These drugs are listed inTable 71-1.

Patients with a deficiency of glucose-6-phosphate dehydrogenase (G6PD) develop hemolytic anemia in response to primaquine and a number of other drugs (Table 71-1) that do not cause hemolysis in patients with adequate quantities of this enzyme (Chap. 108).

DIAGNOSIS

The manifestations of drug-induced diseases frequently resemble those of other diseases, and a given set of manifestations may be produced by different and dissimilar drugs. Recognition of the role of a drug or drugs in an illness depends on appreciation of the possible adverse reactions to drugs in any disease, on identification of the temporal relationship between drug administration and development of the illness, and on familiarity with the common manifestations of the drugs. Many associations between particular drugs and specific reactions have been described, but there is always a "first time" for a novel association, and any drug should be suspected of causing an adverse effect if the clinical setting is appropriate.

Illness related to a drug's pharmacologic action is often more easily recognized than illness attributable to immune or other mechanisms. For example, side effects such as cardiac arrhythmias in patients receiving digitalis, hypoglycemia in patients given insulin, and bleeding in patients receiving anticoagulants are more readily related to a specific drug than are symptoms such as fever or rash, which may be caused by many drugs or by other factors.

Once an adverse reaction is suspected, discontinuance of the suspected drug followed by disappearance of the reaction is presumptive evidence of a drug-induced illness. Confirming evidence may be sought by cautiously reintroducing the drug and seeing if the reaction reappears. However, that should be done only if confirmation would be useful in the future management of the patient and if the attempt would not entail undue risk. With concentration-dependent adverse reactions, lowering the dosage may cause the reaction to disappear, and raising it may cause the reaction to reappear. When the reaction is thought to be allergic, however, readministration of the drug may be hazardous, since anaphylaxis may develop. Readministration is unwise under these conditions unless no alternative drugs are available and treatment is necessary.

If the patient is receiving many drugs when an adverse reaction is suspected, the drugs likeliest to be responsible can usually be identified. All drugs may be discontinued at once, or, if that is not practical, they should be discontinued one at a time, starting with the one that is most suspect, and the patient observed for signs of improvement. The time needed for a concentration-dependent adverse effect to disappear depends on the time required for the concentration to fall below the range associated with the adverse effect, and that, in turn, depends on the initial blood level and on the rate of elimination or metabolism of the drug. Adverse effects of drugs with long half-lives, such as phenobarbital, take a considerable time to disappear.

Drugs recognized as producing a number of reactions are listed inTable 71-1. This table includes both well-documented and some less well-documented reactions, focusing on those that are sufficiently important to require consideration. This information should be used to suggest the drug likely to be causing a reaction; the absence of a drug from the table does not mean that it cannot be responsible for the reaction, however.

Serum antibody has been demonstrated in some persons with drug allergies involving cellular blood elements, as in agranulocytosis, hemolytic anemia, and thrombocytopenia. For example, both quinine and quinidine can produce platelet agglutination in vitro in the presence of complement and the serum from a patient who has developed thrombocytopenia following use of this drug.

Eliciting a drug history from patients is important for diagnosis. Attention must be directed toOTCdrugs and herbal preparations as well as to prescription drugs. Each type can be responsible for adverse drug effects, and adverse interactions may occur between OTC drugs and prescribed drugs. In addition, it is common for patients to be cared for by several physicians, and duplicative, additive, counteractive, or synergistic drug combinations may therefore be administered if the physicians are not aware of the patients' drug histories. Every physician should determine what drugs a patient has been taking, at least during the preceding 30 days, before prescribing any medications. A frequently overlooked source of additional drug exposure is topical therapy; for example, a patient complaining of bronchospasm may not mention that an ophthalmic beta blocker is being used unless specifically asked. A history of previous adverse drug effects in patients is common. Since these patients have shown a predisposition to drug- induced illnesses, such a history should dictate added caution in prescribing drugs.

Patients with biochemical abnormalities such as erythrocyteG6PDdeficiency can be identified. Most patients with the G6PD defect are of African or Mediterranean descent. Drug-induced hemolytic crisis can be avoided by testing for the enzyme defect before administering drugs that could cause the reaction. Similarly, persons with an abnormal serum pseudocholinesterase level may have abnormally prolonged apnea when given succinylcholine.

GENERAL COMMENTS

No drug is completely without side effects, and a side effect in one patient may be the desired pharmacologic effect in another. Current drug regulations allow physicians to have considerable confidence in the purity, bioavailability, and effectiveness of the drugs they prescribe. However, physicians have to weigh potential toxicity against possible benefits. Toxicity that would be acceptable for an effective antineoplastic agent would not be permitted in an oral contraceptive, for example. Because of the necessarily small number of patients treated in premarketing studies, rare adverse reactions may not be identified, so the first responsibility for identifying and reporting these effects must rest with the practicing clinician through the use of the various national adverse reaction reporting systems, such as those operated by the Food and Drug Administration in the United States and the Committee on Safety of Medicines in Great Britain. The publication of a newly recognized adverse reaction can in a short time stimulate many similar such reports of reactions that previously had gone unrecognized.

The prevention of adverse drug reactions first involves a high index of suspicion that the development of a new symptom or sign may be drug-related. Reduction of the dose or discontinuation of the suspected agent usually clarifies the issue in concentration-dependent toxic reactions. Physicians should be familiar with the common adverse effects of the drugs they use and, when in doubt, should consult the literature.

(Bibliography omitted in Palm version)

Back to Table of Contents

72. PHYSIOLOGY AND PHARMACOLOGY OF THE AUTONOMIC NERVOUS SYSTEM - *Lewis Landsberg*, *James B. Young*

FUNCTIONAL ORGANIZATION OF THE AUTONOMIC NERVOUS SYSTEM

The autonomic nervous system innervates vascular and visceral smooth muscle, exocrine and endocrine glands, and parenchymal cells throughout the various organ systems. Functioning below the conscious level, the autonomic nervous system responds rapidly and continuously to perturbations that threaten the constancy of the internal environment. The many functions governed by this system include the distribution of blood flow and the maintenance of tissue perfusion, the regulation of blood pressure, the regulation of the volume and composition of the extracellular fluid, the expenditure of metabolic energy and supply of substrate, and the control of visceral smooth muscle and glands.

ANATOMIC ORGANIZATION

The autonomic neurons, located in ganglia outside the central nervous system (CNS), give rise to the postganglionic autonomic nerves that innervate organs and tissues throughout the body (Fig. 72-1). The activity of autonomic nerves is regulated by central neurons responsive to diverse afferent inputs. After central integration of afferent information, autonomic outflow is adjusted to permit the functioning of the major organ systems in accordance with the needs of the organism as a whole. Connections between the cerebral cortex and the autonomic centers in the brainstem coordinate autonomic outflow with higher mental functions.

The Sympathetic and Parasympathetic Divisions The preganglionic neurons of the parasympathetic nervous system leave theCNS in the third, seventh, ninth, and tenth cranial nerves and in the second and third sacral nerves, while the preganglionic neurons of the sympathetic nervous system exit the spinal cord between the first thoracic and the second lumbar segments (Fig. 72-1). Responses to sympathetic and parasympathetic stimulation are frequently antagonistic, as exemplified by their opposing effects on heart rate and gut motility. This antagonism reflects highly coordinated interactions within the CNS; the resultant changes in parasympathetic and sympathetic activity, often reciprocal, provide more precise control of autonomic responses than could be achieved by the modulation of a single system. Moreover, both sympathetic and parasympathetic portions of the autonomic nervous system are composed of multiple function-specific subdivisions. Neurons with the various subdivisions differ neurochemically and neurophysiologically and are controlled by distinct regions within the CNS. This specialization within sympathetic and parasympathetic divisions contributes to the precision and specificity of autonomic regulation.

Neurotransmitters *Acetylcholine* (ACh) is the preganglionic neurotransmitter for both divisions of the autonomic nervous system as well as the postganglionic neurotransmitter of the parasympathetic neurons. Nerves that release ACh are said to be cholinergic. *Norepinephrine* (NE) is the neurotransmitter of the postganglionic sympathetic neurons; these nerves are said to be adrenergic. Within the sympathetic outflow, postganglionic neurons innervating the eccrine sweat glands (and perhaps some blood vessels supplying skeletal muscle) are of the cholinergic type.

THE SYMPATHETIC NERVOUS SYSTEM AND ADRENAL MEDULLA

CATECHOLAMINES

All three of the naturally occurring catecholamines,NE, *epinephrine* (E), and *dopamine*, function as neurotransmitters within theCNS. NE, the neurotransmitter of postganglionic sympathetic nerve endings, exerts its effects locally, in the immediate vicinity of its release. Epinephrine, the circulating hormone of the adrenal medulla, influences processes throughout the body. A peripheral dopaminergic system also exists but has not been characterized in detail.

Biosynthesis (Fig. 72-2) Catecholamines are synthesized from the amino acid tyrosine, which is sequentially hydroxylated to form dihydroxyphenylalanine (dopa), decarboxylated to form dopamine, and hydroxylated on the b position of the side chain to formNE. The initial step, the hydroxylation of tyrosine, is rate-limiting and is regulated so that synthesis of dopa is coupled to NE release. This regulation is achieved by alterations in both the activity and the amount of tyrosine hydroxylase. In the adrenal medulla and in those central neurons utilizing epinephrine as neurotransmitter, NE is *N*-methylated to epinephrine by the enzyme phenylethanolamine-*N*-methyltransferase (PNMT).

Catecholamine Metabolism The major metabolic transformations of catecholamines involve *O*-methylation at the meta-hydroxyl group and oxidative deamination. *O*-Methylation is catalyzed by the enzyme catechol-*O*-methyltransferase (COMT), and oxidative deamination is promoted by monoamine oxidase (MAO). COMT in liver and kidney is important in the metabolism of circulating catecholamines. MAO, a mitochondrial enzyme present in most tissues, including nerve endings, has a lesser role in the metabolism of circulating catecholamines but is important in regulating the catecholamine stores within the peripheral sympathetic nerve endings. The metanephrines and 3-methoxy-4-hydroxymandelic acid (vanilmandelic acid, VMA) are the major end products of E andNEmetabolism. Homovanillic acid (HVA) is the end product of dopamine metabolism.

STORAGE AND RELEASE OF CATECHOLAMINES

In both the adrenal medulla and sympathetic nerve endings catecholamines are stored in subcellular vesicles and released by exocytosis. The large stores of catecholamines in these tissues provide an important physiologic reserve that maintains an adequate supply of catecholamines in the face of intense stimulation. A variety of substances may be stored along with catecholamines in sympathetic nerve endings and adrenal medulla and released with catecholamines during exocytosis. These substances, which may function as cotransmitters or neuromodulators, include peptides such as neuropeptide Y, substance P, and enkephalins; purines such as ATP and adenosine; and other amines such as serotonin. At the neuroeffector junction, coreleased neuromodulators modify the response toNE, while cotransmitters exert physiologic effects independent of those induced by NE.

Adrenal Medulla The adrenal medullary chromaffin tissue in a pair of normal human adrenal glands weighs about 1 g and contains approximately 6 mg catecholamines, 85% of which is epinephrine.

Catecholamine secretion, stimulated byACh from the preganglionic sympathetic nerves, occurs after calcium influx triggers fusion of the chromaffin granule membrane and cell membrane; obliteration of the cell membrane at the point of fusion and extrusion of the entire soluble contents of the granule into the extracellular space complete the process of exocytosis (Fig. 72-2). Although the molecular mechanisms involved in the exocytotic process are only partially understood, evidence has accumulated that specific calcium-binding proteins are involved. Once bound, calcium induces a conformational change in these proteins that induces fusion of granules and docking of granules at the cell membrane.

Peripheral Sympathetic Nerve Endings The peripheral sympathetic nerve endings form a reticulum or ground plexus that brings the terminal fibers into close contact with effector cells. All theNE in peripheral tissues is in the sympathetic nerve endings, and heavily innervated tissues contain as much as 1 to 2 ug/g of tissue. NE stored in the nerve endings is in discrete subcellular particles analogous to the adrenal medullary chromaffin granules.MAO in the mitochondria of the nerve endings plays an important role in regulating the local concentration of NE (Fig. 72-2). Amines in storage vesicles are protected from oxidative deamination; amines within the cytoplasm, however, are deaminated to inactive metabolites. Release from the nerve ending occurs in response to action potentials propagated in terminal sympathetic fibers.

THE PERIPHERAL ADRENERGIC NEUROEFFECTOR JUNCTION

The peripheral sympathetic nerve endings possess an amine transport system that actively takes up amines from the extracellular fluid. Neuronal uptake or recapture of locally releasedNEterminates the action of the transmitter and contributes to the constancy of the NE stores.

A variety of factors alter the relationship between neuronal impulse traffic andNErelease. Diminished temperature and acidosis, for example, both decrease the amount of NE released in response to sympathetic impulses. Several chemical mediators operate at the peripheral sympathetic nerve ending (referred to as *prejunctional* or *presynaptic sites*) to modify sympathetic neurotransmission by influencing the amount of NE released in response to nerve impulses. Prejunctional modulation may be either inhibitory or facilitatory. Certain modulators, such as catecholamines andACh, may either inhibit or facilitate NE release, antagonistic effects that are mediated by different adrenergic or cholinergic receptors, respectively. Those compounds exerting an *inhibitory* effect on NE release at the prejunctional nerve ending include the following: catecholamines ($a_2$receptor), ACh (muscarinic receptor), dopamine (D_2receptor), histamine (H_2receptor), serotonin, adenosine, enkephalins, and prostaglandins.

Catecholamines reduceNErelease via prejunctional a receptors in a classic negative-feedback system. Feedback regulation is complicated by the fact thatb-receptor activation facilitates NE release.

Though both inhibitory and facilitatory effects ofACh onNErelease have been described, the inhibitory effect of ACh, mediated by the muscarinic cholinergic receptor, occurs at lower ACh concentrations and is probably of greater physiologic significance.

CENTRAL REGULATION OF SYMPATHOADRENAL OUTFLOW

Brainstem Sympathetic Centers Sympathetic outflow is initiated from the reticular formation of the medulla oblongata and pons and from centers in the hypothalamus. The rostral ventral portion of the medulla, particularly the area designated the rostral ventrolateral medulla (RVLM), appears to contain especially important sympathoexcitatory areas. Descending fibers originating from all these centers synapse in the intermediolateral cell column of the spinal cord with the preganglionic sympathetic neurons. Changes in the physical and chemical properties of the extracellular fluid, including the circulating levels of hormones and substrates, also affect sympathetic nervous system outflow. The area postrema, in the floor of the fourth ventricle, along with other circumventricular organs lie outside the blood-brain barrier and may play an important role in this regard. Although the hallmark of intense sympathoadrenal stimulation is a global response (the fight-or-flight reaction of Cannon), discrete changes in sympathetic outflow to different organ systems continuously regulate many autonomic functions.

Relationship Between the Sympathetic Nervous System and the Adrenal Medulla Sympathetic nervous system activity and adrenal medullary secretion are coordinated but not always congruent. During periods of intense sympathetic stimulation, such as cold exposure and exhaustive exercise, the adrenal medulla is progressively recruited, and circulating epinephrine reinforces the physiologic effects of sympathetic stimulation. In other situations, the sympathetic nervous system and the adrenal medulla are stimulated independently. The response to upright posture, for example, involves predominantly the sympathetic nervous system, while hypoglycemia stimulates only the adrenal medulla.

Sympathetic Regulation of the Cardiovascular System Stretch receptors in the systemic and pulmonary arteries and veins continuously monitor intravascular pressures; the resulting afferent impulses, after relay and integration in the brainstem, alter sympathetic activity in defense of blood pressure and blood flow to critical areas (Fig. 72-3).

Arterial Baroreceptors An increase in blood pressure stimulates receptors in the carotid sinus and aortic arch. The ensuing afferent impulses, after relay within the nucleus of the solitary tract (NTS) in the brainstem, suppress the brainstem sympathetic centers (Fig. 72-3). This baroreceptor reflex arc forms a negative-feedback loop in which a rise in arterial pressure results in the inhibition of central sympathetic outflow. A brainstem noradrenergic pathway interacts with the NTS to participate in suppression of sympathetic outflow. This noradrenergic inhibitory pathway is stimulated by centrally actinga-adrenergic agonists and may be involved in the action of certain antihypertensive drugs, such as clonidine, that potentiate the baroreceptor-mediated vasodepressor response (Chap. 246). In the opposite manner, when the blood pressure falls, decreased afferent impulses diminish central inhibition, resulting in an increase in sympathetic outflow and a rise in arterial pressure.

Central Venous Pressure Receptors in the walls of the great veins and within the atria are also involved in the regulation of sympathetic outflow. Stimulation of these receptors by high venous pressure suppresses the brainstem sympathetic centers; when central venous pressure is low, sympathetic outflow increases. The central connections are poorly understood, but the afferent impulses are carried in the vagus (Fig. 72-3).

ASSESSMENT OF SYMPATHOADRENAL ACTIVITY

The clinical assessment of sympathoadrenal activity involves the measurement of catecholamines in plasma and of catecholamines and catecholamine metabolites in urine, and the assessment of sympathetic nerve impulse traffic by microneurography. Microneurography, utilizing microelectrodes implanted in nerves supplying skeletal muscle (such as the peroneal nerve), is primarily a research tool. Quantitation of urinary catecholamines and metabolites is useful in the diagnosis of pheochromocytoma (Chap. 332).

Plasma Catecholamines Catecholamines in human plasma may be measured by radioenzymatic isotope derivative techniques or by high-performance liquid chromatography in conjunction with electrochemical detection. Plasma catecholamine measurements provide an index of sympathetic nervous system and adrenal medullary activity and have been widely used to assess sympathoadrenal activity in clinical investigation in human subjects. The usefulness of plasma catecholamine measurements, however, is compromised by factors that alter the relationship between the plasma concentration of catecholamines and the functional state of the sympathoadrenal system, and also by important regional differences in sympathetic outflow. Techniques utilizing tracer infusions of tritiatedNE, which correct for changes in NE clearance when applied across a particular anatomic region, estimate regional sympathetic outflow with some precision and have helped to define differentiated sympathetic nervous system activity in the investigational setting. The clinical usefulness of plasma catecholamine levels remains limited to the evaluation of patients with autonomic insufficiency and, on occasion, patients with suspected pheochromocytoma (Chap. 332).

Basal plasmaNEconcentrations are in the range of 0.09 to 1.8 nmol/L (150 to 350 pg/mL); basal E levels are about 135 to 270 pmol/L (25 to 50 pg/mL). The half-time of disappearance of NE from the circulation is approximately 2 min. The plasma NE level is markedly affected by a variety of factors, including posture; accordingly, the conditions under which blood is obtained for assay must be controlled. By convention, basal plasma NE levels are those obtained through an indwelling intravenous line after the patient has rested supine in a relaxed environment for 30 min.

Plasma NE response to upright posture The predictable increase in circulating NE concentration during upright posture provides a convenient test of sympathetic nervous system function. Five minutes of quiet standing results in a two- to threefold increase in plasma NE level. A normal response requires an intact afferent system, appropriateCNSrelays, and an intact peripheral sympathetic nervous system; a defect of any of these components reduces the increment in circulating NE.

Plasma E levels are also dependent on the physical and mental state of the subject. Change in plasma E with upright posture is usually small. Hypoglycemia, strenuous exercise, and various types of mental stress, however, can cause large increases in the plasma E level.

PERIPHERAL DOPAMINERGIC SYSTEM

In addition to its role as neurotransmitter in theCNS, dopamine functions as an inhibitory transmitter in the carotid body and the sympathetic ganglia. A distinct peripheral dopaminergic system is also believed to exist. Dopamine elicits a variety of responses not attributable to stimulation of classic adrenergic receptors; it relaxes the lower esophageal sphincter, delays gastric emptying, causes vasodilation in the renal and mesenteric arterial circulation, suppresses aldosterone secretion, directly stimulates renal sodium excretion, and suppressesNErelease at sympathetic nerve terminals by a presynaptic inhibitory mechanism. The mediation of these dopaminergic effects in vivo is poorly understood. Dopamine does not appear to be a circulating hormone.

ADRENERGIC RECEPTORS

Catecholamines influence effector cells by interacting with specific surface *receptors* coupled to G proteins. Two major categories of response to catecholamines reflect the activation of two populations of adrenergic receptors, designateda and b. Botha and b receptors have been further divided into subtypes that serve different functions and are susceptible to differential stimulation and blockade.

a-ADRENERGIC RECEPTORS

a-Adrenergic receptors mediate vasoconstriction, intestinal relaxation, and pupillary dilatation. Epinephrine and NEare approximately equipotent as a-receptor agonists. Distincta_1- anda_2-receptor subtypes are also recognized. Originally the postsynaptic or postjunctionala-adrenergic receptors on effector cells were designateda_1, while the prejunctional a-adrenergic receptors on the sympathetic nerve endings were designateda_2. It is now recognized that nonneuronal (postsynaptic) processes are mediated by the$a_2$receptor as well. The$a_1$receptor mediates the classic a effects, including vasoconstriction; phenylephrine and methoxamine are selective$a_1$agonists, and prazosin is a selective$a_1$antagonist. The$a_2$receptor mediates presynaptic inhibition of NE release from adrenergic nerves and other responses, including inhibition ofAChrelease from cholinergic nerves, inhibition of lipolysis in adipocytes, inhibition of insulin secretion, stimulation of platelet aggregation, and vasoconstriction in some vascular beds. Specific$a_2$agonists include clonidine and a-methylnorepinephrine; these agents, the latter derived from a-methyldopa in vivo, exert an antihypertensive effect by interacting with$a_2$receptors within the brainstem sympathetic centers that regulate blood pressure. Yohimbine is a specific$a_2$antagonist.

b-ADRENERGIC RECEPTORS

Physiologic events associated with b-adrenergic receptor responses include stimulation of heart rate and contractility, vasodilation, bronchodilation, and lipolysis. b-Receptor responses can also be divided into two types. Theb$_1$receptor responds equally to E andNE and mediates cardiac stimulation and lipolysis. Theb$_2$receptor is more responsive to E than to NE and mediates responses such as vasodilation and bronchodilation. Isoproterenol stimulates and propranolol blocks bothb$_1$ andb$_2$receptors. Other agonists and antagonists that have partial selectivity for theb$_1$ orb$_2$receptors have been used therapeutically where the desired response involves predominantly one of the two subtypes.

Both pharmacologic and molecular genetic studies have demonstrated an additional distinctb$_3$-adrenergic receptor that subserves lipolysis in white and brown adipose tissue as well as the stimulation of heat production in brown adipose tissue. The humanb$_3$-adrenergic receptor has been cloned, and a distinct polymorphism noted that may, in some populations, be associated with weight gain, insulin resistance, and type 2 diabetes mellitus. Theb$_3$-adrenergic receptor has a much greater affinity forNE than E and, unlike theb$_1$ andb$_2$receptors, does not undergo desensitization. Synthetic agonists for theb$_3$receptor, currently under development, have a potential role in the treatment of obesity by increasing metabolic rate.

DOPAMINERGIC RECEPTORS

Specific dopaminergic receptors, distinct from the classic a- andb-adrenergic receptors, are present in theCNS and peripheral nervous system and in several nonneural tissues. Two types of dopaminergic receptors serve different functions and have different second messengers. Dopamine is a potent agonist of both types of receptors; the action of dopamine is antagonized by phenothiazines and thioxanthenes. The D$_1$receptor mediates vasodilation in the renal, mesenteric, coronary, and cerebral vascular beds. Fenoldopam is an agonist selective for the D$_1$receptor. The D$_2$receptor inhibits transmission in the sympathetic ganglia, inhibitsNErelease from sympathetic nerve endings by an effect on the presynaptic membrane (Fig. 72-2), inhibits prolactin release from the pituitary, and causes vomiting. Selective agonists of the D$_2$receptor include bromocriptine, cabergoline, and apomorphine, while butyrophenones such as haloperidol (active within the CNS), domperidone (does not cross blood-brain barrier readily), and the benzamide sulpiride are relatively selective D$_2$antagonists.

STRUCTURE AND FUNCTION OF ADRENERGIC RECEPTORS

Adrenergic receptors belong to a superfamily of related membrane proteins, including the visual protein rhodopsin and the muscarinic acetylcholine receptors, that interacts with G proteins. These proteins share significant sequence homologies and, as deduced from the properties of the constituent amino acids, a similar topographic structure in the cell membrane. The postulated structure of this family of receptor proteins is shown schematically inFig. 72-4. The characteristic features include seven membrane-spanning hydrophobic domains containing 20 to 28 amino acids each. The membrane-spanning domains, particularly M-7 (Fig. 72-4), appear to be important in determining the characteristic agonist binding.

Coupling of Receptor Occupancy with Cellular Response The major mediators of adrenergic (as well as many other) cellular responses are a family of regulatory proteins termed *G proteins* that, when activated, bind the nucleotide guanosine triphosphate

(GTP). The best-characterized G proteins are those that stimulate or inhibit adenylyl cyclase, designated G_s or G_i, respectively (Fig. 72-5). Theb_1,b_2, and D_1receptors are coupled to G_s; receptor occupancy is therefore associated with stimulation of adenylyl cyclase and results in an increase in intracellular cyclic adenosine monophosphate (AMP), which in turn results in activation of protein kinase A and other cyclic AMP-dependent protein kinases. The resultant protein phosphorylation alters the activity of enzymes and the function of other proteins, culminating in a cellular response that is characteristic of the tissue being stimulated. Thea_2, M_2subtype of the muscarinic acetylcholine receptor and the D_2receptor are coupled to G_i, resulting in diminished adenylyl cyclase activity and a fall in cyclic AMP. The subsequent alterations in enzyme activity and function of other proteins produce an alternate, frequently opposite, series of cellular responses. Although many$a_2$responses can be explained by inhibition of adenylyl cyclase, other mechanisms may be involved as well.

Thea_1-adrenergic receptor (as well as the M_1subtype of the acetylcholine receptor) is coupled to a different G protein that activates phospholipase C; this G protein has not been as well characterized but is sometimes designated G_q. Receptor occupancy in this system stimulates phospholipase C, which catalyzes the breakdown of membrane-bound phospholipids, particularly phosphatidylinositol-4,5-bisphosphate (PIP_2) with the production of inositol-1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerol (DAG), both of which act as second messengers (Fig. 72-5).IP_3rapidly mobilizes calcium from intracellular stores within the endoplasmic reticulum, producing an increase in free cytoplasmic calcium which by itself and via calcium-calmodulin-dependent protein kinases influences cellular processes appropriate to the stimulated cell. The transient rise in calcium induced by IP_3 from the intracellular stores is reinforced in the presence of continued agonist stimulation by alterations in membrane calcium flux that result eventually in net calcium uptake from the extracellular fluid by mechanisms that have been incompletely defined.

DAG, the other second messenger produced by the action of phospholipase C onPIP_2(as well as other membrane phospholipids), remains associated with the cell membrane and activates protein kinase C, which has different substrates than the calcium-calmodulin kinases stimulated byIP_3. Protein phosphorylation stimulated by protein kinase C contributes to the tissue-specific response in ways that remain poorly understood. Increases in intracellular calcium also potentiate the activation of protein kinase C (Fig. 72-5).

REGULATION OF ADRENERGIC RECEPTORS

Prolonged exposure to a- or b-adrenergic agonists decreases the number of corresponding adrenergic receptors on effector cells. Although the biochemical mechanisms involved are obscure, internalization of the b-adrenergic receptor within the cell occurs during agonist exposure in some systems, suggesting that internal translocation contributes to the decrease in receptor number under these circumstances.

Alteration in agonist concentration may also affect the affinity of the receptor for the agonist. Adrenergic receptors that utilize adenylyl cyclase for the second messenger (b receptors,$a_2$receptors) exist in high- and low-affinity states; exposure to agonist diminishes the proportion of receptors in the high-affinity state. Such alterations in adrenergic receptors induced by adrenergic agonists are termed *homologous regulation*. Agonist-induced alterations in adrenergic-receptor density and affinity are believed to contribute to the diminished physiologic response that occurs after prolonged exposure of an effector tissue to adrenergic agonist, a phenomenon known as *tachyphylaxis* or *desensitization*.

Adrenergic receptors are also influenced by factors other than adrenergic agonists, so-called *heterologous regulation*. Enhanceda-adrenergic-receptor affinity, for example, may underlie the potentiation of a-adrenergic responses that occur in response to lowered environmental temperatures. Thyroid hormones potentiate b-receptor responses by alterations inb-receptor number and in the efficiency of coupling receptor occupancy with physiologic response. Estrogen and progesterone alter the sensitivity of the myometrium to catecholamines by effects ona-adrenergic receptors. Glucocorticoids may influence adrenergic function by antagonizing agonist-induced decreases in adrenergic receptors, thereby counteracting tachyphylaxis in response to intense adrenergic stimulation.

Alterations in sensitivity to catecholamines also occur as a consequence of postreceptor changes, although the latter remain poorly characterized.

PHYSIOLOGY OF THE SYMPATHOADRENAL SYSTEM

Catecholamines influence all of the major organ systems. The effects take place in seconds and may occur in anticipation of physiologic requirement. An increase in sympathoadrenal activity prior to strenuous exercise, for example, lessens the impact of exercise on the internal environment.

DIRECT EFFECTS OF CATECHOLAMINES

Cardiovascular System Catecholamines stimulate vasoconstriction in the subcutaneous, mucosal, splanchnic, and renal vascular beds bya-receptor-mediated mechanisms. Although vasoconstriction was originally considered ana$_1$-receptor response, vascular tone appears to be more complexly regulated and, in many areas, involvesa$_2$-mediated responses as well. The venous portion of the circulation, in particular, is endowed witha$_2$receptors. Differential regulation of the two types ofa receptors, under certain circumstances, contributes to an integrated physiologic response. Since vasoconstriction in the coronary and cerebral circulations is minimal, flow to these areas is maintained during sympathetic stimulation. Skeletal muscle vasculature containsb receptors sensitive to low circulating levels of epinephrine so that skeletal muscle blood flow is augmented during adrenal medullary activation.

The effects of catecholamines on the heart are mediated byb$_1$receptors and include increase in heart rate, enhancement of cardiac contractility, and increase in conduction velocity. The increase in myocardial contractility is illustrated by a leftward and upward shift of the ventricular function curve (see Fig. 231-6) that relates cardiac work to ventricular diastolic fiber length; at any initial fiber length, catecholamines increase cardiac work. Catecholamines also enhance cardiac output by stimulating venoconstriction, enhancing venous return, and increasing the force of atrial contraction, thereby augmenting diastolic volume and hence fiber length. The acceleration of conduction in the junctional tissues results in a more synchronous, and hence more effective, ventricular contraction. Cardiac stimulation increases myocardial oxygen consumption, a major factor in the pathogenesis and treatment of myocardial ischemia.

Metabolism Catecholamines increase metabolic rate. In small mammals, mitochondrial respiration in brown adipose tissue is functionally uncoupled byNE. In a reaction unique to brown adipose tissue, NE stimulates theb$_3$-adrenergic receptor that activates a specific mitochondrial uncoupling protein that dissipates the proton gradient between the inner mitochondrial matrix and the cytoplasm, thereby uncoupling substrate utilization and ATP synthesis. In humans, a functional role for brown adipose tissue has not been established with certainty.

Substrate Mobilization In a variety of tissues, catecholamines stimulate the breakdown of stored fuel with the production of substrate for local consumption; glycogenolysis in the heart, for example, provides substrate for immediate metabolism by the myocardium. Catecholamines also accelerate fuel mobilization in liver, adipose tissue, and skeletal muscle, liberating substrates (glucose, free fatty acids, lactate) into the circulation for use throughout the body.

Fluids and Electrolytes By a direct action on the renal tubule,NEstimulates sodium reabsorption, thereby defending extracellular fluid volume. Dopamine, in contrast, promotes sodium excretion. NE and E also promote cellular uptake of potassium.

Viscera Catecholamines affect visceral function by actions on smooth muscle and glandular epithelium. Urinary bladder and intestinal smooth muscle are relaxed while the corresponding sphincters are stimulated. Gallbladder emptying also involves sympathetic mechanisms. Catecholamine-mediated smooth-muscle contraction in the female aids ovulation and ovum transport along the fallopian tubes, and in the male provides propulsive force for the seminal fluid during ejaculation. Inhibitorya$_2$receptors on cholinergic neurons within the gut contribute to intestinal relaxation. Catecholamines induce bronchodilation by ab$_2$-receptor mechanism.

INDIRECT EFFECTS OF CATECHOLAMINES

The ultimate physiologic response induced by catecholamines involves changes in hormone secretion and in blood flow distribution, both of which support and amplify the direct effects of catecholamines.

Endocrine System Catecholamines influence the secretion of renin, insulin, glucagon, calcitonin, parathormone, thyroxine, gastrin, erythropoietin, progesterone, and, possibly, testosterone. Secretion of each of these hormones is governed by complex feedback loops. With the exception of thyroxine and the gonadal steroids, each is a polypeptide not under the direct control of the pituitary gland. Sympathoadrenal input into the secretion of these hormones provides a mechanism for regulation by theCNS and ensures a coordinated hormonal response in accord with the homeostatic needs of the organism.

Renin (See alsoChap. 246) Sympathetic stimulation increases renin release by a direct b-receptor effect independent of vascular changes within the kidney. The renin response to volume depletion is sympathetically mediated and is initiated by a fall in central venous pressure. Since renin secretion activates the angiotensin-aldosterone system, angiotensin-induced vasoconstriction supports the direct effects of catecholamines on blood vessels, while aldosterone-mediated sodium reabsorption complements the direct increase in sodium reabsorption induced by sympathetic stimulation. b-receptor blocking agents suppress renin secretion.

Insulin and Glucagon Stimulation of pancreatic sympathetic nerves or an elevation in circulating catecholamines suppresses insulin and increases glucagon release. Inhibition of insulin secretion is mediated by thea$_2$receptor, and stimulation of glucagon is mediated by theb receptor. This combination of effects supports substrate mobilization, reinforcing the direct effects of catecholamines on hepatic glucose output and lipolysis. Althougha-receptor-mediated suppression of insulin release usually predominates, a b-receptor mechanism may augment insulin secretion under some circumstances.

SYMPATHOADRENAL FUNCTION IN SELECTED PHYSIOLOGIC AND PATHOPHYSIOLOGIC STATES

Support of the Circulation The sympathetic nervous system functions to maintain an adequate circulation. During upright posture and volume depletion, reduction of afferent venous and arterial baroreceptor impulse traffic diminishes an inhibitory input to the vasomotor center, thereby increasing sympathetic activity (Fig. 72-3) and reducing efferent vagal tone. As a result, heart rate is increased, and cardiac output is diverted from the skin, subcutaneous tissues, mucosa, and viscera. Sympathetic stimulation of the kidney increases sodium reabsorption, and sympathetically mediated venoconstriction enhances venous return (Fig. 72-6). With pronounced hypotension, the adrenal medulla is recruited and epinephrine reinforces the effects of the sympathetic nervous system.

The intense sympathoadrenal stimulation that accompanies severe volume depletion may contribute to the development of ketoacidosis in alcoholics as well as to the ketoacidosis sometimes seen in association with hyperemesis gravidarum. Catecholamine-mediated suppression of insulin and stimulation of glucagon markedly potentiate ketogenesis in these disease states. Volume resuscitation and provision of adequate glucose promptly reverse the ketoacidosis in most cases.

Congestive Heart Failure The sympathetic nervous system also provides circulatory support during congestive heart failure (Chap. 232). Venoconstriction and sympathetic stimulation of the heart increase cardiac output while peripheral vasoconstriction directs blood flow to the heart and brain. The afferent signals are less clear than in simple volume depletion because the venous pressure is usually elevated. In severe heart failure, depletion of cardiacNE may impair the effectiveness of sympathetic circulatory support. On the other hand, the possibility has been raised that intense sympathetic stimulation may further impair cardiac function, suggesting possible benefit fromb-adrenergic blockade. The use of beta blockers in the treatment of congestive heart failure, in fact, has increased in recent years.

Trauma and Shock In acute traumatic injury or shock, adrenal catecholamines support the circulation and mobilize substrates. In the chronic, reparative phase following injury, catecholamines also contribute to substrate mobilization and to the elevation in metabolic rate.

Exercise Sympathetic activation during exercise increases cardiac output and ensures sufficient substrate to meet the increased metabolic needs. Central neural factors, such as anticipation, and circulatory factors, such as fall in venous pressure, trigger the sympathetic response. Mild degrees of exercise stimulate the sympathetic nervous system alone; during more severe exertion the adrenal medulla is activated as well. Cardiovascular conditioning is associated with a decrease in sympathetic nervous system activity both at rest and during exercise, in comparison with the untrained state.

Hypoglycemia (See alsoChap. 334) Hypoglycemia causes a marked increase in adrenal medullary epinephrine secretion. When glucose concentrations fall below overnight fasting levels, regulatory glucose-sensitive neurons in theCNSinitiate a prompt increase in adrenal medullary secretion. The increase is especially intense at plasma glucose levels below 2.8 mmol/L (50 mg/dL), when plasma E levels increase 25 to 50 times above baseline, thereby increasing hepatic glucose output, providing alternative substrate in the form of free fatty acids, suppressing endogenous insulin release, and inhibiting insulin-mediated glucose utilization in muscle. Many clinical manifestations of hypoglycemia, such as tachycardia, palpitations, nervousness, tremor, and widened pulse pressure, are secondary to increased E secretion, and these manifestations constitute an "early warning" system in insulin-requiring diabetics. In patients with long-standing diabetes mellitus, however, the E response to hypoglycemia may be diminished or absent, leaving affected patients at greater risk to develop severe hypoglycemia.

Cold Exposure The sympathetic nervous system plays a critical role in the maintenance of normal body temperature during exposure to a cold environment. Receptors in the skin and CNSrespond to a fall in temperature by activating hypothalamic and brainstem centers that increase sympathetic activity. Sympathetic stimulation leads to vasoconstriction in the superficial vascular beds, thereby diminishing heat loss. The sympathetic response involves a complex interaction between lowered environmental temperatures anda$_2$-adrenergic receptors. Acclimatization during chronic cold exposure increases the capacity for metabolic heat production in response to sympathetic stimulation.

Dietary Intake Fasting suppresses and overfeeding stimulates the sympathetic nervous system. The reduction in sympathetic activity during fasting or starvation contributes to the decrease in metabolic rate, bradycardia, and hypotension in these states. Enhanced sympathetic activity during periods of increased caloric intake contributes to the elevation in metabolic rate associated with a chronic increase in dietary intake.

Hypoxia Chronic hypoxia is associated with stimulation of the sympathoadrenal system, and some of the cardiovascular changes attendant to hypoxia are dependent on catecholamines.

Orthostatic Hypotension The maintenance of arterial pressure during upright posture depends on an adequate blood volume, an unimpaired venous return, and an intact sympathetic nervous system. Significant postural hypotension, therefore, often reflects extracellular fluid volume depletion or dysfunction of the circulatory reflexes. Diseases of the nervous system, such as tabes dorsalis, syringomyelia, or diabetes mellitus, may disrupt these sympathetic reflexes with resultant orthostatic hypotension. Although any antiadrenergic agent may impair the postural sympathetic response, orthostatic hypotension is most prominent with drugs that block neurotransmission within the ganglia or adrenergic neurons.

The term *idiopathic orthostatic hypotension* refers to a group of degenerative diseases involving either the pre- or postganglionic sympathetic neurons (Chaps. 21 and365).

Treatment of orthostatic hypotension is usually unsatisfactory except in the mildest cases. There is no way of reestablishing the normal relationship between fall in venous return and sympathetic neuronal activation. Volume expansion with fludrocortisone and a liberal salt diet in conjunction with fitted stockings to the waist, as well as elevation of the head of the bed to avoid recumbency, will maintain plasma volume and venous return and frequently provide symptomatic improvement.

PHARMACOLOGY OF THE SYMPATHOADRENAL SYSTEM

A variety of therapeutic agents affect sympathetic nervous system function or interact with adrenergic receptors, making it possible to stimulate or suppress effects mediated by catecholamines with some degree of specificity (Table 72-1).

SYMPATHOMIMETIC AMINES

Sympathomimetic amines may directly activate adrenergic receptors (direct acting) or releaseNE from the sympathetic nerve endings (indirect acting). Many agents have both direct and indirect effects.

Epinephrine and Norepinephrine The naturally occurring catecholamines act predominantly by the direct stimulation of adrenergic receptors.NE is employed to support the circulation and elevate the blood pressure in hypotensive states (Chap. 38). Peripheral vasoconstriction is the major effect, although cardiac stimulation occurs as well. Epinephrine, also employed as a pressor, has special usefulness in the treatment of allergic reactions, especially those associated with anaphylaxis. Epinephrine antagonizes the effects of histamine and other mediators on vascular and visceral smooth muscle and is useful in the treatment of bronchospasm.

Dopamine *Dopamine* is used in treating hypotension, shock (Chap. 38), and certain forms of heart failure (Chap. 232). At low infusion rates it exerts a positive inotropic effect both by a direct action on the cardiacb$_1$receptors and by the indirect release ofNE from sympathetic nerve endings in the heart. At low doses direct stimulation of dopaminergic receptors in the renal and mesenteric vasculature also results in vasodilation in the gut and kidney and facilitates sodium excretion. Athigher infusion rates interaction with a-adrenergic receptors results in vasoconstriction, an increase in peripheral resistance, and an elevation of blood pressure.

b-Receptor Agonists *Isoproterenol*, a direct-acting b-receptor agonist, stimulates the heart, decreases peripheral resistance, and relaxes bronchial smooth muscle. It raises the cardiac output and accelerates atrioventricular conduction while increasing the automaticity of ventricular pacemakers. Isoproterenol was formerly used in the treatment of heart block and bronchoconstriction. *Dobutamine*, a congener of dopamine with relative selectivity for theb$_1$receptor and with a greater effect on myocardial contractility than on heart rate, is also used in the treatment of congestive heart failure, often in combination with vasodilators (Chap. 232). In conjunction with radionuclide imaging or echocardiographic assessment of wall motion, dobutamine, as well as other investigational congeners that have a relatively greater effect on heart rate, is used in the diagnosis of demand-induced myocardial ischemia.

Selectiveb$_2$-Receptor Agonists The cardiac stimulation caused by nonselectiveb agonists, such as isoproterenol or epinephrine, is occasionally dangerous when these agents are used in the treatment of bronchoconstriction (Chap. 252). Selectiveb$_2$agonists, administered by inhalation for bronchoconstriction, include agents with an intermediate duration of action (*metaproterenol*, *albuterol*, *terbutaline*, *pirbuterol*, *isoetharine*, and *bitolterol*) and the newer long-acting agents (*salmeterol* and *formoterol*); these drugs improve the therapeutic ratio by achieving bronchial dilatation with less activation of the cardiovascular system (Chaps. 252 and258). Although selectivity is relative and cardiac stimulation may occur with these agents at higher dose levels, inhaled agonists at the usual doses result in relatively little cardiac stimulation. Oral administration, which is no longer preferred, is associated with more systemic b-agonist effects. *Ritodrine*, another selectiveb$_2$agonist, is used as a tocolytic agent (as is *terbutaline*) to relax the uterus and antagonize premature labor.

a-Adrenergic Agonists *Phenylephrine* and *methoxamine* are direct-actinga agonists that elevate blood pressure by increasing peripheral vasoconstriction. They are used primarily in the treatment of hypotension and paroxysmal supraventricular tachycardia (Chap. 230), in the latter case by increasing cardiac vagal tone through reflex baroreceptor stimulation. Phenylephrine and a related proprietary compound, *phenylpropanolamine*, are common constituents of decongestant medications (often combined with antihistamines) for the treatment of allergic rhinitis and upper respiratory infections.

Miscellaneous Sympathomimetic Amines with Mixed Actions *Ephedrine* has both directb-receptor agonist properties and an indirect effect on sympathetic nerve endings, from which it releasesNE, and is used primarily as a bronchodilator. *Sudephedrine*, a congener of ephedrine, is less potent at dilating bronchi and serves as a nasal decongestant. *Metaraminol* has both direct and indirect effects on sympathetic nerve endings and is employed in the treatment of hypotensive states.

Dopaminergic Agonists The D$_2$-receptor agonists, *bromocriptine* and *cabergoline*, are used to suppress prolactin secretion (Chap. 328). *Apomorphine*, another D$_2$-receptor agonist, is used to induce emesis. The D$_1$receptor agonist, *fenoldapam*, has recently been approved for the short-term in-hospital treatment of severe hypertension.

ANTIADRENERGIC OR SYMPATHOLYTIC AGENTS (See alsoChap. 246)

Agents Inhibiting Central Sympathetic Outflow The antihypertensive agents *methyldopa*, *clonidine*, *guanabenz*, and *guanfacine* diminish central sympathetic outflow by stimulating a centrala-adrenergic pathway (a$_2$receptor) that diminishes vasomotor outflow.CNS side effects such as sedation are common. When administration of clonidine is stopped abruptly, a withdrawal syndrome characterized by rebound hyperactivity of the sympathetic nervous system can produce a state resembling the crises of patients with pheochromocytoma. *Opiates* also may exert a central sympatholytic effect; the sympathetic excitation of morphine withdrawal responds to clonidine and vice versa. *Propranolol* and *reserpine* may exert some sympatholytic effects at the level of the CNS.

Ganglionic Blocking Agents Ganglionic transmission may be antagonized by drugs that block the (nicotinic) cholinergic synapse between the pre- and postganglionic autonomic nerves. These agents inhibit the parasympathetic as well as the sympathetic nervous system. Only *trimethaphan* is in general clinical use; its major application is in the treatment of hypertensive crises, particularly aortic dissection, when controlled hypotension and decreased myocardial contractility are desirable (Chap. 246).

Agents Acting at the Peripheral Sympathetic Nerve Endings Adrenergic neuron-blocking agents depress the function of the peripheral sympathetic nerves by decreasing the amount of neurotransmitter released. *Guanethidine*, the prototype of this class of drugs, is concentrated in the sympathetic nerve endings by the amine-uptake mechanism. Within the terminal it blocks the release of NE in response to nerve impulses and eventually depletes the nerve of NE by displacing it from the intraneuronal storage granules. The drug was formerly useful in the management of severe hypertension, although orthostatic hypotension was a limiting side effect. *Bretylium*, an agent whose effects are similar to those of guanethidine, is employed in the treatment of ventricular fibrillation (Chap. 230). Both guanethidine and bretylium are antagonized by agents that affect the amine-uptake transport process such as sympathomimetic amines, tricyclic antidepressants, phenoxybenzamine, and phenothiazines. The antihypertensive action of guanethidine may be rapidly reversed by these drugs.

Reserpine depletes catecholamines from the peripheral sympathetic nerve endings, the brain, and the adrenal medulla. Its antihypertensive effect in humans is usually attributed to depletion of peripheralNEstores within sympathetic nerve endings. The sedation and occasionally morbid depression attending its use result from NE depletion within the CNS.

Adrenergic-Receptor Blocking Agents Adrenergic blocking agents antagonize the effects of catecholamines at the level of the peripheral tissue.

a-Adrenergic-receptor blocking agents *Phenoxybenzamine* and *phentolamine* are utilized principally in treating pheochromocytoma(Chap. 332). Phenoxybenzamine produces prolonged, noncompetitive alpha blockade, while phentolamine leads to reversible, competitive blockade. Because of its rapid action and short duration, phentolamine is commonly used in the treatment of acute hypertensive paroxysms secondary to catecholamine excess, such as occur with pheochromocytoma. Both phentolamine and phenoxybenzamine antagonizea$_1$ anda$_2$receptors, although phenoxybenzamine is more potent at thea$_1$-receptor site. *Prazosin*, an a-adrenergic blocking agent with selectivity for thea$_1$receptor, possesses properties that resemble those of primary vasodilators and has been used in the treatment of essential hypertension, as an afterload-reducing agent in congestive heart failure, and as an adjunct in the treatment of pheochromocytoma (Chap. 332). *Doxazosin* and *terazosin*, long-acting selectivea$_1$blockers, are more useful in the treatment of essential hypertension because of better dosing schedule and less orthostatic hypotension. These agents also lower triglyceride levels and raise high-density lipoprotein (HDL) cholesterol levels. These selectivea$_1$blockers, along with *tamsulosin* are useful in the symptomatic treatment of urinary outflow track obstruction and prostatism because they antagonize contraction of the sphincter at the bladder trigone and the prostate smooth muscle.

b-Adrenergic-receptor blocking agents These drugs antagonize the effects of catecholamines on the heart and are useful in the treatment of angina pectoris, hypertension, and cardiac arrhythmias. The benefit of beta blockade in angina derives from the decrease in myocardial oxygen consumption following reduction in heart rate and myocardial contractility (Chap. 244). The hypotensive effect of beta blockade is not clearly understood (Chap. 246). Diminished cardiac output, decreasedNErelease at postganglionic sympathetic nerve endings, reduced renin secretion, and suppressed central sympathetic outflow are possible mechanisms. The efficacy of b-blocking agents in the treatment of arrhythmias depends on reduction of the rate of spontaneous depolarization of pacemaker cells in the sinus node and junctional pacemakers and on slowing conduction within the atria and atrioventricular node. Beta blockade is also effective in the symptomatic management of hyperthyroidism and the control of tachycardia and arrhythmias in patients with pheochromocytoma.b-adrenergic blocking agents are also useful in the treatment of migraine, essential tremor, idiopathic hypertrophic subaortic stenosis, and aortic dissection. Several trials have demonstrated thatb-receptor blocking agents, administered long-term, diminish mortality following acute myocardial infarction. The mechanism of this cardioprotective effect may involve antiarrhythmic action, prevention of reinfarction, and reduction in infarct size (Chap. 243).

Pharmacologic Properties of b-Receptor Blocking Agents Fourteen beta-blocking agents (atenolol, acebutolol, betaxolol, bisoprolol, carvedilol, carteolol, esmolol, metoprolol, nadolol, pindolol, penbutolol, propranolol, sotalol, and timolol) are available for use in the United States. Other agents (alprenolol, bevantolol, dilevalol, oxprenolol, etc.) are in use in other countries and investigational within the United States. The utility of these agents is derived predominantly from blockade ofb-adrenergic receptors.

Although much has been written about other pharmacologic properties, including cardioselectivity, membrane-stabilizing (local anesthetic) effects, intrinsic sympathomimetic (partial-agonist) activity, and lipid solubility, the clinical significance of these additional properties is small. Local anesthetic properties are most prominent with propranolol; however, membrane stabilization probably does not contribute substantially to the clinical utility. The various beta blockers do differ in their water and lipid solubility. The lipophilic agents (propranolol, metoprolol, oxprenolol, bisoprolol, carvedilol) are readily absorbed from the gastrointestinal tract, metabolized by the liver, have large volumes of distribution, and penetrate theCNSwell; the hydrophilic agents (acebutolol, atenolol, betaxolol, carteolol, nadolol, sotalol) are less readily absorbed, not extensively metabolized, and have relatively long plasma half-lives. As a consequence, the hydrophilic agents may be administered once per day. Hepatic failure may prolong the plasma half-life of the lipophilic agents, whereas renal failure may prolong the action of the hydrophilic group. The degree of lipid solubility, therefore, provides a basis for choice of a particular agent in patients with hepatic or renal insufficiency. Although the hydrophilic agents penetrate the CNS less well, CNS side effects (sedation, depression, hallucinations) are well described with the hydrophilic as well as with the lipophilic agents.

Someb-adrenergic blocking agents possessb-agonist activity. This has been referred to as "intrinsic sympathomimetic activity" (ISA). Agents with partial agonist activity (pindolol, alprenolol, acebutolol, carteolol, dilevalol, oxprenolol) cause little or no depression of resting heart rate (partial agonist effect) while blocking the increase in heart rate that occurs in response to exercise or the administration of a beta agonist such as isoproterenol. The presence of partial agonist activity may be useful when bradycardia limits treatment in patients with slow resting heart rates. Pindolol also produces mild vasodilation, perhaps in part related to peripheral$b_2$stimulation. Agents with partial agonist activity cause less change in blood lipid levels than agents without agonist properties. On theoretical grounds, intrinsic sympathomimetic activity would be undesirable in the treatment of thyrotoxicosis, idiopathic hypertrophic subaortic stenosis, aortic dissection, and tachyarrhythmias.

Cardioselective (b_1) Adrenergic-Receptor Blocking Agents Propranolol, the prototype of the nonselective b-adrenergic blocking agent, induces a competitive blockade of bothb_1 and$b_2$receptors. Other nonselective beta-blocking agents include alprenolol, carteolol, dilevalol, nadolol, oxprenolol, penbutolol, pindolol, sotalol, timolol, and carvedilol. Metoprolol, esmolol, acebutolol, atenolol, and betaxolol possess relative selectivity for the$b_1$receptor. Althoughb_1-(cardio-) selective agents have the theoretical advantage of producing less bronchoconstriction and less peripheral vasoconstriction, a clear-cut clinical advantage of the cardioselective agents has not been decisively demonstrated, since the$b_1$selectivity is only relative. Bronchoconstriction may occur whenb_1-selective agents are administered in full therapeutic doses.

Adverse Effects ofb-Receptor Blocking Agents Aside from the effects on theCNS, most adverse reactions to beta-blocking agents are consequences ofb-adrenergic blockade. These include the precipitation of heart failure in patients in whom cardiac compensation depends on enhanced sympathetic drive; the aggravation of bronchospasm in patients with asthma; predisposition to the development of hypoglycemia in insulin-requiring diabetics (blockade of catecholamine-mediated counterregulation and antagonism of the adrenergic warning signs of hypoglycemia); the development of hyperkalemia in diabetic or uremic patients with impaired potassium tolerance; the enhancement of coronary or peripheral arterial vasospasm; and elevation in triglycerides and depression ofHDLlevels. The lipid (and perhaps the peripheral vascular) effects are less (or absent) in agents with partial (b_2) agonist activity or alpha-blocking properties (carvedilol).

Miscellaneous Adrenergic Blocking Agents Labetalol, approved for use in the United States as an antihypertensive agent, is a competitive antagonist of botha- and b-adrenergic receptors. Although labetalol induces relatively moreb- than a-receptor blockade, fall in peripheral resistance may be marked following acute administration of the drug. Vasodilation may be mediated in part by a partial agonist effect on theb_2-adrenergic receptor; labetalol does not possess partial agonist activity for theb_1(cardiac) receptor.

Metoclopramide is a dopaminergic antagonist with cholinergic agonist properties. It enhances gastric emptying, increases the tone of the lower esophageal sphincter, increases prolactin and aldosterone secretion, and antagonizes emesis induced by apomorphine. It is useful clinically in enhancing gastric emptying (in the absence of organic obstruction such as in diabetic gastroparesis), in antagonizing gastroesophageal reflux, and as an antiemetic during cancer chemotherapy.

THE PARASYMPATHETIC NERVOUS SYSTEM

ACETYLCHOLINE

AChserves as the neurotransmitter at all autonomic ganglia, at the postganglionic parasympathetic nerve endings, at the postganglionic sympathetic nerve endings innervating the eccrine sweat glands, and at the skeletal muscle end plate (neuromuscular junction). The enzyme choline acetyltransferase catalyzes the synthesis of ACh from acetyl coenzyme A (CoA) produced within the nerve ending and from choline, actively taken up from the extracellular fluid. Within the cholinergic nerve endings, ACh is stored in discrete synaptic vesicles and released in response to nerve impulses that depolarize the nerve terminals and increase calcium influx.

Cholinergic Receptors Different receptors forAChexist on the postganglionic neurons within the autonomic ganglia and at the postjunctional autonomic effector sites. Those within the autonomic ganglia and adrenal medulla are stimulated predominantly by nicotine (*nicotinic receptors*) and those on autonomic effector cells by the alkaloid muscarine (*muscarinic receptors*). Ganglionic blocking agents antagonize the nicotinic receptors, while atropine blocks the muscarinic receptors. The muscarinic (M) receptor, furthermore, has been recently subdivided into additional types. The M_1receptor is localized to theCNS and perhaps parasympathetic ganglia; the M_2receptor is the nonneuronal muscarinic receptor on smooth muscle, cardiac muscle, and glandular epithelium. Bethanechol is a selective agonist of the M_2receptor; pirenzepine, an investigational agent, is a selective antagonist of the M_1receptor that markedly reduces gastric acid secretion. The M_2receptor inhibits adenylyl cyclase and utilizes the regulatory $G_i$protein; the M_1receptor interacts with G_q and stimulates phospholipase C (Fig. 72-5). The M_3receptor, present on smooth muscle and secretory glands, is antagonized by atropine and utilizes phospholipase C,IP_3, and DAGas second messengers. Other subtypes have been identified by molecular biologic techniques but have not yet been fully characterized.

Acetylcholinesterase Hydrolysis ofACh by acetylcholinesterase inactivates the neurotransmitter at cholinergic synapses. This enzyme (also known as specific or true cholinesterase) is present within neurons and is distinct from butyrocholinesterase (serum cholinesterase or pseudocholinesterase). The latter enzyme is present in plasma and nonneuronal tissues and is not primarily involved in the termination of the effects of ACh at autonomic effector sites. The pharmacologic effects of anticholinesterase agents are due to inhibition of neuronal (true) acetylcholinesterase.

PHYSIOLOGY OF THE PARASYMPATHETIC NERVOUS SYSTEM

The parasympathetic nervous system participates in the regulation of the cardiovascular system, the gastrointestinal tract, and the genitourinary system. Tissues such as liver, kidney, pancreas, and thyroid also receive parasympathetic innervation, suggesting a role for the parasympathetic nervous system in metabolic regulation as well, although cholinergic effects on metabolism are not well characterized.

Cardiovascular System Parasympathetic effects on the heart are mediated by the vagus nerve.AChreduces the rate of spontaneous depolarization of the sinoatrial node and decreases heart rate. ACh also delays impulse conduction within the atrial musculature while shortening the effective refractory period, a combination of factors that may initiate or perpetuate atrial arrhythmias. At the atrioventricular node, ACh reduces conduction velocity, increases the effective refractory period, and thus diminishes the ventricular response during atrial flutter or fibrillation (Chap. 230). The decrease in inotropy induced by ACh is related to a prejunctional inhibitory effect on sympathetic nerve endings as well as to a direct inhibitory effect on the atrial myocardium. The ventricular myocardium is not much affected since innervation by cholinergic fibers is minimal. A direct cholinergic contribution to the regulation of peripheral resistance appears unlikely since parasympathetic innervation of the vasculature is not extensive. The parasympathetic nervous system, however, may influence peripheral resistance indirectly by inhibitingNErelease from sympathetic nerves.

Gastrointestinal Tract Parasympathetic innervation of the gut is via the vagus nerve and the pelvic sacral nerves. The parasympathetic nervous system increases the tone of gastrointestinal smooth muscle, enhances peristaltic activity, and relaxes the gastrointestinal sphincters.AChstimulates exocrine secretion from the glandular epithelium and enhances the secretion of gastrin, secretin, and insulin.

Genitourinary and Respiratory Systems Sacral parasympathetic nerves supply the urinary bladder and genitalia.AChincreases ureteral peristalsis, contracts the urinary detrusor muscle, and relaxes the trigone and sphincter, thereby playing a critical role in the coordination of urination. The respiratory tract is innervated with parasympathetic fibers derived from the vagus nerve. ACh increases tracheobronchial secretions and stimulates bronchial constriction.

PHARMACOLOGY OF THE PARASYMPATHETIC NERVOUS SYSTEM

Cholinergic Agonists AChitself has no therapeutic role because of its widespread effects and short duration of action. Congeners of ACh are less susceptible to hydrolysis by cholinesterase and have a narrower range of physiologic effects. Bethanechol, the only systemic cholinergic agonist in general use, stimulates gastrointestinal and genitourinary smooth muscle with minimal effect on the cardiovascular system. It is used in the treatment of urinary retention in the absence of outflow tract obstruction and, less commonly, in gastrointestinal disorders such as postvagotomy gastric atony. Pilocarpine and carbachol are topical cholinergic agonists used in the treatment of glaucoma.

Acetylcholinesterase Inhibitors Cholinesterase inhibitors enhance the effects of parasympathetic stimulation by diminishing the inactivation ofACh. The therapeutic application of reversible cholinesterase inhibitors depends on the role of ACh as neurotransmitter at the skeletal muscle neuroeffector junction and within theCNS and includes the treatment of myasthenia gravis (Chap. 380), the termination of neuromuscular blockade following general anesthesia, and the reversal of intoxication by agents with a central anticholinergic action. Physostigmine, a tertiary amine, penetrates the CNS well, while related quaternary amines (neostigmine, pyridostigmine, ambenonium, and edrophonium) do not. Organophosphorous cholinesterase inhibitors produce irreversible cholinesterase blockade; these agents are used principally as insecticides and are primarily of toxicologic interest. With regard to the autonomic nervous system, cholinesterase inhibitors are of limited use in the treatment of intestinal and bladder smooth-muscle dysfunction such as occurs in paralytic ileus and atonic urinary bladder. Cholinesterase inhibitors induce a vagotonic response in the heart and may be useful in terminating attacks of paroxysmal supraventricular tachycardia (Chap. 230).

Cholinergic-Receptor Blocking Agents *Atropine* blocks muscarinic cholinergic receptors, with little effect on cholinergic transmission at the autonomic ganglia and the neuromuscular junctions. Many of theCNSactions of atropine and atropine-like drugs are attributable to blockade of central muscarinic synapses. The related alkaloid, *scopolamine*, is similar to atropine but causes drowsiness, euphoria, and amnesia, effects that make it suitable as a preanesthetic medication.

Atropine increases heart rate and enhances atrioventricular conduction, actions that may be useful in combating the bradycardia or heart block associated with heightened vagal tone. In addition, atropine reverses cholinergically mediated bronchoconstriction and diminishes respiratory tract secretions. These effects contribute to its utility as a preanesthetic medication.

Atropine also decreases gastrointestinal tract motility and secretion. Although various derivatives and congeners of atropine (such as *propantheline*, *isopropamide*, and *glycopyrrolate*) have been advocated in patients with peptic ulcer or with diarrheal syndromes, the chronic use of such agents is limited by other manifestations of parasympathetic inhibition such as dry mouth and urinary retention. The investigational selective M_1inhibitor pirenzepine inhibits gastric secretion at doses that have minimal anticholinergic effects at other sites; this agent may be useful in the treatment of peptic ulcer. Atropine and its congener *ipratropium*, when given by inhalation, cause bronchodilation and have been used experimentally in the treatment of asthma.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART FIVE -NUTRITION

73. NUTRITIONAL REQUIREMENTS AND DIETARY ASSESSMENT - *Johanna Dwyer*

Nutrients are substances that are not synthesized in the body in sufficient amounts and therefore must be supplied by the diet. Nutrient requirements for groups of healthy persons have been thoroughly defined on the basis of experimental evidence. For good health we require energy-providing nutrients (protein, fat, and carbohydrate), vitamins, minerals, and water. Specific nutrient requirements include 9 essential amino acids, several fatty acids, 4 fat-soluble vitamins, 10 water-soluble vitamins, and choline. Several inorganic substances, including four minerals, seven trace minerals, three electrolytes, and the ultratrace elements, must also be supplied in the diet (Chap. 75).

The required amounts of the essential nutrients differ by age and physiologic state. Conditionally essential nutrients are not required in the diet but must be supplied to individuals who do not synthesize them in adequate amounts, such as those with genetic defects, those having pathologic states with nutritional implications, and developmentally immature infants (Chap. 74). Many organic phytochemicals and zoochemicals present in foods have various health effects. For example, dietary fiber has been shown to have beneficial effects on gastrointestinal function.

ESSENTIAL NUTRIENT REQUIREMENTS

Energy For weight to remain stable, energy intake must match energy output (Chap. 77). The major categories of energy output are resting energy expenditure (REE) and physical activity; minor sources include the energy cost of metabolizing food (thermic effect of food or specific dynamic action) and shivering thermogenesis (e.g., cold-induced thermogenesis). The average energy intake is about 2800 kcal/d for American men and about 1800 kcal/d for American women, though these estimates vary with body size and activity level. Formulas for estimating REE are useful for assessing the energy needs of an individual whose weight is stable. Thus, for males, REE= 900 + 10*w*, and for females, REE = 700 +7*w*, where *w* is weight in kg. The calculated REE is then adjusted for physical activity level by multiplying by 1.2 for sedentary, 1.4 for moderately active, or 1.8 for very active individuals. The final figure provides an estimate of total caloric needs in a state of energy balance.

Illness often alters energy needs. Unstressed hospitalized patients at bed rest usually require 1.2 times theirREE, whereas those who are stressed, febrile, and catabolic require 1.5 to 2 times their REE (Chap. 74). Intestinal malabsorption may decrease net utilizable energy to as little as 25% of ingested energy and may necessitate feeding by parenteral routes (Chap. 76). Fever increases energy expenditure by 10 to 13% per degree Celsius above normal. Other diseases increase energy needs by varying amounts, such as burns (40 to 100%), trauma (40 to 100%), and hyperthyroidism (10 to 100%). Hypothyroidism and adrenal insufficiency decrease resting energy needs, but these alterations are corrected after adequate hormone replacement. In obese patients, weight reduction can be accomplished by reducing energy intakes by approximately 500 kcal/d to achieve a loss of 0.5 kg of fat per week, or 1000 kcal/d to lose 1 kg per week (Chap. 77).

Protein Dietary protein consists of both essential and nonessential amino acids that are required for protein synthesis, whereas certain amino acids can also be used for energy and gluconeogenesis (Chap. 334). The nine essential amino acids are histidine, isoleucine, leucine, lysine, methionine/cystine, phenylalanine/tyrosine, threonine, tryptophan, and valine. When energy intake is inadequate, protein intake must be increased, since ingested amino acids are diverted into pathways of glucose synthesis and oxidation. In extreme energy deprivation, protein-calorie malnutrition may ensue (Chaps. 74 and76).

For adults, the recommended dietary allowance (RDA) for protein is about 0.6 g/kg desirable body weight per day, assuming that energy needs are met and that the protein is of relatively high biologic value. Current recommendations for a healthy diet call for at least 10 to 14% of calories from protein. Biologic value tends to be highest for animal proteins, followed by proteins from legumes (beans), cereals (rice, wheat, corn), and roots. Combinations of plant proteins that complement one another in biologic value or combinations of animal and plant proteins can increase biologic value and lower total protein requirements.

Protein needs increase during growth, pregnancy, lactation, and rehabilitation during treatment of malnutrition. The tolerance of dietary protein is decreased in renal insufficiency and liver failure. Normal protein intake can precipitate encephalopathy in patients with cirrhosis of the liver (Chap. 299) or worsen uremia in those with renal failure (Chap. 270).

Fat and Carbohydrate Fats are a concentrated source of energy and constitute on average 34% of calories in U.S. diets. However, for optimal health, fat intake should total no more than 30% of calories. Saturated fat and trans-fat should be limited to <10% of calories, and polyunsaturated fats to <10% of calories, with monounsaturated fats comprising the remainder of fat intake. At least 55% of total calories should be derived from carbohydrates. The brain requires about 100 g/d of glucose for fuel; other tissues use about 50 g/d. Over time, adaptations in carbohydrate needs are possible in hypocaloric states. For example, reduced insulin levels lead to adipose tissue breakdown and cause the body to burn more fatty acids. However, some tissues (e.g., brain and red blood cells) rely on glucose supplied either exogenously or from muscle proteolysis (Chap. 334).

Water For adults, 1 to 1.5 mL water per kcal of energy expenditure is sufficient under usual conditions to allow for normal variations in physical activity levels, sweating, and solute load of the diet. Water losses include 50 to 100 mL/d in the feces, 500 to 1000 mL/d by evaporation or exhalation, and, depending on the renal solute load, ³1000 mL/d in the urine. If external losses increase, intakes must increase accordingly to avoid underhydration. Fever increases water losses by approximately 200 mL/d per°C; diarrheal losses vary but may be as great as 5 L/d in severe diarrhea. Heavy sweating and vomiting also increase water losses. When renal function is normal and solute intakes are adequate, the kidneys can adjust to increased water intake by excreting up to 18 L/d of excess water (Chap. 329). However, obligatory urine outputs can compromise hydration status when there is inadequate intake or when losses increase in disease or kidney damage.

Infants have high requirements for water because of their large ratio of surface area to volume, the limited capacity of the immature kidney to handle high renal solute loads, and their inability to communicate their thirst. Increased water needs during pregnancy are low, perhaps an additional 30 mL/d; but during lactation, milk production increases water requirements so that approximately 1000 mL/d of additional water is needed, or 1 mL for each mL of milk produced. Special attention must be paid to the water needs of the elderly, who have reduced total body water and blunted thirst sensation and may be taking diuretics.

Other Nutrients The vitamins and minerals required for health and the clinical disorders caused by vitamin deficiency or excess are discussed inChap. 75.

DIETARY REFERENCE INTAKES, RECOMMENDED ALLOWANCES, AND TOLERANCES

Fortunately, human life and well-being can be maintained within a fairly wide range for most nutrients. However, the capacity for adaptation is not infinite -- too much, as well as too little, intake of a nutrient may have adverse effects or alter the health benefits conferred by another nutrient (Chap. 75). Therefore, benchmark recommendations on nutrient intakes have been developed to guide clinical practice. These quantitative estimates of nutrient intakes are collectively referred to as the *dietary reference intakes* (DRIs). The DRIs supplant theRDAs, the single reference values used in the United States since 1989. DRIs include the estimated average requirement (EAR) for nutrients, as well as three other reference values used for dietary planning for individuals: the RDAs, the adequate intake (AI), and the safe upper level (UL). The current RDAs and AIs are provided in Tables 73-1 and73-2, respectively.

Estimated Average Requirement When florid dietary deficiency diseases such as rickets, scurvy, xerophthalmia, and protein-calorie malnutrition were common, nutrient adequacy was assumed by the absence of clinical signs of a dietary deficiency disease. Later, it was determined that biochemical and other changes were evident long before the clinical deficiency became apparent. Consequently, criteria of adequacy are chosen using such biologic markers when they are available. Current efforts focus on the amount of a nutrient that reduces the risk of chronic degenerative diseases. Priority is given to sensitive biochemical, physiologic, or behavioral tests that reflect early changes in regulatory processes or maintenance of body stores of nutrients.

TheEAR is the amount of a nutrient estimated to be adequate for half of the healthy individuals of a specific age and sex. The types of evidence and criteria used to establish nutrient requirements vary by nutrient, age, and physiologic group. The EAR is not an effective estimate of nutrient adequacy in individuals because it is a median requirement for a group, and the variation around this number is considerable. As the EAR specifies, 50% of individuals in a group fall below the requirement and 50% fall above it. Thus, a person with a usual intake at the EAR has a 50% risk of an inadequate intake during the reporting period. For these reasons, other standards, described below, are more useful for clinical purposes.

Recommended Dietary Allowances The RDAis the average daily dietary intake level that meets the nutrient requirements of nearly all healthy persons of a specific sex, age, life stage, or physiologic condition (such as pregnancy or lactation). The RDA is commonly used as a nutrient-intake goal for planning diets of individuals.

TheRDA is defined statistically as 2 standard deviations (SD) above theEAR to ensure that the needs of any given individual are met. The RDAs are used to formulate food guides such as the U.S. Department of Agriculture (USDA) Food Guide Pyramid for individuals, food exchange lists for therapeutic diet planning, and as a standard for describing the nutritional content of processed foods and nutrient supplements. The nutrient content in a food is stated by weight or as a percent of the daily value (DV), a varient of the RDA which, for an adult, represents the highest RDA for an adult consuming 2000 kcal/d.

The risk of dietary inadequacy increases as intakes fall further below theRDA. However, the RDA is an overly generous criterion for evaluating nutrient adequacy. For example, by definition the RDA exceeds the actual requirements of all but about 2 to 3% of the population. Therefore, many people whose intakes fall below the RDA may still be getting enough of the nutrient.

Adequate Intake It is not possible to set anRDA for some nutrients that do not have an establishedEAR. In this circumstance, theAI is based on observed, or experimentally determined, approximations of nutrient intakes in healthy people. In theDRIsestablished to date, AIs rather than RDAs are proposed for infants up to age 1, as well as for calcium, vitamin D, fluoride, pantothenic acid, biotin, and choline for persons of all ages.

Tolerable Upper Levels of Nutrient Intake Excessive nutrient intake can disturb body functions and cause acute, progressive, or permanent disabilities (Chap. 75). Some diseases of nutritional excess include fluorosis, hypervitaminosis A, hypervitaminosis D, and obesity. The tolerableUL is the highest level of chronic nutrient intake (usually daily) that is unlikely to pose a risk of adverse health effects for most of the population. An uncertainty factor is applied to ensure that even very sensitive persons would not experience adverse effects at the UL dose chosen. For many nutrients, data on the adverse effects of large amounts of the nutrient are unavailable or too limited establish a UL. Therefore, the lack of a UL does *not* mean that the risk of adverse effects from high intakes is nonexistent; caution is warranted in those who consume large amounts of such nutrients. Healthy individuals derive no established benefit from consuming nutrient levels above the RDAor AI. Individual nutrients in foods that most people eat rarely reach levels that exceed the UL. However, nutritional supplements provide more concentrated amounts of nutrients per dose and, as a result, pose a potential risk of toxicity. Nutrient supplements are labeled with "supplement facts" that express the amount of nutrient in absolute units or as the percent of theDVprovided per recommended serving size. Those who use supplements should be advised that total nutrient consumption, including both food and supplements, should not exceed RDA levels.

FACTORS ALTERING NUTRIENT NEEDS

TheDRIs are affected by age, sex, rate of growth, pregnancy, lactation, physical activity, composition of diet, concomitant diseases, and drugs. When only slight differences exist between the requirements for nutrient sufficiency and excess, dietary planning becomes more difficult. Renal insufficiency provides one example in which protein intakes must be sufficient to maintain protein nutritional status, while avoiding exacerbation of uremic symptoms because of protein excess.

Physiologic Factors Growth, strenuous physical activity, pregnancy, and lactation increase needs for energy and several essential nutrients. Energy needs rise during pregnancy, due to the demands of fetal growth, and during lactation, because of the increased energy required for milk production. Energy needs decrease with loss of lean body mass, the major determinant ofREE. Because both health and physical activity tend to decline with age, energy needs in older persons, especially those over 70, tend to be less than those of younger persons.

Dietary Composition Dietary composition affects the biologic availability and utilization of nutrients. For example, the absorption of iron may be impaired by high amounts of calcium or lead; non-heme iron uptake may be impaired by the lack of ascorbic acid and amino acids in the meal. The absorption of calcium and magnesium is decreased by large amounts of phytates in the diet. Protein utilization by the body may be decreased when essential amino acids are not present in sufficient amounts. Animal foods, such as milk, eggs, and meat, have high biologic values with most of the needed amino acids present in adequate amounts. Plant proteins in corn (maize), soy, and wheat have lower biologic values and must be combined with other plant or animal proteins to achieve optimal utilization by the body.

Route of Administration TheRDAsapply only to oral intakes. When nutrients are administered parenterally, similar values can sometimes be used for amino acids, carbohydrates, fats, sodium, chloride, potassium, and most of the vitamins, since their intestinal absorption is nearly 100% (Chap. 75). However, the oral bioavailability of most mineral elements may be only half that obtained by parenteral administration. For some nutrients that are not readily stored in the body, or cannot be stored in large amounts, timing of administration may also be important. For example, amino acids cannot be used for protein synthesis if they are not supplied together; instead they will be used for energy production.

Disease Specific dietary deficiency diseases include protein-calorie malnutrition; iron, iodine, and vitamin A deficiency; megaloblastic anemia due to vitamin B_{12} or folic acid deficiency; vitamin D deficiency rickets; and scurvy, beriberi, and pellagra (Chaps. 74 and75). Each deficiency disease is characterized by imbalances at the cellular level between the supply of nutrients or energy and the body's nutritional needs for growth, maintenance, and other functions. Imbalances in nutrient intakes are recognized as risk factors for certain chronic degenerative diseases, such as saturated fat and cholesterol in coronary artery disease; sodium in hypertension; obesity in hormone-dependent endometrial, breast, and prostate cancers; and ethanol in alcoholism. Since the etiology and pathogenesis of these disorders are multifactorial, diet is only one of many risk factors. Osteoporosis, for example, is associated with calcium deficiency, as well as risk factors related to environment (e.g., smoking, sedentary lifestyle), physiology (e.g., estrogen deficiency), genetic determinants (e.g., defects in collagen metabolism), and drug use (chronic steroids) (Chap. 342).

DIETARY ASSESSMENT

In clinical situations, nutritional assessment is an iterative process that involves: (1) screening for malnutrition, (2) assessing the diet and other data to establish either the absence or presence of malnutrition and its possible causes, and (3) planning for the most appropriate nutritional therapy. Some disease states affect the bioavailability, requirements, utilization, or excretion of specific nutrients. In these circumstances, specific measurements of various nutrients may be required to assure adequate replacement (Chap. 75).

Most health care facilities have a nutrition screening process in place for identifying possible malnutrition after hospital admission. Nutritional screening is required by the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), but there are no universally recognized or validated standards, so techniques vary. The factors that are usually assessed include: abnormal weight for height or body mass index (e.g., BMI <19 or>25); reported weight change (involuntary loss or gain of >5 kg in past 6 months) (Chap. 43); diagnoses with known nutritional implications (metabolic disease, any disease affecting the gastrointestinal tract, alcoholism, and others); present therapeutic dietary prescription; chronic poor appetite; presence of chewing and swallowing problems or major food intolerances; need for assistance with preparing or shopping for food, eating, or other aspects of self care; and social isolation. Reassessment of nutrition status should occur periodically in hospitalized patients -- at least once every week.

A more complete dietary assessment is indicated for patients who exhibit a high risk of malnutrition on nutrition screening. The type of assessment varies based on the clinical setting, severity of the patient's illness, and stability of his or her condition.

Acute Care Settings In acute care settings, anorexia, various diseases, test procedures, and medications can compromise dietary intake. Under such circumstances, the goal is to identify and avoid inadequate intake and assure appropriate alimentation. Dietary assessment in acute care situations focuses on what patients are currently eating, whether they are able and willing to eat, and whether they experience any problems with eating. Dietary intake assessment is based on information from observed intakes; medical record; history; clinical examination; and anthropometric, biochemical, and functional status. The objective is to gather enough information to establish the likelihood of malnutrition due to poor dietary intake or other causes in order to determine whether nutritional therapy is indicated.

Simple observations may suffice to suggest inadequate oral intake. These include dietitians' and nurses' notes, the amount of food eaten on trays, frequent tests and procedures that are likely to cause meals to be skipped, nutritionally inadequate diet orders such as clear liquids or full liquids for more than a few days, fever, gastrointestinal distress, vomiting, diarrhea, or a comatose state. Patients with diseases or treatments that involve any part of the alimentary tract are at high nutritional risk. Acutely ill patients with diet-related diseases such as diabetes need assessment because an inappropriate diet may exacerbate these conditions and adversely affect other therapies. Abnormal biochemical values [serum albumin levels <35 g/L (<3.5 mg/dL); serum cholesterol levels <3.9 mmol/L (<150 mg/dL)] are nonspecific but may also indicate a need for further nutritional assessment.

Most therapeutic diets offered in hospitals are calculated to meet individual nutrient requirements and theRDA. Exceptions include clear liquids, some full liquid diets, and test diets, which are inadequate for several nutrients and should not be used, if possible, for more than 24 h. As much as half of the food served to hospitalized patients is not eaten, and so it cannot be assumed that the intakes of hospitalized patients are adequate. The dietary assessment should therefore compare how much and what food the patient has consumed with the diet that has been provided in the hospital. Major deviations in intakes of energy, protein, fluids, or other nutrients of special concern for the patient's illness should be noted and corrected.

Nutritional monitoring is especially important for patients who are very ill and who have extended lengths of stay. Patients who are fed by special enteral and parenteral routes also require special nutritional assessment and monitoring by physicians with training in nutrition support and/or dietitians with certification in nutrition support (Chap. 76).

Ambulatory Settings The aim of dietary assessment in the outpatient setting is to determine whether the patient's usual diet is a health risk in itself or if it contributes to existing chronic disease-related problems. It also provides the basis for planning a diet that fulfills therapeutic goals while ensuring patient compliance. The outpatient dietary assessment should review the adequacy of present and usual food intakes, including vitamin and mineral supplements, medications, and alcohol, as all of these may affect the patient's nutritional status. The dietary assessment should focus on the dietary constituents that are most likely to be involved or compromised by a specific diagnosis, as well as any comorbidities that are present. More than one day's intake should be reviewed to provide a better representation of the usual diet.

There are many ways to assess the adequacy of the patient's habitual diet. These include a food guide, a food exchange list, a diet history, or a food frequency questionnaire. A commonly used food guide for healthy persons is theUSDA's food pyramid, which is useful as a basis for identifying inadequate intakes of essential nutrients, as well as likely excesses in fat, saturated fat, sodium, sugar, and alcohol (Table 73-3). The guide is calculated to provide approximately 1600 kcal for sedentary women and some older adults; 2200 kcal for most children, teenage girls, active women, and many sedentary men (women who are pregnant or breastfeeding may need somewhat more); and 2800 kcal for teenage boys, most active men, and some very active women. Results provide a rough guide to food groups that may be eaten either in excess of recommendations or in insufficient quantities. Respondents who follow ethnic or unusual dietary patterns may need extra instruction on how foods should be categorized, as well as the appropriate portion sizes that constitute a serving. The process of reviewing the guide with patients helps them transition to healthier dietary patterns. For those on therapeutic diets, assessment against food exchange lists may be useful. These include, for example, the American Diabetes Association food exchange lists for diabetes, or the American Dietetic Association food exchange lists for renal disease.

Nutritional Status Assessment Full nutritional status assessment is a complex, time-consuming, and expensive process that requires considerable expertise.

Candidates include seriously ill patients and those at very high nutritional risk when the cause of malnutrition is still uncertain after initial clinical evaluation and dietary assessment. Full nutritional status assessment involves multiple dimensions, including documentation of dietary intake, anthropometric measurements, biochemical measurements of blood and urine, clinical examination, health history, and functional.**For further discussion of Nutritional Assessment, see Chap. 74.*

(Bibliography omitted in Palm version)

Back to Table of Contents

74. MALNUTRITION AND NUTRITIONAL ASSESSMENT - *Charles H. Halsted*

Malnutrition is a frequent and integral component of acute and chronic illness. When recognized by appropriate clinical assessment, malnutrition is found in>50% of all hospitalized adults. It contributes to increased in-hospital morbidity and mortality in both medical and surgical patients, and leads to more frequent hospital admissions among the elderly. Malnutrition results from various combinations of starvation, including inadequate intake or abnormal gastrointestinal assimilation of the diet, the stress response to acute injury or chronic inflammation, and abnormal nutrient metabolism. Nutritional assessment should be considered an integral part of the clinical evaluation and be used as a basis for nutritional support in the overall therapeutic plan.

DEFINITIONS OF MALNUTRITION

In the strict sense, the term *malnutrition* includes extremes of underweight and overweight. The current chapter, however, focuses on the evaluation of the undernourished patient who presents with diminished body protein and energy stores and micronutrient deficiencies.

To the practicing physician, both outpatients and inpatients should be considered at risk for malnutrition if they meet one or more of the following criteria: (1) unintentional loss of>10% of usual body weight in the preceding 3 months, (2) body weight <90% of ideal for height, or (3) body mass index (BMI; the weight in kilograms divided by the height in square meters) <18.5. With regard to varying levels of severity, body weight <90% of ideal for height represents risk of malnutrition, body weight <85% of ideal constitutes *malnutrition,* <70% of ideal represents *severe malnutrition,* and<60% of ideal is usually incompatible with survival.

Malnutrition may be endemic in regions of famine, and two forms of severe malnutrition are recognized under conditions of inadequate food supply or distribution: *marasmus* refers to generalized starvation with loss of body fat and protein, whereas *kwashiorkor* refers to selective protein malnutrition with edema and fatty liver. The latter form occurs following restriction of dietary protein among children in settings of recurrent diarrheal illness. These distinctions, however, seldom apply to malnourished patients in more developed societies. In this setting, features of combined protein-calorie malnutrition (PCM) are more commonly seen in the context of a wide variety of acute and chronic illnesses that lead to depletion of body fat, muscle wasting, multiple signs of micronutrient deficiencies, decubitus ulcers, and life-threatening infections. An overview of the evaluation of malnutrition in the sick adult is depicted inFig. 74-1.

PATHOPHYSIOLOGY AND ETIOLOGIES OF MALNUTRITION

In simple terms, patients lose weight when: (1) the intake or gastrointestinal assimilation of dietary calories is insufficient to meet normal energy expenditure; (2) the expenditure of body energy stores is greater than energy normally consumed and assimilated by the body; or (3) the metabolism of energy supplies, protein, and other nutrients is significantly impaired by the intrinsic disease process.

Body Composition As depicted in Fig. 74-2, the human body stores between 15 and

25% of its energy as fat (greater in women than men), which is available for the metabolism of endogenous fatty acids during starvation. The remaining fat-free mass (FFM) is composed of extracellular and intracellular water, the bony skeleton, glycogen, and skeletal and visceral protein. Aside from body fat, energy reserves are also provided by intracellular glycogen and protein, which, together with intracellular water, constitute the body cell mass (BCM). Thus, in addition to the enzymes that support the normal metabolic machinery of the body, the BCM provides reserve protein for energy production by gluconeogenesis during the stress response.

The Metabolic Response to Starvation and Stress The expenditure of body stores of energy (as fat, glycogen, and protein) is different during *starvation* (due to decreased intake and/or assimilation of the diet) and *stress* (due to excessive expenditure of energy and body protein). Consequently, these events affect body compartments differently. Starvation decreases the size of all body compartments, whereas stress reducesBCM, increases extracellular water, and has variable effects on body fat.

A normal 70-kg man stores fuel at about 15 kg as fat, 6 kg as protein, and 0.4 kg as glycogen. During a 24-h fast, energy needs are met by the consumption of liver glycogen stores and the conversion of up to 75 g of body protein to glucose (by gluconeogenesis). During prolonged starvation, metabolism is supported by stores of body fat (about 150 g/d), which provides fatty acid-derived ketones, and muscle protein (about 20 g/d), which is used for gluconeogenesis. Under these conditions, total energy expenditure is decreased in order to conserve energy. While normal-weight individuals can sustain total fasting for about 2 months, obese individuals can fast for periods>12 months, depending on their fat stores.

The metabolic responses to the stress of acute critical illness (e.g., following accidental or surgical trauma or sepsis) significantly modify this sequence of events. In contrast to the hypometabolism, protein conservation, and reliance on body fat stores for energy needs during starvation, the acute stress response is characterized by hypermetabolism, in which the demands of accelerated energy expenditure are met by skeletal and visceral proteolysis to provide amino acid substrate for gluconeogenesis. Muscle proteolysis and gluconeogenesis are promoted by high levels of circulating catecholamines, glucagon, cortisol, and cytokines, including tumor necrosis factor (TNF) a and interleukins 1 and 6, in the setting of insulin resistance. When untreated, body protein catabolism is accelerated to 240 g/d, which is sufficient to deplete 50% of body protein stores within 3 weeks.

A more common clinical situation is the malnourished patient with chronic illness in whom acute trauma or sepsis superimposes cytokine-mediated proteolysis with increased metabolic demands. If unchecked by appropriate therapy, the process of progressivePCM in such patients is associated with decreased cardiac and renal function, fluid retention, intestinal mucosal atrophy, loss of intracellular minerals (zinc, magnesium, and phosphorus), diminished cell-mediated immune functions, increased risk of infection, and eventual death (Fig. 74-3).

Etiologies of Malnutrition The causes of decreased dietary intake are diverse and include social and economic conditions, psychiatric diseases, neurodegenerative dementias, cytokine-mediated appetite suppression in chronic infections such as AIDS or in disseminated cancer, and self-limited food intake in abdominal pain syndromes (Table 74-1;Chap. 43). Given the central role of the gastrointestinal tract in the assimilation of nutrients,PCM is a predictable component of many chronic gastrointestinal diseases. These diseases promote starvation through decreased assimilation of the diet by: (1) blocking the transit of dietary constituents to the intestinal absorbing surface, (2) impairing normal processes of pancreatic or biliary digestion, or (3) preventing the intestinal mucosal transport of dietary constituents. Diseases that are characterized by increased catabolism of stored energy and protein include acute surgical or medical critical illness and acute or chronic inflammatory or infectious disorders affecting diverse organ systems. Other chronic diseases promote malnutrition through mixed mechanisms that contribute to abnormal nutrient metabolism. Both AIDS and disseminated malignancy, for example, cause progressive malnutrition through combinations of anorexia and futile cycles of fatty acid and glucose metabolism. Chronic obstructive pulmonary disease increases risk of malnutrition through the increased energy expenditure of labored respiration, chronic indolent bronchial infection, and the anorexic side effects of many bronchodilating drugs. Chronic liver disease is often associated with PCM caused by the cumulative effects of anorexia; decreased biliary circulation; and abnormal lipid, carbohydrate, and protein metabolism. The chronic intestinal inflammation of Crohn's disease or ulcerative colitis accelerates fecal losses of protein, electrolytes, and zinc.

CLINICAL EVALUATION OF THE MALNOURISHED PATIENT

THE PATIENT HISTORY

The clinical nutritional history should include diet and weight change, socioeconomic conditions, and symptoms unique to each clinical setting (Table 74-2). Social and economic conditions that may lead to poverty include inadequate income, homelessness, and activities that restrict real income and promote involuntary diet restriction, such as drug abuse or chronic alcoholism. Anorexia, or loss of appetite, is a feature of psychiatric disorders, such as anorexia nervosa and neurodegenerative dementia in the elderly. Many self-selected, inadequate diets may promote malnutrition. During binge drinking, chronic alcoholics typically substitute more than half their daily food calories with excessive amounts of ethanol, the metabolism of which consumes energy and promotes unbalanced metabolism of fat and carbohydrates. Other inadequate diets include unbalanced and commercially promoted formulas for rapid weight loss and strict vegetarianism, which may lead to selective deficiencies of specific nutrients such as vitamin B_{12} and iron.

Digestive diseases are major causes of malnutrition, both in the inpatient and outpatient settings. The malnourished patient with digestive disease may present with symptoms of: (1) dysphagia or recurrent vomiting due to benign or malignant esophageal or gastrointestinal obstruction; (2) chronic diarrhea due to abnormal pancreatic or biliary digestion, intestinal mucosal malabsorption, or protein-losing enteropathy; or (3) recurrent abdominal pain exacerbated by eating, as occurs in patients with chronic pancreatitis, inflammatory bowel disease, or intestinal ischemia.

On the general medical service,PCM is prevalent in patients with multiple chronic illnesses that are associated with anorexia, recurrent stress, and abnormal nutrient metabolism. In addition, PCM is comorbid with chronic recurrent pancreatitis, renal failure, chronic liver disease, chronic obstructive pulmonary disease, disseminated cancer, and chronic infections such as AIDS and tuberculosis. Depending on the severity of injury or illness, critically ill surgical and medical patients predictably develop stress-related PCM if increased nutritional needs are not met after 5 to 10 days.

THE PHYSICAL EXAMINATION

A careful physical examination can both characterize and define the extent of malnutrition. Measurements of unclothed weight and height are essential for establishing the severity of malnutrition in all patients but may be confounded by the effects of fluid overload as a result of edema and ascites. The normal values for weight (in kg) and height (in cm) in men and women are provided inTable 74-3. These values can be adjusted by±10% to account for variability in body frame.

Anthropometry Measurements of subcutaneous fat and skeletal muscle are important to determine the severity ofPCM. Using specialized calipers and a tape measure, anthropometry estimates body fat from the thickness of the skin-fold of the posterior mid-upper arm. Anthropometric measurements in healthy and malnourished adults are shown inTable 74-4. Mid-arm muscle circumference is estimated from the equation:

The use of anthropometry is limited by the requirement for specialized calipers, the experience of the observer, and potential confounding effects of edema or dehydration.

Specific Physical Findings of Malnutrition During the conventional physical examination, the observant and experienced clinician can identify multiple and specific findings ofPCM and its associated micronutrient deficiencies (Chap. 75). A variety of nutritional deficiencies can be identified by examination of the patient's general appearance, including skin, hair, nails, mucus membranes, and neurologic system (Table 74-5). Initially, a pinch of the posterior upper arm may reveal loss of subcutaneous fat in the malnourished patient. Hollowing of the temporal muscles, wasting of upper arms and thigh muscles, easily plucked hair, and peripheral edema are all consistent with protein deficiency. Examination of the skin may reveal the papular keratitis ("goose bump rash") of vitamin A deficiency, perifollicular hemorrhages of vitamin C deficiency, ecchymoses of vitamin K deficiency, the "flaky paint" lower extremity rash of zinc deficiency, hyperpigmentation of skin-exposed areas from niacin deficiency, seborrhea of essential fatty acid deficiency, spooning of nails in iron deficiency, and transverse nail pigmentation in protein deficiency. The eye examination yields conjunctival pallor of anemia, pericorneal and corneal opacities of severe vitamin A deficiency ("Bitot spots"), and nystagmus and isolated ocular muscle paresis of thiamine deficiency. The oral examination may reveal angular stomatitis and cheilosis of either riboflavin or niacin deficiency; glossitis with smooth and red tongue of riboflavin, niacin, vitamin B_{12}, or pyridoxine deficiency; and hypertrophied bleeding gums of vitamin C deficiency. Examination of the neurologic system, particularly in the setting of chronic alcohol abuse, may detect memory loss with confabulation, a wide-based gait, and past pointing, which, together with ophthalmoplegia and peripheral neuropathy, constitute the Wernicke-Korsakoff syndrome of thiamine deficiency. Other neurologic causes of dementia include pellagra due to niacin and/or tryptophan deficiency. Additional causes of peripheral neuropathy include deficiencies of pyridoxine or vitamin E; loss of distal vibratory and position sense is characteristic of the subacute combined degeneration of vitamin B_{12}deficiency.

LABORATORY ASSESSMENT

Selected use of laboratory tests, most of which are widely available, is essential for characterizing and quantifying malnutrition. Laboratory findings that are often attributed to chronic disease may, in actuality, reflect the response toPCM or selected micronutrient deficiencies in the setting of chronic illness.

Serum Visceral Proteins Serum albumin, which has a 2- to 3-week half-life, is a highly sensitive but nonspecific measure ofPCM. A normal serum albumin level in a well-hydrated patient is inconsistent with PCM. On the other hand, a low serum albumin level must be interpreted in its clinical context, since the concentration of albumin is decreased in the setting of increased plasma volume (as seen in acute trauma or sepsis and in chronic liver, renal, or cardiopulmonary failure). The acute stress of surgery, sepsis, or other acute inflammatory illness lowers the serum albumin level because of a combination of increased circulating extracellular volume andTNF-a-mediated inhibition of albumin synthesis. Hepatic albumin synthesis is inhibited in the setting of liver cirrhosis, AIDS, and disseminated cancer, whereas albumin loss from the body is accelerated in inflammatory bowel diseases, including ulcerative colitis, Crohn's disease, and radiation enteritis. Several shorter-lived visceral proteins can also be measured for estimation of the severity of PCM. These include transferrin (1-week half-life), prealbumin or retinol-binding protein complex (2-day half-life), and fibronectin (1-day half-life). However, like the serum albumin level, the circulating level of each of these proteins is affected by the changes in extracellular volume that occur in acute and chronic illnesses.

Vitamin and Mineral Assays Specific micronutrient deficiencies can be measured by a variety of serum and red blood cell assays, often utilizing high-performance liquid chromatography or enzyme or microbiologic assays (Chap. 75). Commonly available assays and their interpretations are listed in Table 74-6.PCM is typically associated with low serum levels of vitamin A, zinc, and magnesium. Abnormal digestion and absorption of dietary fat are associated with deficiencies of fat-soluble vitamins A, D, and E, whereas intestinal mucosal malabsorption (as in celiac disease) is commonly associated with additional deficiencies of iron and folic acid. Chronic alcoholism is frequently associated with thiamine, folate, vitamin A, and zinc deficiencies. Vitamin B_{12}deficiency due to achlorhydria occurs in up to 15% of elderly individuals as well as in those with pernicious anemia or with diseases involving the terminal ileum. As described in Chap. 105, both folate and vitamin B_{12}deficiencies are associated with elevations in plasma homocysteine; vitamin B_{12}deficiency can also elevate the plasma level of methylmalonic acid.

Assessment of Immune Function PCM is associated with atrophy of thymic-dependent lymphoid structures and reduced T cell-mediated immunity. Conversely, B cell-mediated production of immunoglobulins is usually unaffected. Total lymphocyte count (total white cell count ´ fraction as lymphocytes) is often<1000/uL in

PCM and may be accompanied by anergy to common skin test antigens. While sensitive for PCM, these measures of cell-mediated immunity are nonspecific and can be affected by other disorders such as acute or chronic infections, uremia, or immunosuppressive therapy.

SPECIALIZED PROCEDURES FOR NUTRITIONAL ASSESSMENT

Several specialized procedures are used to assess energy and protein stores and energy expenditure in malnourished patients. These procedures may be employed during the initial nutritional assessment or may serve as an index of the efficacy of nutritional support during the treatment of malnourished patients.

Bioelectric Impedance Analysis Bioelectric impedance analysis (BIA) is a simplified and portable method for measurement of body fat,FFM, and total-body water. BIA is based on differences in the electric conductivity of a weak current between electrodes placed on the dorsal surfaces of the hands and feet. The measurement reflects differences in the impedance to electric current, which is greatest through fat and least through water. Lean body mass can be calculated as the difference between fat mass and body weight or as total-body water divided by 0.73.

Overall,BIA is most useful in assessing body fat andFFM in stable patients and in those who suffer from conditions leading to relative starvation. However, BIA can also be used to assess critically ill patients with decreased intracellular water space andBCM and expanded extracellular compartment size. Reduced BCM correlates inversely with increased metabolic rate. BIA may be confounded in AIDS patients receiving protease-inhibitor therapy, if they exhibit lipodystrophy with associated redistribution of interscapular, abdominal, and breast fat (Chap. 309).

Energy Expenditure Body weight and energy balance are sustained in health by the consumption of dietary calories in an amount equal to the daily expenditure of energy. Therefore, caloric needs can be determined from the estimated daily total energy expenditure (TEE), which is composed of basal or resting energy expenditure (REE, about 75% of total), the thermic expenditures of digestion (about 10% of total), and modest physical activity (about 15% of total). The REE is directly proportional to both theFFM andBCM and can be estimated in healthy people using the Harris and Benedict formula on the basis of weight in kg (*W*), height in cm (*H*), and age in years (*A*):

A simplified bedside estimation forTEE in sick patients is 25 kcal/kg of body weight, to which is added 10% for digestion or metabolism of intravenous or enteral nutrition. In the acutely ill patient, one should include an additional 12.5% for each degree of fever over 37°C, as well as an additional multiplier commensurate with the severity of illness (e.g., 25% for general surgery, 50% for sepsis, and 100% for extensive third-degree burns).

WhileREE can be predicted by the Harris-Benedict equations in healthy persons, it is decreased in starvation because of hypometabolism. In contrast, REE is increased in the hypermetabolic stress that accompanies critical illness. REE and caloric requirements cannot be predicted in certain clinical conditions. These include the relatively starved, chronically ill patient admitted with a critical illness, the obese patient who develops a critical illness on the background of both increased body fat andFFM, or the patient with chronic liver disease accompanied by combinations of anorexia and ongoing hepatic inflammation. In these situations, REE can be measured accurately by the gas-exchange method of indirect calorimetry. In practice, indirect calorimetry is performed at the bedside using a mobile metabolic cart. This procedure is applicable to ventilator-independent and -dependent patients whose fractional intake of oxygen is less than 0.45. Because the goal is to reach an accurate approximation of the 24-h energy requirement, measurements must be taken at intervals during the day and must account for several variables, including food intake and activity. To calculate the energy cost of metabolism by indirect calorimetry, the volumes (*V*) of oxygen consumed and carbon dioxide produced are measured over a given period of time, according to the modified Weir equation where

Indirect calorimetry also provides the respiratory quotient (RQ), which is the ratio of carbon dioxide produced to oxygen consumed during the process of gas collection. The RQ decreases when fat is the predominant substrate for metabolism (as in starvation) and increases when the contribution of carbohydrate increases (as during stress with gluconeogenesis). In healthy individuals, the RQ usually falls between 0.80 and 0.90. A RQ <0.7 is consistent with active ketogenesis from endogenous fatty acid metabolism with limited generation of carbon dioxide. An RQ >1.0 indicates net lipogenesis, or the conversion of substrate carbohydrate to fat, a situation that occurs with overfeeding. Values that fall outside the range of 0.65 to 1.25 suggest an error in measurement technique.

Creatinine Excretion in the 24-h Urine Creatinine, the metabolic product of skeletal muscle creatine, is produced at a constant rate and in an amount directly proportional to skeletal muscle mass. With steady-state day-to-day renal function, each gram of creatinine in the 24-h urine collection represents 18.5 g of fat-free skeletal muscle. Since skeletal muscle is the major component ofFFM andBCM, measurement of creatinine in the 24-h urine collection can be used as a relative measure of these body compartments during the initial assessment and/or during the course of nutritional support. The *creatinine coefficient* represents the amount of creatinine excreted per kilogram of body weight; it is equal to 23 mg/kg of ideal body weight in men and 18 mg/kg of ideal body weight in women. The *creatinine-height index* represents the ratio of the measured 24-h urine creatinine excretion to the value predicted by the creatinine coefficient for the patient's ideal body weight. These values can be calculated from estimation of the patient's ideal body weight (Table 74-3) or from tables that relate creatinine excretion to height in men and women (Table 74-7). In practice, the accuracy of the 24-h urine creatinine depends primarily on completeness of the urine collection. Together with variations due to fever and fluctuations in dietary intake, inaccuracies of urine collections may result in as much as 10% error in the quantitative 24-h urine creatinine measurement. The constancy of creatinine excretion depends on steady-state renal function, and unpredictable creatinine excretion may occur through feces or skin in patients with serum creatinine levels>530 umol/L (>6 mg/dL). The presence of ascites, however, apparently does not compromise the accuracy of the 24-h urine creatinine as a reflection of FFM or BCM in patients with chronic liver disease.

Urine Nitrogen Excretion and Nitrogen Balance Nitrogen balance provides an index of protein gain or loss: 1 g nitrogen is equivalent to 6.25 g protein. Nitrogen balance can be assessed by measuring the difference between nitrogen consumed through the mouth, enteral tube, or intravenous sources and nitrogen excreted in the urine, feces, and other intestinal sources. Protein requirements to achieve zero or positive balance are less in starvation states, where daily protein losses are minimized because of hypometabolism, than in clinical states of stress, where the catabolism of skeletal muscle is accelerated for gluconeogenesis. Accurate measurement of nitrogen balance requires complete measurement of nitrogen losses from all possible excretory routes. In most cases, total urine nitrogen can be calculated by dividing 24-h urinary urea nitrogen by 0.85 and assuming approximately 2 g/d for nitrogen losses in feces and sweat. On the other hand, when the clinical condition includes extensive diarrhea and/or protein losses from pancreatic or enterocutaneous fistulas, the accuracy of nitrogen balance requires measurement of total nitrogen by the modified Kjeldahl technique in both urine and enteric sources. Total nitrogen measurements are also advisable in patients with liver failure, where urinary ammonia becomes a major and alternative source of nitrogen.

INTEGRATED BEDSIDE NUTRITIONAL ASSESSMENT

Several different approaches have been developed in order to simplify the process of nutritional assessment by using selective measurements that relate malnutrition to the specific medical condition and the severity of the underlying disease process.

Subjective Global Assessment This approach incorporates historic and physical findings as a basis for nutrition assessment by the trained physician. Major components in the history include evaluation of the extent of recent weight loss, changes in dietary intake, presence of significant gastrointestinal symptoms persisting more than 2 weeks, alterations in functional status, and the metabolic demand of the patient's underlying disease. Emphasis in the physical examination is placed on findings of depletion of subcutaneous body fat; skeletal muscle wasting; typical changes in skin, mucus membranes, and neurologic examination; as well as the presence of edema. Integration of the historic and physical data permits ranking of patients according to the following categories: adequate nutrition, moderate malnutrition, or severe malnutrition. Though the developers of the subjective global assessment have reported good sensitivity and specificity, the approach is still quite dependent on the training and experience of the clinician.

Prognostic Nutritional Assessment Several paradigms have been developed to link different parameters of nutritional assessment with clinical prognosis. Each approach links specific features of malnutrition with certain measurements of cell-mediated immunity, since abnormal immune function is a common pathway for increased risk in the malnourished patient (Fig. 74-3). A surgical prognostic nutritional index predicts morbidity based on preoperative measurements of serum albumin, transferrin, triceps skin-fold thickness, and delayed hypersensitivity to skin-test antigens.

AnotherPCMscore was developed to link survival in alcoholic liver disease to both skin-fold and mid-arm muscle measurements; the creatinine-height index; values for serum albumin, transferrin, prealbumin, and retinol-binding protein; the total lymphocyte count; and the skin-test response to a series of antigens. The Maastricht index predicts survival in patients with serious gastrointestinal diseases on the basis of factors related to serum albumin, retinol-binding protein, lymphotyce count, and deviation from the patient's ideal body weight.

(Bibliography omitted in Palm version)

Back to Table of Contents

75. VITAMIN AND TRACE MINERAL DEFICIENCY AND EXCESS- *Robert M. Russell*

Vitamins and trace minerals are required constituents of the human diet since they are either inadequately synthesized or not synthesized in the human body. Only small amounts of these substances are needed for carrying out essential biochemical reactions (e.g., acting as coenzymes or prosthetic groups). Overt vitamin or trace mineral deficiencies are rare in western countries due to a plentiful, varied, and inexpensive food supply; however, multiple nutrient deficiencies may appear together in persons who are ill or alcoholic. Moreover, subclinical vitamin and trace mineral deficiencies, as diagnosed by laboratory testing, are quite common in the normal population -- especially the geriatric population.

Body stores of vitamins and minerals vary tremendously. For example, vitamin B_{12} and vitamin A stores are large, and an adult may not become deficient for 1 or more years after being on a depleted diet. However, folate and thiamine may become depleted within weeks when eating a deficient diet. Therapeutic modalities can deplete essential nutrients from the body; for example, hemodialysis removes water-soluble vitamins, which must be replaced by supplementation.

There are several roles for vitamins and trace minerals in diseases: (1) deficiencies of vitamins and minerals may be caused by disease states such as malabsorption; (2) both deficiency and excess of vitamins and minerals can cause disease in and of themselves (e.g., vitamin A intoxication and liver disease); and (3) vitamins and minerals in high doses may be used as drugs (e.g., niacin for hypercholesterolemia). The hematologic-related vitamins and minerals (Chaps. 105,107) are not considered in this chapter, nor are the bone-related vitamins and minerals (vitamin D, calcium, phosphorus; Chap. 340), as they are covered elsewhere.

VITAMINS

THIAMINE (VITAMIN B_1)

Thiamine was the first B vitamin to be identified and is therefore also referred to as vitamin B_1. Thiamine pyrophosphate, the coenzyme form of thiamine, is required for branched-chain amino acid metabolism and carbohydrate metabolism (Fig. 75-1). Thiamine functions in the decarboxylation of a-ketoacids, such as pyruvate anda-ketoglutarate, and branched-chain amino acids and thus is a source of energy generation. In addition, thiamine pyrophosphate acts as a coenzyme for a transketolase reaction that mediates the conversion of hexose and pentose phosphates. It has also been postulated that thiamine plays a role in peripheral nerve conduction, although the exact chemical reactions underlying this function are unknown.

Absorption and Requirements At high doses, thiamine is absorbed by a passive mechanism; at low doses, it is absorbed by a carrier-mediated, active transport system and becomes phosphorylated in the process. Once absorbed, thiamine circulates bound to plasma proteins (mainly albumin) and erythrocytes. Storage sites for thiamine include muscle, heart, liver, kidney, and brain, although muscle is the principal storage site. The total body store of thiamine, mainly in the form of thiamine pyrophosphate, is approximately 30 mg, and its biologic half-life ranges between 9 and 18 days.

Given the fact that thiamine is involved in carbohydrate metabolism and energy generation, the Recommended Dietary Allowance (RDA) for males has been adjusted upward to account for increased energy utilization. Experiments have shown that heavy athletic training also increases thiamine utilization slightly. The RDA for thiamine is 1.2 mg/d for males and 1.1 mg/d for females. The median intake of thiamine in the United States from food alone is 2 mg/d. There is a 10% increase in the need for thiamine in pregnancy, and a small further increase in lactating females.

Primary food sources for thiamine include yeast, pork, legumes, beef, whole grains, and nuts. Milled and polished rice contain little, if any, thiamine. Thiamine deficiency is therefore more common in cultures that rely heavily on a rice-based diet. The molecule is heat-sensitive and is destroyed at pH > 8. Tea, coffee (caffeinated and decaffeinated), raw fish, and shellfish contain thiamineases, which can destroy the vitamin. Thus, drinking large amounts of tea or coffee can theoretically lower thiamine body stores.

Deficiency Most dietary deficiency of thiamine worldwide is the result of poor dietary intake. In western countries, the primary causes of thiamine deficiency are alcoholism and chronic illness, such as cancer. Alcohol is known to interfere directly with the absorption of thiamine and with the synthesis of thiamine pyrophosphate. Malnourished individuals with alcoholic liver disease are also at increased risk of thiamine deficiency because of diminished storage sites in liver and muscle. Thiamine should always be replenished when refeeding a patient with alcoholism, as carbohydrate repletion without adequate thiamine can precipitate acute thiamine deficiency.

Thiamine deficiency in its early stage induces anorexia, irritability, apathy, and generalized weakness. Prolonged thiamine deficiency causes beriberi, which is classically categorized as wet or dry, although there is considerable overlap. In either form of beriberi, patients may complain of pain and parathesia. *Wet beriberi* presents primarily with cardiovascular symptoms, due to impaired myocardial energy metabolism and dysautonomia, and can occur after 3 months of a thiamine-deficient diet. Patients present with an enlarged heart, tachycardia, high-output congestive heart failure, peripheral edema, and peripheral neuritis. Patients with *dry beriberi* present with a symmetric peripheral neuropathy of the motor and sensory systems with diminished reflexes. The neuropathy affects the legs most markedly, and patients have difficulty rising from a squatting position.

Alcoholic patients with chronic thiamine deficiency may also have central nervous system manifestations known as *Wernicke's encephalopathy*, consisting of horizontal nystagmus, ophthalmoplegia (due to weakness of one or more extraocular muscles), cerebellar ataxia, and mental impairment (Chap. 387). When there is an additional loss of memory and a confabulatory psychosis, the syndrome is known as *Wernicke-Korsakoff syndrome*. Although this syndrome is generally described in alcoholic patients, there may be a genetic predisposition to Wernicke-Korsakoff that involves a variant transketolase isozyme.

In severely malnourished infants 2 to 3 months old, thiamine deficiency may occur precipitously with sudden cardiovascular failure and collapse, resulting in death within hours. In addition, infants with thiamine deficiency may present with features suggesting meningitis, including vomiting, nystagmus, and convulsions. An aphonic presentation has also been described in which there is extreme irritability and either a very hoarse cry or total inability to emit any noise whatsoever (a silent scream).

The laboratory diagnosis of thiamine deficiency is usually made by a functional enzymatic assay of transketolase activity measured before and after the addition of thiamine pyrophosphate. A >25% stimulation by the addition of thiamine pyrophosphate (an activity coefficient of 1.25) is taken as abnormal. Thiamine or the phosphorylated esters of thiamine in serum or blood can also be measured by high-performance liquid chromatography (HPLC) to detect deficiency. Moreover, a urinary level of thiamine<27 ug per gram of creatinine per day is abnormal. In measuring urinary excretion of thiamine, one should make sure the patient is not taking diuretics, which increase thiamine excretion.

TREATMENT

In acute thiamine deficiency with either cardiovascular or neurologic signs, 100 mg/d of thiamine should be given parenterally for 7 days, followed by 10 mg/d orally until there is complete recovery. Cardiovascular improvement occurs in£12 h, and ophthalmoplegic improvement occurs within 24 h. Other manifestations gradually clear, although psychosis in the Wernicke-Korsakoff syndrome may be permanent or persist for several months. Consistent with this, pathologic changes occur in the cortex, cerebellum, and mammillary bodies of the thalamus. Parenteral thiamine should be given prophylactically to all chronic alcoholic patients in the emergency room, or as soon as they are admitted, to prevent precipitation of thiamine deficiency after the provision of glucose-containing solutions.

Thiamine-responsive conditions requiring pharmacologic doses of thiamine include branched-chain ketoaciduria (maple sugar urine disease), subacute necrotizing encephalopathy due to thiamine triphosphate deficiency in the brain (Leigh syndrome), thiamine-responsive lactic acidosis, and thiamine-responsive megaloblastic anemia associated with diabetes mellitus and deafness (Chap. 353). The gene for this recessive disorder, *SLC19A2*, encodes a thiamine transporter.

Toxicity Although anaphylaxis has been reported after high doses of thiamine, no adverse effects have been recorded from either food or supplements at high doses. Thiamine supplements may be bought over the counter in doses of up to 50 mg/d.

RIBOFLAVIN (VITAMIN B_2)

Riboflavin is important for the metabolism of fat, carbohydrate, and protein, reflecting its role as a respiratory coenzyme and an electron donor. Riboflavin is esterified with phosphoric acid in the body to form two coenzymes, flavin-mononucleotide (FMN) and flavin adenine dinucleotide (FAD) which are involved in a variety of cellular oxidation-reduction processes. Enzymes that contain FAD or FMN as prosthetic groups are known as *flavoenzymes* (e.g., succinic acid dehydrogenase, monoamine oxidase, glutathione reductase). Riboflavin plays an important role in niacin metabolism, since flavoenzymes act as intermediaries in the oxidation of the reduced forms of nicotinamide adenine dinucleotide (NAD) and NAD phosphate (NADP).

Riboflavin is normally absorbed by active and carrier-mediated saturable mechanisms, whereas diffusion is the principal mechanism of absorption at high concentrations. Riboflavin phosphorylation takes place mainly in the wall of the small intestine. BothFMN andFAD are bound to immunoglobulins and albumin in the circulation, and both forms are stored to some degree in liver and muscle.

Although much is known about the chemical and enzymatic reactions of riboflavin, the clinical manifestations of riboflavin deficiency are nonspecific and similar to those of other B vitamin deficiencies. Further, riboflavin deficiency usually occurs in combination with other water-soluble vitamin deficiencies (Chap. 74). Riboflavin deficiency is manifested principally by lesions of the mucocutaneous surfaces of the mouth (angular stomatitis, cheilosis, atrophic glossitis, magenta tongue, pharyngitis) and skin (seborrhea, genital dermatitis). In addition to the mucocutaneous lesions, corneal vascularization, anemia, and personality changes have been described with riboflavin deficiency.

Deficiency and Excess Riboflavin deficiency is almost always due to dietary deficiency. The requirement for riboflavin is increased during pregnancy and lactation and possibly by heavy exercise. The use of phenothiazines and antibiotics also appears to increase the need for riboflavin. Milk, other dairy products, and enriched breads and cereals are the most important dietary sources of riboflavin in the United States, although lean meat, fish, eggs, broccoli, and legumes are also good sources. Riboflavin is extremely sensitive to light, and milk should be stored in containers that protect against photodegradation. In non-milk-drinking societies (e.g., Central America), the laboratory diagnosis of riboflavin deficiency is common. Laboratory diagnosis of riboflavin deficiency can be made by measurement of red blood cell or urinary riboflavin concentrations or by measurement of erythrocyte glutathione reductase activity, with and without addedFAD. A stimulation (activity coefficient) of >1.4 is diagnostic of a deficient state. TheRDA for riboflavin is 1.1 to 1.3 mg/d in adults, with slightly higher recommendations for lactating and pregnant women. Rare genetic defects of flavoprotein synthesis may require pharmacologic doses of riboflavin for treatment. Because the capacity of the gastrointestinal tract to absorb riboflavin is limited (~20 mg if given in one oral dose), riboflavin toxicity has not been described. Thus, the most recent revision of the RDAs did not set an upper limit for this nutrient.

NIACIN (VITAMIN B$_3$)

The term *niacin* refers to nicotinic acid and nicotinamide and their biologically active derivatives. Nicotinic acid and nicotinamide serve as precursors of two coenzymes,NAD andNADP. These coenzymes are important in numerous oxidation and reduction reactions in the body. NAD and NADP serve as cofactors for dehydrogenases and are involved in the transfer of the hydride ion in many redox reactions. Thus, niacin is important in pentose, steroid, and fatty acid biosynthesis; glycolysis; protein metabolism; and the oxidation of fuels such as lactate, pyruvate, and alcohol. In addition, NAD and NADP are active in adenine diphosphate-ribose transfer reactions involved in DNA repair and calcium mobilization.

Absorption, Metabolism, and Requirements Nicotinic acid and nicotinamide are absorbed well from the stomach and small intestine. Both forms of niacin are absorbed by a sodium-dependent, facilitated diffusion mechanism at low doses, whereas passive diffusion occurs at high doses. Some storage ofNADtakes place in the liver. The amino acid tryptophan can be converted to niacin with an efficiency of 60:1 by weight. Thus, theRDA for niacin is expressed in niacin equivalents. Greater conversion of tryptophan to niacin occurs in niacin-deficient states, pregnancy, and in women using oral contraceptives. However, a lower conversion efficiency occurs if a patient is vitamin B_6- or riboflavin-deficient. The drug isoniazid inhibits the conversion of tryptophan to niacin. The urinary excretion products of niacin include nicotinic acid and niacin oxide; however, the major urinary metabolites are 2-pyridone and 2-methyl nicotinamide, measurements of which are used in diagnosis of niacin deficiency.

TheRDA for niacin is 16 niacin equivalents per day for men and 14 niacin equivalents for women. Median intakes of niacin in the United States considerably exceed these values. Diets that are corn-based can predispose to niacin deficiency due to the low tryptophan and niacin content. Niacin bioavailability is high from beans, milk, meat, and eggs; bioavailability from cereal grains is lower. Since flour is enriched with the "free" niacin (i.e., non-coenzyme form), bioavailability is excellent.

Deficiency Niacin deficiency causes *pellagra*, which is mostly found among people eating corn-based diets in parts of China, Africa, and India. Pellagra in North America is found mainly among alcoholics; in patients with congenital defects of intestinal and kidney absorption of tryptophan (Hartnup's disease;Chap. 352); and in patients with carcinoid syndrome (Chap. 93), in which there is increased conversion of tryptophan to serotonin. The early symptoms of pellagra include loss of appetite, generalized weakness and irritability, abdominal pain, and vomiting. Epithelial cell changes then ensue with stomatitis and bright-red glossitis, followed by a characteristic skin rash that is pigmented and scaling, particularly in skin areas exposed to sunlight. This rash is known as "Casal's necklace," when it rings the neck, and is seen in advanced cases. Vaginitis and esophagitis may also occur. Diarrhea (in part due to proctitis and in part due to malabsorption), depression, seizures, and dementia are also part of the pellagra syndrome -- the four D's: *d*ermatitis, *d*iarrhea, and *d*ementia leading to *d*eath.

The diagnosis of niacin deficiency is based on low levels of the urinary metabolites 2-methyl nicotinamide and 2-pyridone. Treatment of pellagra consists of oral supplementation of 100 to 200 mg of nicotinamide or nicotinic acid three times daily for 5 days. High doses of nicotinic acid (^{3}3 g nicotinic acid per day) are used for the treatment of elevated cholesterol levels and in the treatment of types 2, 4, and 5 hyperlipidemias (Chap. 344).

Toxicity Prostaglandin-mediated flushing has been observed at daily doses as low as 50 mg of niacin when taken as a supplement or as therapy for hypertriglyceridemia. No toxicity has been seen from niacin derived from food sources. Flushing may be accompanied by skin dryness, itching, and headache. Premedication with aspirin may alleviate these symptoms. Nausea, vomiting, and abdominal pain also occur at similar doses of niacin. Hepatic toxicity is the most serious toxic reaction due to niacin and may present as jaundice with elevated AST and ALT levels. A few cases of fulminant hepatitis requiring liver transplantation have been reported at doses of 3 to 9 g/d. Other toxic reactions include glucose intolerance, macular edema, and macular cysts. It is not clear whether sustained-release forms of nicotinic acid are more toxic than regular forms. The upper limit for daily niacin intake has been set at 35 mg. However, this upper limit does not pertain to the therapeutic use of niacin.

PYRIDOXINE (VITAMIN B_6)

Vitamin B_6refers to a family of compounds including pyridoxine, pyridoxal, pyridoxamine, and their 5¢-phosphate derivatives. 5¢-Pyridoxal phosphate (PLP) is a cofactor for more than 100 enzymes involved in amino acid metabolism (e.g., 5¢-PLP is a cofactor for the transulfuration enzymes involved in the conversion of homocysteine to cystathionine;Chap. 352). Vitamin B_6 is also involved in heme and neurotransmitter synthesis and in the metabolism of glycogen, lipids, steroids, sphingoid bases, and several vitamins, including the conversion of tryptophan to niacin.

Absorption and Metabolism Approximately 75% of vitamin B_6 is absorbed from a mixed diet by a nonsaturable, passive process. Much of dietary vitamin B_6 is in the phosphorylated form, and the phosphate must be removed by intestinal alkaline phosphatase before absorption takes place. Once absorbed, the vitamin becomes rephosphorylated in the liver, where the various forms can be interconverted. In the liver,PLPbinds avidly to cellular proteins and albumin. Since these binding proteins protect it from phosphatase activity, tissue levels of PLP can become quite high with continuous supplementation. Sixty mg of vitamin B_6 is stored in the body, and much is in the form of PLP bound to phosphorylase A in muscle. The biologic half-life of vitamin B_6 is 25 days.

Dietary Sources Plants contain vitamin B_6 in the form of pyridoxine, whereas animal tissues containPLP and pyridoxamine phosphate. The vitamin B_6contained in plants is less bioavailable than that from animal tissues. All forms of vitamin B_6 are labile in alkaline conditions. Rich food sources of vitamin B_6include legumes, nuts, wheat bran, and meat, although the vitamin is present in all food groups. TheRDA for young adults (both males and females) has been set at 1.3 mg/d. For older adults, the RDA is slightly higher (1.5 mg/d for women, 1.7 mg/d for men).

Deficiency Symptoms of vitamin B_6deficiency include seborrheic dermatitis, glossitis, stomatitis, and cheilosis, as frequently seen with other B vitamin deficiencies (Chap. 74). In addition, severe vitamin B_6deficiency can lead to generalized weakness, irritability, peripheral neuropathy, abnormal electroencephalograms, and personality changes including depression and confusion. In infants, diarrhea, seizures, and anemia have been reported. Microcytic, hypochromic anemia is due to diminished hemoglobin synthesis, since the first enzyme involved in heme biosynthesis (amino-levulinate synthase) requiresPLP as a cofactor (Chap. 104). In some case reports, platelet dysfunction has also been reported. Since vitamin B_6 is necessary for the conversion of homocysteine to cystathionine, it is possible that chronic low-grade vitamin B_6deficiency may result in hyperhomocystinemia and increased risk of cardiovascular disease (Chaps. 242 and352).

Certain medications such as isoniazid, L-dopa, penicillamine, and cycloserine interact with PLPdue to a reaction with carbonyl groups. Oral contraceptives have been reported to decrease vitamin B_6status indicators, although the mechanism for this is uncertain. Alcoholism also decreases vitamin B_6status due to poor diet, liver disease, and the fact that acetaldehyde can compete with PLP for protein binding, leading to increased degradation and excretion. The increased ratio of aspartate aminotransferase (AST or SGOT) to alanine aminotransferase (ALT or SGPT) seen in alcoholic liver disease reflects the relative vitamin B_6dependence of ALT. Vitamin B_6requirements are higher in preeclampsia, eclampsia, and hemodialysis. Vitamin B_6dependency syndromes that require pharmacologic doses of vitamin B_6 are rare, but include cystathionine b-synthase deficiency, pyridoxine-responsive (primarily sideroblastic) anemias, and gyrate atrophy with chorioretinal degeneration due to decreased activity of the mitochondrial enzyme ornithine aminotransferase. In these situations, 100 to 200 mg/d of oral vitamin B_6 are required for treatment.

High doses of vitamin B_6 have been used to treat carpal tunnel syndrome, premenstrual tension, schizophrenia, autism, and diabetic neuropathy but have not been found to be effective.

The laboratory diagnosis of vitamin B_6deficiency is generally made on the basis of low plasmaPLPvalues (<20 nmol/L). Other measures of vitamin B_6deficiency include low erythrocyte levels of PLP, low plasma pyridoxal, and low urinary levels of 4-pyridoxic acid. Treatment of vitamin B_6deficiency is 50 mg/d; higher doses of 100 to 200 mg/d are given if vitamin B_6deficiency is related to medication use. Vitamin B_6should not be given with L-dopa, since the vitamin interferes with the action of this drug.

Toxicity The safe upper limit for vitamin B_6 has been set at 100 mg/d, although the lowest dose at which toxicity (sensory neuropathy) has been seen is 500 mg/d. No adverse effects have been associated with high intakes of vitamin B_6 from food sources only. When toxicity occurs, it causes a severe sensory neuropathy, leaving patients unable to walk. Some cases of photosensitivity and dermatitis have also been reported.

VITAMIN C

Both ascorbic acid and its oxidized product dehydroascorbic acid are biologically active. Vitamin C participates in oxidation-reduction reactions and hydrogen ion transfer reactions. As an antioxidant, vitamin C donates electrons to quench reactive free radical and oxygen species. It also acts to regenerate other antioxidants such as vitamin E, flavonoids, and glutathione. Other actions of vitamin C include promotion of nonheme iron absorption, carnitine biosynthesis, and the conversion of dopamine to norepinephrine. Vitamin C is also important for connective tissue metabolism and cross-linking and is a component of many drug-metabolizing enzyme systems, particularly the mixed-function oxidase systems. As such, the vitamin participates in the synthesis of corticosteroids, aldosterone, and the metabolism of cholesterol. Vitamin C also participates in enzymatic reactions requiring a reduced metal, although the exact molecular basis for this role has not been delineated.

Absorption and Physiology Vitamin C is absorbed by an energy-dependent, saturable transport system, and a progressively smaller proportion of the vitamin is absorbed with increasing dose. Almost complete absorption of the vitamin occurs if <100 mg is administered in a single dose; however, only 50% or less is absorbed at doses >1 g.

Enhanced degradation and fecal and urinary excretion of vitamin C occur at higher intake levels. High levels of the reduced form of vitamin C are contained in white blood cells, lens tissue, and brain. The maximum body pool in adult males is approximately 1500 mg, and 3% of this body pool is turned over each day, resulting in a half-life of approximately 18 days.

Dietary Sources and Requirements Good dietary sources of vitamin C include citrus fruits, green vegetables (especially broccoli), tomatoes, and potatoes. Appreciable amounts of vitamin C may be consumed as an antioxidant food additive, and the consumption of five servings of fruits and vegetables a day provides vitamin C in excess of theRDA of 60 mg/d for males and females. Moreover, approximately 40% of the U.S. population takes vitamin C as a dietary supplement. Vitamin C requirements are increased slightly to 70 mg in pregnancy and are increased further to 90 to 95 mg/d during lactation. Smoking, hemodialysis, and stress (e.g., infection, trauma) appear to increase vitamin C requirements. "Natural forms" of vitamin C are no more bioavailable than synthetic forms.

Deficiency Vitamin C deficiency causes scurvy; in the United States, this is seen primarily among poor and elderly people and alcoholics who consume <10 mg/d of vitamin C. Vitamin C deficiency has also been described among individuals consuming macrobiotic diets. Scurvy occurs when the body pool for vitamin C drops to <300 mg/d and plasma levels drop to<11 umol/L. Symptoms of scurvy primarily reflect impaired formation of mature connective tissue and include bleeding into skin (petechiae, ecchymoses, perifollicular hemorrhages); inflamed and bleeding gums; and manifestations of bleeding into joints, the peritoneal cavity, pericardium, and the adrenal glands. Other generalized symptoms include weakness, fatigue, and depression. In children, vitamin C deficiency may cause impaired bone growth. Laboratory diagnosis of vitamin C deficiency is made on the basis of low plasma or leukocyte levels.

Administration of vitamin C (200 mg/d) results in marked improvement in the symptoms of scurvy in a matter of several days. High-dose vitamin C supplementation (e.g., 1 to 2 g/d) has been shown to slightly decrease the symptoms and duration of upper respiratory tract infections and to improve glycemic control. Vitamin C supplementation has also been reported to be useful in Chediak-Higashi syndrome (Chap. 64) and osteogenesis imperfecta (Chap. 351). It has been claimed that foods high in vitamin C may lower the incidence of certain cancers, particularly esophageal and gastric cancers. If proven, this effect may be due to the fact that vitamin C can prevent the conversion of nitrites and secondary amines to carcinogenic nitrosomines. However, one intervention study from China did not show vitamin C to be protective. Other chronic diseases for which diets high in vitamin C have been reported to be protective include cardiovascular disease, stroke, and cataracts. However, these studies are correlational, and no large-scale intervention studies have been reported.

Toxicity Taking>2 g of vitamin C in a single dose may result in abdominal pain, diarrhea, and nausea; doses >3 g have been reported to elevate blood levels of alanine aminotransferase, lactic acid dehydrogenase, and uric acid. Since vitamin C may be metabolized to oxalate, it has been feared that chronic, high-dose vitamin C supplementation could result in an increased prevalence of kidney stones. However, this has not been borne out in several trials, except in individual patients with preexisting renal disease. Thus, it is reasonable to advise patients with a past history of kidney stones not to take large doses of vitamin C. There is also an unproven, but possible risk that chronic high doses of vitamin C could promote iron overload in patients taking supplemental iron. High doses of vitamin C can induce hemolysis in patients with glucose-6-phosphate dehydrogenase deficiency, and doses >1 g/d can cause false-negative guaiac reactions as well as interfering with tests for urinary glucose.

BIOTIN

Biotin is a water-soluble vitamin with a bicyclic structure. The vitamin plays an important role in gluconeogenesis and fatty acid synthesis and serves as a CO_2carrier on the surface of both cytosolic and mitochondrial carboxylase enzymes. The vitamin also functions in the catabolism of specific amino acids (e.g., leucine).

Biotin in food sources is bound to protein from which it must be cleaved in order to be absorbed. The enzyme biotinidase dissociates the vitamin and facilitates its subsequent transport. Excellent food sources of biotin include liver, soy, beans, yeast, and egg yolks, although egg white contains the protein avidin that strongly binds the vitamin and reduces its bioavailability. Biotin is contained in moderate amounts in legumes, nuts, mushrooms, cauliflower, and certain cereals. Although biotin is synthesized by intestinal bacteria, the relative importance of this source in humans is uncertain. The recommended intake of biotin for adults is 30 ug/d and 35 ug/d in lactating women.

Biotin deficiency has been induced by experimental feeding of egg white diets and in patients with short bowels who received biotin-free parenteral nutrition. In the adult, biotin deficiency results in mental changes (depression, hallucinations), paresthesia, anorexia, and nausea. A scaling, seborrheic, and erythematous rash may occur around the eyes, nose, and mouth as well as on the extremities. In infants, biotin deficiency presents as hypotonia, lethargy, and apathy. In addition, the infant may develop alopecia and a characteristic rash that includes the ears. Two types of inherited infantile biotin deficiency states have been described. Multiple carboxylase deficiency syndrome is an autosomal recessive disorder that is expressed during the first week of life and is characterized by severe metabolic ketoacidosis and dermatitis. Treatment requires pharmacologic doses of biotin, using up to 10 mg/d. Late-onset infantile biotin deficiency due to absorptive and transport defects occurs between 3 and 6 months with dermatitis, seizures, ataxia, hypotonia, and variable metabolic acidosis. The laboratory diagnosis of biotin deficiency can be established based on a decreased urinary concentration.

PANTOTHENIC ACID

Pantothenic acid is a component of coenzyme A and phosphopantetheine, which are involved in fatty acid metabolism and the synthesis of cholesterol, steroid hormones, and all compounds formed from isoprenoid units. In addition, pantothenic acid is involved in the acetylation of proteins. Pantothenic acid is actively transported when given at low doses, but it is passively absorbed when given at high doses. The vitamin is excreted in the urine, and the laboratory diagnosis of deficiency is made on the basis of low urinary vitamin levels.

The vitamin is ubiquitous in the food supply. Liver, yeast, egg yolks, and vegetables are particularly good sources. The recommended adequate intake for adults is 5 mg/d. Human pantothenic acid deficiency has only been demonstrated in experimental feeding of diets low in pantothenic acid or by giving a specific pantothenic acid antagonist. The symptoms of pantothenic acid deficiency are nonspecific and include gastrointestinal disturbance, depression, muscle cramps, paresthesia, ataxia, and hypoglycemia. Pantothenic acid deficiency was thought to cause the burning feet syndrome seen in prisoners of war during World War II. No toxicity of this vitamin has been reported.

CHOLINE

Choline is a precursor for acetylcholine, phospholipids, and betaine. Choline is necessary for the structural integrity of cell membranes, cholinergic neurotransmission, lipid and cholesterol metabolism, and transmembrane signaling. Recently, a recommended adequate intake was set at 550 mg/d for adult males and 425 mg/d for adult females. Choline is thought to be a "conditionally essential" nutrient, in that de novo synthesis occurs in the liver and is less than the vitamin's utilization only under certain stress conditions. Choline deficiency has occurred in patients receiving parenteral nutrition devoid of choline. Deficiency results in fatty liver and elevated transaminase levels. The diagnosis of choline deficiency is made on the basis of low plasma levels.

Toxicity from choline results in hypotension, cholinergic sweating, diarrhea, salivation, and a fishy body odor. The upper limit for choline has been set at 3.5 g/d. Therapeutically, choline has been suggested for patients with dementia and for patients at high risk of cardiovascular disease, due to its ability to lower cholesterol and homocysteine levels. However, such benefits have yet to be documented.

VITAMIN A

Vitamin A, in the strictest sense, refers to retinol. However, the oxidized metabolites, retinaldehyde and retinoic acid, are also biologically active compounds. The term *retinoids* includes synthetic molecules that are chemically related to retinol. Retinaldehyde is the essential form of vitamin A that is required for normal vision, whereas retinoic acid is necessary for normal morphogenesis, growth, and cell differentiation. Retinoic acid does not function in vision and, in contrast to retinol, is not involved in reproduction. Vitamin A also plays a role in iron utilization, humoral immunity, T cell-mediated immunity, natural killer cell activity, and phagocytosis. Vitamin A is commercially available in esterified forms (e.g., acetate, palmitate) since it is more stable as an ester.

There are over 600 carotenoids in nature, and approximately 50 of these can be metabolized to vitamin A.b-Carotene is the most prevalent carotenoid in the food supply that has provitamin A activity. Although the breakdown ofb-carotene should theoretically yield two molecules of vitamin A, the conversion of carotenoids to vitamin A, in fact, is much less efficient. It is estimated that 6 ug or greater of dietaryb-carotene is equivalent to 1 ug of retinol, whereas 12 ug or greater of other dietary provitamin A carotenoids (e.g., cryptoxanthin,a-carotene) is equivalent to 1 ug of retinol.

Absorption and Metabolism Approximately 80% of preformed vitamin A is absorbed from food, and absorption is via a carrier-mediated mechanism at low concentrations and passive diffusion at high concentrations. Approximately 15 to 30% of provitamin A carotenoids are absorbed passively from the diet, and the absorption becomes much less efficient at high dosage. The absorption of both vitamin A and carotenoids are partially dependent on an adequate bile concentration within the intestinal lumen for the formation of micelles. Once a provitamin A carotene is absorbed into the epithelial cell, a small proportion of it is split to form vitamin A. At higher doses of b-carotene, the conversion to vitamin A is less efficient, thereby preventing vitamin A toxicity. The absorption of both vitamin A and intact b-carotene is via the lymphatics after chylomicron formation.

Hepatic clearance of vitamin A in chylomicrons is efficient, and the liver contains approximately 90% of the vitamin A reserves. Approximately 10 to 40% of a vitamin A dose is oxidized or conjugated in the liver and excreted in urine or bile. Of a given dose of vitamin A, approximately 50% enters the liver storage pool. Storage of vitamin A takes place in the lipid storage (Ito) cell of the liver, which is also a collagen-producing cell. The liver secretes vitamin A in the form of retinol, which is bound to retinol-binding protein. Once this has occurred, the retinol-binding protein complex interacts with a second protein, transthryetin. This trimolecular complex functions to prevent vitamin A from being filtered by the kidney glomerulus, to protect the body against the toxicity of retinol and to allow retinol to be taken up by specific cell-surface receptors that recognize retinol-binding protein. A certain amount of vitamin A enters peripheral cells even if it is not bound to retinol-binding protein. After retinol is internalized by the cell, it becomes bound to a series of cellular retinol-binding proteins, which function as sequestering and transporting agents as well as coligands for enzymatic reactions. Certain cells also contain retinoic acid-binding proteins, which have the same sequestering functions as well as enabling retinoic acid metabolism.

11-*cis*-Retinaldehyde functions as a visual pigment chromophore to capture light. Rhodopsin is composed of the protein opsin and retinaldehyde and is contained in the rod cells, whereas iodopsin is contained in cones. When the dark-adapted retina is exposed to light, the 11-*cis*-retinaldehyde contained in rhodopsin isomerizes to an all-*trans* form. This conformational change causes dissociation from the opsin, resulting in a nerve impulse and a visual response. Once the retina returns to dim light conditions, rhodopsin is regenerated.

Retinoic acid is a ligand for certain nuclear receptors that act as transcription factors. Two families of receptors (RAR and RXR receptors) are active in retinoid-mediated gene transcription. Retinoid receptors regulate transcription by binding as dimeric complexes to specific DNA sites, the retinoic acid response elements, in target genes (Chap. 327). The receptors can either stimulate or repress gene expression in response to their ligands. RAR binds all-*trans* retinoic acid and 9-*cis* retinoic acid, whereas RXR binds only 9-*cis* retinoic acid.

The retinoid receptors play an important role in controlling cell proliferation and differentiation. Retinoic acid is useful in the treatment of promyeolcytic leukemia (Chap. 111). In this case, a gene rearrangement fuses the RAR to one of several other genes [e.g., t(15;17)], causing an apparent block in cell differentiation. Treatment with retinoic acid activates the RAR, dissociating repressor complexes and leading to cell differentiation and more normal cell turnover. Retinoic acid is also used in the treatment of cystic acne because it inhibits keratinization, decreases sebum secretion, and possibly alters the inflammatory reaction (Chap. 56). RXRs dimerize with other nuclear receptors to function as coregulators of genes responsive to retinoids, thyroid hormone, and calcitriol. RXR agonists induce insulin sensitivity experimentally, perhaps because RXR is a cofactor for the peroxisome-proliferator-activated receptors (PPARs), which are targets for the thiazolidinedione drugs such as rosiglitazone and troglitazone (Chap. 333).

Dietary Sources The retinol equivalent (RE) is used to express the vitamin A value of food. One RE is defined as 1 ug of retinol (0.003491 mmol). In the past, 1 RE was considered to be equal to 6 ug of b-carotene, but additional studies indicate that 1 RE may, in fact, be equal to 12 to 20 ug ofb-carotene from a dietary source. In older literature, vitamin A was often expressed in international units (IU), with 1 RE being equal to 3.33 IU of retinol and 12 IU of b-carotene, but these units are no longer in current medical or scientific use. The RDA for vitamin A is set at 1000 RE for adult males and 800 RE for adult females.

Liver and fish are excellent food sources for preformed vitamin A; vegetable sources of provitamin A carotenoids include dark-green and -colored fruits and vegetables. Diets consisting mainly of rice, wheat, maize, and tubers can produce vitamin A deficiency, as few carotenoids are contained in these foods. In areas where these foods are staples, children are particularly susceptible to vitamin A deficiency because neither breast nor cow's milk supplies enough vitamin A to prevent deficiency. Areas of the world where vitamin A deficiency is particularly prevalent include parts of Africa, South America, and Southeast Asia. Vitamin A deficiency occurs in more than 250,000 children each year, resulting in blindness and a 50% mortality rate within the year. In western countries, vitamin A deficiency is seen primarily among patients with diseases associated with fat malabsorption (e.g., celiac sprue, short-bowel syndrome). Concurrent zinc deficiency can interfere with the mobilization of vitamin A from liver stores as well as the synthesis of rhodopsin in the eye; thus vitamin A deficiency is exacerbated by concurrent zinc deficiency. Alcohol also interferes with the conversion of retinol to retinaldehyde in the eye by competing for alcohol (retinol) dehydrogenase. Drugs that interfere with the absorption of vitamin A include mineral oil, neomycin, and cholestyramine.

Deficiency Symptoms of vitamin A deficiency include hyperkeratotic skin lesions, night blindness (inability to see in dim light), dryness of the eyes, xerosis, and Bitot spots, which are white patches of keratinized epithelium appearing on the sclera (Fig. 75-CD1). Aggressive xerophthalmia can result in corneal ulceration. If untreated, proteolytic destruction and rupture of the cornea ensues with permanent blindness, although vitamin A treatment of patients with corneal ulcers can also result in blindness due to permanent corneal scarring. Children with vitamin A deficiency have increased mortality, primarily from infectious diseases, measles, respiratory diseases, and diarrhea. Extremely low birth weight infants (<1000 g) should be treated parenterally with 5000 IU (1500 ug or RE) of vitamin A three times a week for 4 weeks.

There are no specific deficiency signs or symptoms that result from carotenoid deficiency. However, dietary carotenoids have been suggested to protect against cataract formation, low-density lipoprotein (LDL) oxidation, and certain cancers. It was hoped thatb-carotene would be an effective chemopreventive for cancer because numerous epidemiologic studies had shown that diets high inb-carotene were associated with lower incidences of cancers of the respiratory and digestive system. However, intervention studies using high doses of b-carotene actually resulted in more lung cancers than in placebo-treated groups. Non-provitamin A carotenoids, such as lutein and zeaxanthin, have been suggested to protect against macular degeneration. The non-provitamin A carotenoid lycopene has been suggested to protect against prostate cancer. However, the effectiveness of these agents has not been proven by intervention studies, and the mechanisms underlying these purported biologic actions are unknown.

The diagnosis of vitamin A deficiency is made by measurement of serum retinol (normal range, 30 to 65 ug/dL), tests of dark adaptation, impression cytology of the conjunctiva (decreased numbers of mucous-secreting cells), or measurement of body storage pools, either directly by liver biopsy or by isotopic dilution after administering a stable isotope of vitamin A.

Vitamin A deficiency with ocular changes should be treated by administering 100,000 IU (30 mg) of vitamin A intramuscularly, or 200,000 IU (60 mg) orally. In areas of endemic vitamin A deficiency, this is followed by vitamin A capsules of 200,000 IU at 6-month intervals. Vitamin A deficiency in patients with malabsorptive diseases, who have abnormal dark adaptation or symptoms of night blindness without ocular changes, should be treated for 1 month with 50,000 IU/d (15 mg/d) orally of a water micelle preparation of vitamin A. This is followed by lower maintenance doses with the exact amount determined by monitoring serum retinol.

Toxicity Acute toxicity of vitamin A was first noted in Arctic explorers after eating polar bear liver and has been seen after administration of 150 mg in adults or 100 mg in children. Acute toxicity is manifest by increased intracranial pressure, vertigo, diplopia, bulging fontanels in children, seizures, and exfoliative dermatitis; it may result in death. Chronic vitamin A intoxication has been seen in normal adults who ingest 50,000 IU/d (15 mg/d) of vitamin A for a period of several months and in children who ingest 20,000 IU/d (6 mg/d). Manifestations include dry skin, cheilosis, glossitis, vomiting, alopecia, bone pain, hypercalcemia, lymph node enlargement, hyperlipidemia, amenorrhea, and features of pseudotumor cerebri with increased intracranial pressure and papilledema. Liver fibrosis with portal hypertension and bone demineralization may also result from chronic vitamin A intoxication. When vitamin A is provided in excess of pregnant women, congenital malformations have included spontaneous abortions, craniofacial abnormalities, and valvular heart disease. In pregnancy, the daily dose of vitamin A should not exceed 10,000 IU (3 mg). Elderly individuals appear to be more prone to vitamin A intoxication, as are alcoholics and patients with liver disease. In fact acute hepatitis may precipitate vitamin A intoxication in patients who have extremely high vitamin A stores in the liver. It should be noted that the commercially available retinoid derivatives are also toxic, including 13-*cis*-retinoic acid, which has been associated with birth defects. As a result, contraception should be continued for a least 1 year, and possibly longer, in women who have taken 13-*cis* retinoic acid.

High doses of carotenoids do not result in toxic symptoms. However, carotenemia, which is characterized by a yellowing of the skin (creases of the palms and soles) but not the sclerae, may be seen after ingestion of >30 mg ofb-carotene on a daily basis. Hypothyroid patients are particularly susceptible to the development of carotenemia due to impaired breakdown of carotene to vitamin A. Reduction of carotenes from the diet results in the disappearance of skin yellowing and carotenemia over a period of 30 to 60 days.

VITAMIN D (SeeChap. 340).

VITAMIN E

Vitamin E is a collective name for a group of tocopherols and tocotrionols, the latter having an unsaturated sidechain. There are eight naturally occurring plant compounds with vitamin E activity. RRR-a tocopherol is the most active, while synthetic stereoisomers of vitamin E are less biologically active. Vitamin E acts as a chain-breaking antioxidant and is an efficient pyroxyl radical scavenger, which protectsLDLs and polyunsaturated fats in membranes from oxidation. A network of other antioxidants (e.g., vitamin C, glutathione) and enzymes maintains vitamin E in a reduced state. Vitamin E also inhibits prostaglandin synthesis and the activities of protein kinase C and phospholipase A_2.

Absorption and Metabolism Vitamin E is a fat-soluble vitamin and requires all the processes needed for micelle formation to be absorbed. About 15 to 40% is absorbed passively from a single physiologic dose, and there is less efficient absorption at high doses. Polyunsaturated fat may inhibit absorption. Vitamin E is taken up from chylomicrons by the liver, and an hepatica tocopherol transport protein is involved in intracellular vitamin E transport and incorporation into very low density lipoprotein (VLDL). The transport protein has particular affinity for the RRR isomeric form of a tocopherol; thus this natural isomer has the most biologic activity. In the circulation, vitamin E is bound to all lipoprotein classes and becomes widely distributed in tissues, with fat and muscle being the most important storage depots. Vitamin E metabolites are mainly excreted in feces, although some are also excreted in urine.

Requirement TheRDA for vitamin E is currently 10 mg for adults. Additional vitamin E is recommended during pregnancy (12 mg/d) and lactation (14 mg/d). Vitamin E is widely distributed in the food supply. The RRR-a isomers are particularly high in sunflower oil, safflower oil, and wheat germ oil;g tocotrionols are notably present in soybean and corn oils. Vitamin E is also found in meats, nuts, and cereal grains, and small amounts are present in fruits and vegetables. Vitamin E pills containing doses of 50 to 1000 mg are ingested by a large fraction of the U.S. population. In the older literature, 1 IU of vitamin E is equal to 1 mg *all*-racemica tocopherol acetate. Diets high in polyunsaturated fats may necessitate a slightly higher requirement for vitamin E.

Dietary deficiency of vitamin E does not exist. Vitamin E deficiency is seen only in severe and prolonged malabsorptive diseases, such as celiac disease, or after small-intestinal resection, leading to short-bowel syndrome. Children with cystic fibrosis or prolonged cholestasis may develop vitamin E deficiency characterized by areflexia and hemolytic anemia. Children with abetalipoproteinemia cannot absorb or transport vitamin E and become deficient quite rapidly. A familial form of isolated vitamin E deficiency also exists, which is due to a defect in thea tocopherol transport protein.

Vitamin E deficiency causes axonal degeneration of the large myelinated axons and results in posterior column and spinocerebellar symptoms. Peripheral neuropathy is initially characterized by areflexia, with progression to an ataxic gait, and by decreased vibration and position sensations. Ophthalmoplegia, skeletal myopathy, and pigmented retinopathy may also be features of vitamin E deficiency. The laboratory diagnosis of vitamin E deficiency is made on the basis of low blood levels of a tocopherol (<5 ug/mL, or <0.8 mg of a tocopherol per gram of total lipids).

TREATMENT

Symptomatic vitamin E deficiency should be treated with 800 to 1200 mg ofa tocopherol per day. Patients with abetalipoproteinemia may need as much as 5000 to 7000 mg/d. Children with symptomatic vitamin E deficiency should be treated with 400 mg/d orally of water-soluble esters; alternatively, 2 mg/kg per day may be administered intramuscularly. Vitamin E in high doses may protect against oxygen-induced retrolental fibroplasia and bronchopulmonary dysplasia in prematurity, as well as intraventricular hemorrhage of prematurity. Vitamin E has been suggested to increase sexual performance, to treat intermittent claudication, and to slow the aging process, but evidence for these properties is lacking. High doses (60 to 800 mg/d) of vitamin E have been shown in controlled trials to improve parameters of immune function, and there are two intervention studies showing that vitamin E at 400 to 800 mg/d may be protective against cardiovascular disease, possibly by inhibitingLDLoxidation. Also, supplemental intake of vitamin E (100 to 200 mg/d) has been associated with a decreased risk of cataracts.

Toxicity High doses of vitamin E (>800 mg/d) may reduce platelet aggregation and interfere with vitamin K metabolism and are therefore contraindicated in patients taking coumadin. Nausea, flatulence, and diarrhea have been reported at doses >1 g/d.

VITAMIN K

There are two natural forms of vitamin K: vitamin K I, also known as *phylloquinone*, from vegetable and animal sources, and vitamin K II, or *menaquinone*, which is synthesized by bacterial flora and found in hepatic tissue. *Menadione*, or vitamin K III, is a chemically synthesized pro-vitamin that can be converted to menaquinone by the liver. Phylloquinone and menaquinones differ only in their lipophilic sidechains, and both are destroyed in an alkaline pH and by ultraviolet light.

Absorption and Physiology As with other fat-soluble vitamins, vitamin K absorption is dependent on normal pancreatic function and the presence of bile salts. Phylloquinones are absorbed by a saturable energy-dependent mechanism in the proximal small intestine, whereas menaquinones are absorbed by passive diffusion in the small intestine and colon. Approximately 100 ug of vitamin K is stored in the liver as well as in lung, bone marrow, kidneys, and adrenal glands. Most vitamin K circulates bound toVLDL, although it is also carried byLDL and high-density lipoprotein (HDL). The half-life of vitamin K is only 1$_1$/$_2$days, despite the presence of a vitamin K regeneration cycle.

Vitamin K is necessary for the posttranslational carboxylation of glutamic acid, which is necessary for calcium binding to g-carboxylated proteins such as prothrombin (factor II); factors VII, IX, and X; protein C; protein S; and proteins found in bone (bone gla, matrix gla protein, and osteocalcin). The importance of vitamin K for bone mineralization is not known. Warfarin-type drugs inhibitg carboxylation by preventing the conversion of vitamin K to its active hydroquinone form. Vitamin E, at high doses, may act as a vitamin K antagonist.

Dietary Sources Vitamin K is found in green leafy vegetables such as kale and spinach, but appreciable amounts are also present in butter, margarine, liver, milk, ground beef, coffee, and pears. Vitamin K is present in vegetable oils and is particularly rich in olive oil and soybean oil. The recommended intake of vitamin K is 70 ug/d in adults. The average daily intake by Americans is estimated to be approximately 100 ug/d.

Deficiency The symptoms of vitamin K deficiency are due to hemorrhage, and newborns are particularly susceptible because of low fat stores, low breast milk levels of vitamin K, sterility of the infantile intestinal tract, liver immaturity, and poor placental transport. Intracranial bleeding, as well as gastrointestinal and skin bleeding, can be seen in vitamin K-deficient infants 1 to 7 days after birth. Thus, vitamin K (1 mg intramuscularly) is given prophylactically at the time of delivery.

Vitamin K deficiency in adults may be seen in patients with chronic small-intestinal disease (e.g., celiac disease, Crohn's disease), obstructed biliary tracts, or after small-bowel resection. Broad-spectrum antibiotic treatment can precipitate vitamin K deficiency by reducing gut bacteria, which synthesize menaquinones, as well as by inhibiting the metabolism of vitamin K. The diagnosis of vitamin K deficiency is usually made on the basis of an elevated prothrombin time or reduced clotting factors. Vitamin K may also be measured directly byHPLC. In addition, undercarboxylated prothrombin and low gla levels in urine are indicative of vitamin K deficiency. Vitamin K deficiency is treated using a parenteral dose of 10 mg. For patients with chronic malabsorption, 1 to 2 mg/d of vitamin K may be given orally, or 1 to 2 mg/week can be taken parenterally. Patients with liver disease may have an elevated prothrombin time because of liver cell destruction as well as vitamin K deficiency. If an elevated prothrombin time does not improve on vitamin K therapy, it can be assumed that it is not the result of vitamin K deficiency.

Toxicity Parenteral doses of the water-soluble vitamin K derivative (menadione) have been reported to cause hemolytic anemia and hypobilirubinemia in infants. Toxicity from dietary phylloquinones and menaquinones has not been described. High doses of vitamin K can impair the actions of oral anticoagulants.

TRACE MINERALS (See Table 75-1)

ZINC

Zinc is an integral component of many metalloenzymes in the body; it is involved in the synthesis and stabilization of proteins, DNA, and RNA and plays a structural role in ribosomes and membranes. Zinc is necessary for the binding of steroid hormone receptors and several other transcription factors to DNA and thereby plays an important role in the regulation of gene transcription. Zinc is absolutely required for normal spermatogenesis, fetal growth, and embryonic development.

Absorption and Physiology Zinc is absorbed in the small intestine by a carrier-mediated mechanism. The absorption of zinc from the diet is inhibited by dietary phytate, fiber, oxalate, iron, and copper, as well as by certain drugs including penicillamine, sodium valproate, and ethambutol. TheRDA for zinc is 15 mg in males and 12 mg in females, with an additional 3 mg in pregnancy and 4 to 7 mg during lactation. Supplemental zinc is recommended for women taking ³60 mg/d of iron during pregnancy.

Meat, shellfish, nuts, and legumes are good sources of bioavailable zinc, whereas zinc in grains is less available for absorption. Zinc is excreted mainly in the feces but also in urine and sweat. The body contains approximately 2 g of zinc, and high concentrations are found in liver, prostate, pancreas, bone, and brain (hippocampus and cerebral cortex), where the metal may function in neural transmission.

Deficiency Mild zinc deficiency has been described in many diseases including diabetes mellitus, AIDS, cirrhosis, alcoholism, inflammatory bowel disease, malabsorption syndromes, and sickle cell anemia (Figs. 75-CD2 and75-CD3). In these diseases, mild chronic zinc deficiency can cause stunted growth in children, decreased taste sensation (hypogusia), impaired immune function, and night blindness due to impaired conversion of retinol to retinaldehyde. Severe chronic zinc deficiency has been described as a cause of hypogonadism and dwarfism in several Middle Eastern countries. In these children, hypopigmented hair is also part of the syndrome. Acrodermatitis enteropathica is a rare autosomal recessive disorder characterized by abnormalities in zinc absorption. Clinical manifestations include diarrhea, alopecia, muscle wasting, depression, irritability, and a rash involving the extremities, face, and perineum. The rash is characterized by vesicular and pustular crusting with scaling and erythema. In addition, hypopigmentation and corneal edema have been described in these patients. Occasional patients with Wilson's disease have developed zinc deficiency as a consequence of penicillamine therapy. Patients on long-term parenteral nutrition have developed deficiency when zinc has been omitted from the total parenteral nutrition (TPN) solution.

The diagnosis of zinc deficiency is usually made by a serum zinc level of <12 umol/L (<70 ug/dL). Pregnancy and birth control pills may cause a slight depression in serum zinc levels, and hypoalbuminemia from any cause can result in hypozincemia. In acute stress situations, zinc may be redistributed from serum into tissues. Zinc deficiency may be treated with 60 mg elemental zinc, given orally twice a day. Zinc gluconate lozenges (13 mg elemental Zn every 2 h while awake) have been reported to reduce the duration and symptoms of the common cold in adults, but these studies are conflicting.

Toxicity Acute zinc toxicity after oral ingestion causes nausea and vomiting, fever, and respiratory distress. Zinc fumes from welding may also be toxic and cause fever, chills, excessive salivation, sweating, and headache. Chronic large doses of zinc may depress immune function and cause hypochromic anemia as a result of copper deficiency.

COPPER

Copper is an integral part of numerous enzyme systems including amine oxidases, ferrooxidase (ceruloplasmin), cytochrome-*c* oxidase, superoxide dismutase, and dopamine hydroxylase. As such, copper plays a role in iron metabolism, melanin synthesis, and central nervous system function; the synthesis and cross-linking of elastin and collagen; and the scavenging of superoxide radicals.

Copper is absorbed in the proximal small intestine, and 90% of circulating copper is bound to ceruloplasmin. The body contains 50 to 120 mg of copper, and high concentrations are found in liver, brain, heart, spleen, kidney, and blood. The U.S.RDA is 1.5 to 3 mg of copper intake per day, although World Health Organization recommendations are somewhat lower. Dietary sources of copper include shellfish, liver, nuts, legumes, bran, and organ meats, whereas milk is a very poor source. Copper is primarily excreted in the feces, and small amounts are also excreted in urine.

Deficiency Dietary copper deficiency is relatively rare, although it has been described in premature infants fed milk diets and in infants with malabsorption. Signs and symptoms of copper deficiency include a hypochromic-normocytic anemia, osteopenia, depigmentation, mental retardation, and psychomotor abnormalities. Copper deficiency anemia has been reported in patients with malabsorptive diseases and nephrotic syndrome and in patients treated for Wilson's disease with chronic high doses of oral zinc, which can interfere with copper absorption. Menkes kinky hair syndrome is an X-linked metabolic disturbance of copper metabolism characterized by mental retardation, hypocupremia, and decreased circulating ceruloplasmin (Chap. 351). It is caused by mutations in a copper-transporting *ATP7A* gene. Children with this disease often die within 5 years due to dissecting aneurysms or cardiac rupture.

The diagnosis of copper deficiency is usually made on the basis of low serum levels of copper (<65 ug/dL) and low ceruloplasmin levels (<18 mg/dL). Serum levels of copper may be elevated in pregnancy or stress conditions since ceruloplasmin is an acute-phase reactant.

Toxicity Toxicity due to copper is usually accidental and may include nausea, vomiting, diarrhea, and hemolytic anemia. In severe cases, kidney failure, liver failure, and coma may ensue. In Wilson's disease, mutations in the copper-transporting *ATP7B* gene lead to accumulation of copper in the liver and brain, with low blood levels due to decreased ceruloplasmin (Chap. 348). Indian childhood cirrhosis is another hereditary disease characterized by extremely high copper levels in the liver. The World Health Organization recommends that adult females should not ingest>10 mg/d and males should not take in>12 mg/d of copper.

SELENIUM

Selenium, in the form of selenocysteine, is a component of the enzyme glutathione peroxidase, which serves to protect proteins, cell membranes, lipids, and nucleic acids from oxidant molecules. Selenocysteine is also found in the deiodinase enzymes, which mediate the deiodination of thyroxine to the more active triiodothyronine (Chap. 330). Rich sources of selenium include seafood, muscle meat, and cereals, although the selenium content of cereal is determined by the soil concentration. Countries with low soil concentrations include parts of Scandinavia, China, and New Zealand. *Keshan disease* is an endemic cardiomyopathy found in children and young women residing in regions of China where dietary intake of selenium is low (<20 ug/d). Concomitant deficiencies of iodine and selenium may worsen the clinical manifestations of cretinism. The adultRDAs for selenium in the United States are 55 and 70 ug/d for females and males, respectively. Low blood levels of selenium in various populations have been correlated with an increase in coronary artery disease and certain cancers, although the data are not consistent. Selenosis occurs at intakes of³400 ug/d and can result in nausea, vomiting, loss of hair, nail changes, peripheral neuropathy, and fatigue.

CHROMIUM

Chromium potentiates the action of insulin in patients with impaired glucose tolerance, presumably by increasing insulin receptor-mediated signaling. In addition, in some patients, improvement in blood lipid profiles has been reported. The usefulness of chromium supplements in muscle building are not substantiated. Rich food sources of chromium include yeast, meat, and grain products. Chromium deficiency has been reported to cause glucose intolerance, peripheral neuropathy, and confusion. The suggested intake of chromium for adults is 50 to 200 ug/d. Chromium in the trivalent state is found in supplements and is largely nontoxic; however, chromium-6 is a product of stainless steel welding and is a known pulmonary carcinogen, as well as causing liver, kidney, and central nervous system.

MAGNESIUM SeeChap. 340

FLUORIDE, MANGANESE, AND ULTRATRACE ELEMENTS

An essential function for *fluoride* in humans has not been described, although it is useful for the maintenance of structure in teeth and bone. An adequate intake for fluoride (on the basis of protection against dental caries) has been set at 3.1 and 3.8 mg/d in adult females and males, respectively. Adult fluorosis can occur at an intake of 10 mg/d for prolonged periods and results in mottled and pitted defects in tooth enamel as well as brittle bone (skeletal fluorosis). Much lower doses of fluoride (0.7 to 2 mg) can cause dental fluorosis or mottled enamel in infants and children.

Manganese and molybdenum deficiencies have been reported in patients with rare genetic abnormalities as well as in a few patients receiving prolongedTPN. Several manganese-specific enzymes have been identified (e.g., manganese superoxide dismutase). The estimated adequate daily dietary *manganese* intake for adults is 2 to 3 mg/d. Deficiencies of manganese have been reported to result in bone demineralization, poor growth, ataxia, and convulsions.

Ultratrace elements are those for which the need is <1 mg/d. Essentiality has not been established for most ultratrace elements, although *iodine* is clearly essential (Chap. 330). *Molybdenum* is necessary for the activity of sulfite and xanthine oxidase, and molybdenum deficiency may result in skeletal and brain lesions. The minimum required daily molybdenum intake is estimated to be ~25 ug/d. There is circumstantial evidence to suggest that *arsenic* (impaired growth, infertility), *boron* (impaired energy metabolism, impaired brain function), *nickel* (impaired-growth and reproduction), *silicon* (impaired growth) and *vanadium* (impaired skeletal formation) might also be essential.

(Bibliography omitted in Palm version)

Back to Table of Contents

76. ENTERAL AND PARENTERAL NUTRITION THERAPY - *Lyn Howard*

Parenteral and enteral nutrition provide life-sustaining therapy for patients who cannot take adequate food by mouth and who consequently are at risk for malnutrition and its effects, including susceptibility to infection, weakness and immobility; these features predispose the patient to aspiration pneumonia, pulmonary embolism, and pressure sores, all of which delay recovery from illness and increase mortality.

The term *enteral* refers to feeding via the gut and hence includes normal eating, but in the present context implies the infusion of formulas via a tube into the upper gastrointestinal tract. *Parenteral* refers to the infusion of nutrient solutions into the bloodstream. While these are different approaches to nutritional support, their goals are the same. Where feasible, enteral nutrition is the preferred route because it sustains the digestive, absorptive, and immunologic barrier functions of the gastrointestinal tract. The cost of enteral tube feeding is about one-tenth the cost of parenteral feeding.

Several developments have made tube feeding easier and more acceptable to patients. Small-bore pliable tubes have largely replaced large-bore rubber tubes, and double-lumen tubes are now available for simultaneous gastric suction and jejunal feeding when there is concern about gastric retention and aspiration. Enteral tubes can be inserted into the stomach or jejunum through the nose or, for long-term use, directly through the abdominal wall, using endoscopic, radiologic, or surgical techniques. Once the enterocutaneous tract is established, the protruding tube can be replaced by a "button" entry port, flush with the abdominal wall.

Complete nutrition by vein with sufficient calories, amino acids, minerals, and vitamins to permit wound healing, restoration of normal body composition of a cachectic patient, or growth in children became feasible in the 1960s with the development of high-flow central vein catheters. Parental nutrition is now available in all large hospitals and for some patients at home. Adequate calories and other nutrients can be delivered in the form of high-energy, isotonic intravenous fat solutions via a peripheral vein. However, peripheral veins usually cannot sustain such infusions indefinitely, and long-term support requires central venous access.

THE DECISION PROCESS FOR USING PARENTERAL OR ENTERAL NUTRITION

The decision to use specialized nutrition support should be based on the likelihood that averting or redressing malnutrition will improve the quality of life or the ability to recover from a serious illness.

Approximately 15 to 20% of hospitalized patients have evidence of malnutrition. Some malnourished patients benefit from specialized nutrition support; for others, wasting is an inevitable component of a terminal disease. Selecting the appropriate form of nutritional support for the patient requires knowledge of the potential benefits and risks of nutritional support, and the physician must inform the patient and family of these issues. A flow diagram of the steps involved in deciding whether specialized nutrition support should be undertaken and, if so, how, is depicted in Fig. 76-1. Like all life-support measures, these therapies are difficult to withdraw once started.

The first step requires consideration of the nutritional implications of the disease process. Is the condition or its treatment likely to impair appetite or food ingestion and absorption for a prolonged period of time? Because prevention of malnutrition is easier than repleting a cachectic patient, this issue must be considered in the initial evaluation (Chap. 74). The second step is to determine whether the patient is already sufficiently malnourished that lean body mass is decreased and critical functions such as healing and ventilation are impaired. The presence or absence of metabolic stress should be noted, since injury or infection can evoke the secretion of hormonal and cytokine factors that reduce the efficiency of nutrition repletion.

Weight loss without physiologic impairment is probably of no consequence. Physiologic impairment usually develops when more than 20% of body protein is lost and is more likely if key organ systems, such as the gut or liver, are directly affected by disease. Once it is recognized that the patient is malnourished or at major risk, the next question is whether specialized nutritional support will impact positively on the patient's response to the disease, improving the quality of life. While the provision of food and water is part of basic medical care, nutrition support by enteral or parenteral means is associated with risk and discomfort and should be recommended only when potential benefit exceeds risk and undertaken only with the consent of the patient.

If it is decided that preventing or treating malnutrition with specialized nutrition support would improve the prognosis and quality of life, the nutritional requirements must be determined and the route of nutrient delivery must be selected.

RISKS AND BENEFITS OF NUTRITION SUPPORT

The risks are determined primarily by the route required to deliver nutrition support. Providing nutritional requirements by special attention to oral intake of food, or by adding oral liquid supplements and monitoring food intake with frequent calorie counts, is the safest and least costly approach. It is also the most metabolically efficient since normal eating initiates the cephalic phase of digestion. Tube-fed infants grow better if the cephalic phase is stimulated by having the infant suck on a pacifier.

Anorexia, impairment of swallowing, or bowel disease may limit oral intake or the absorption of oral nutrients, in which case tube enteral nutrition is the next consideration. The bowel and its associated digestive organs derive 70% of their required nutrients directly from food in the lumen. In addition, glutamine, short-chain fatty acids, and nucleotides may have particular importance in maintaining gut integrity. Enteral feeding also supports gut function by stimulating splanchnic blood flow, neuronal activity, IgA antibody release, and secretion of gastrointestinal hormones such as epidermal growth factor that stimulate gut trophic activity. All these factors support the gut as an immunologic barrier against enteric pathogens, reducing the likelihood of bacterial overgrowth. For these reasons, some enteral nutrition should always be provided if possible, even when parenteral nutrition is required to provide most of the support. In the past, bowel rest through parenteral nutrition was thought to be the cornerstone of treatment of many severe gastrointestinal disorders, but the value of some enteral nutrition is now widely accepted, and strict bowel rest is rarely appropriate. Parenteral nutrition alone is necessary in severe hemorrhagic pancreatitis, necrotizing enterocolitis, prolonged ileus, and distal bowel obstruction.

Specialized nutrition support is expensive, accounting for >1% of all health care dollars. Consequently, hard clinical endpoints such as mortality rate, incidence of major complications, and duration of hospital stay are required of risk-benefit studies. Better nitrogen balance, increased levels of serum albumin, and improved delayed hypersensitivity are softer endpoints.Table 76-1 summarizes clinical trials that evaluate the use of specialized nutrition support in different disease states.

Perioperative Nutrition There is a clear-cut association between preoperative malnutrition and poor surgical outcome, but it has been difficult to demonstrate the benefit of preoperative parenteral nutrition on the outcome of surgery in malnourished patients. However, a meta-analysis of several small studies and a large cooperative Veterans Administration study indicates that preoperative parenteral nutrition does improve the outcome of severely malnourished surgical patients. In treated patients, noninfectious complications (e.g., pulmonary emboli and delayed wound healing) are reduced in the postoperative period. Effective preoperative restoration of nutrition by the parenteral route requires at least 7 to 14 days. If feasible, a safer and less costly approach is preoperative enteral nutrition, especially if provided at home.

Immediate postoperative nutritional support is appropriate for patients who received preoperative support and for patients unlikely to resume oral feeding within 10 days. The parenteral route is commonly used because of postoperative ileus or concern about disrupting a new bowel anastomosis. However, cautious jejunal feeding is often tolerated. Specialized enteral formulas supplemented with conditionally essential nutrients may be particularly beneficial in debilitated and immunosuppressed postoperative patients (Table 76-2).

Critical Illness Very early nutrition support (within the first 48 h) improves survival and reduces infections and length of hospital stay in patients with severe head injuries, burns, and major abdominal trauma. Enteral therapy, where feasible, is superior to parenteral therapy in several randomized trials. Enteral nutrition equally benefits malnourished and well-nourished injured patients. Animal studies show that enteral feeding reduces translocation of gut bacteria and the systemic catabolic response; however, these phenomena have not been substantiated in humans. Early enteral feeding may prevent bacterial overgrowth and decrease aspiration pneumonia.

The practical issue is obtaining jejunal access in a critically ill patient, who is not easily transferred out of the intensive care unit for endoscopic or radiologic tube placement. Sometimes a nasal or percutaneous combined gastric suction and jejunal feeding tube can be inserted at the bedside. If a surgical laparotomy is indicated, a feeding tube can be placed simultaneously.

Most studies of enteral feeding in critically ill patients used either a general polymeric formula or one with hydrolyzed protein. Formulas supplemented with conditionally essential nutrients reduce infections and length of hospital stay. Parenteral formulas enriched with large amounts of branched-chain amino acids (BCAA) improve nitrogen balance but do not appear to affect clinical outcome.

Cancer Cachexia Early nonrandomized studies suggested that patients with cancer cachexia benefitted from parenteral nutrition, but randomized trials demonstrated more risk than benefit for patients receiving chemotherapy or radiation. Severely malnourished cancer patients undergoing surgery benefit from preoperative parenteral nutrition, as do other malnourished patients.

In two randomized, prospective trials, patients undergoing bone marrow transplantation had better long-term survival after parenteral or enteral nutrition in the cytoreduction phase; nutrition support did not influence the initial infection rate or the frequency of graft rejection or graft-versus-host disease. Immediate morbidity is reduced if glutamine supplements are added to the parenteral nutrition solution. Parenteral nutrition continued at home delays return to oral feeding. For cancer patients with unresectable upper gastroinstestinal cancer, enteral feeding is usually justified if it is desired by the patient and family. Parenteral feeding should be provided only if clinical improvement can be expected and when quality survival at home for several months is predicted.

Liver Failure Malnutrition is common in advanced liver disease. Patients with acute or chronic liver failure have decreased levels ofBCAA and elevated levels of aromatic amino acids (AAA) in plasma and cerebrospinal fluid. Randomized, prospective trials with parenteral and enteral formulas high in BCAA and low in AAA have demonstrated better nitrogen balance and less risk of encephalopathy. One large multicenter study also reported improved survival. The BCAA-enriched formulas are expensive and should be used only in patients who have encephalopathy or who develop encephalopathy when fed a standard protein formula providing 0.8 g protein/kg per day.

Renal Failure Since renal failure is associated with impaired nitrogen excretion, it is rational to assume that protein restriction might benefit patients with both acute and chronic renal failure. Patients with acute renal failure given parenteral calories and amino acids have fewer infectious complications and a better chance of leaving the hospital than similar patients given only calories. An early randomized study showed benefit when essential amino acids were the sole source of nitrogen, but in other studies, standard solutions supplying both essential and nonessential amino acids provide similar advantage. Thus, the benefit of using expensive formulas containing only essential amino acids or their keto analogues is not established. A large national study failed to show any benefit of a low-protein diet on slowing progression of renal impairment in patients on chronic dialysis. Some 15 to 20% of patients on chronic dialysis have significant nutritional impairment, usually due to profound anorexia. The anorexia may improve with stepped-up dialysis or treatment of gastritis but usually persists. The resulting growth impairment in younger patients has been treated with supplemental tube enteral nutrition. This approach has not been widely used in adult patients. Limited parenteral calories and amino acids can be provided in the last 90 min of hemodialysis treatment (intradialytic parenteral nutrition). This may improve appetite, serum protein levels, and body weight. No randomized studies have documented better survival, so the appropriateness of this regimen is not established. Standard dialysis uses glucose to provide an osmotic load, and some glucose calories are absorbed. During continuous ambulatory peritoneal dialysis, amino acids can be substituted for glucose and are also partly absorbed, offsetting the loss of endogenous amino acids into the dialysate. This approach to nutrition repletion is expensive and also awaits a randomized study.

Pancreatitis Parenteral nutrition does not improve the outcome of patients with mild or moderate pancreatitis. However, in severe pancreatitis, survival decreases as malnutrition becomes more severe. When parenteral nutrition support was delayed beyond 72 h, patients with severe pancreatitis had a threefold higher complication and mortality rate, compared to similar patients treated earlier. In the absence of severe hyperlipidemia or thrombocytopenia, intravenous lipids appear safe and are especially useful if glucose intolerance is present. Several studies report successful enteral jejunal feeding in acute pancreatitis and, compared to parenteral nutrition, the inflammatory response and infectious complications are less.

Inflammatory Bowel Disease Evidence of nutritional deficiencies such as weight loss, growth failure, anemia, and hypoalbuminenia are common in inflammatory bowel disease (IBD), more so in Crohn's disease than in ulcerative colitis (Chap. 287). Nutrition support plays a role in correcting these nutritional deficiencies, particularly prior to elective surgery. Since IBD often improves with diversion of the fecal stream, the question is whether bowel rest and parenteral nutrition have a role as primary treatment. However, randomized, prospective studies have shown no special benefit from bowel rest. Elemental diets are not quite as effective as glucocorticoids for inducing remission in acute Crohn's disease but may be preferable in children to avoid growth impairment. Relapse is common when a regular diet is resumed. In controlled studies, remissions are prolonged if the Crohn's patient does not return to a regular diet but instead eliminates from the diet those foods that induce gastrointestinal symptoms. For the majority of Crohn's patients this leads to avoidance of cereals, yeast, green vegetables, and, early on, dairy products. Because of the possibility that diets high in omega-3 fatty acids have a beneficial effect in immune disorders by altering prostaglandin synthesis, their value in IBD is under investigation. Some studies suggest that high-fiber diets benefit IBD patients, but fiber can also cause obstruction in patients with bowel strictures.

Short Bowel Syndrome Before the advent of parenteral nutrition, patients with acute short bowel syndrome from mesenteric vascular infarction or massive small bowel surgical resection seldom survived. Parenteral nutrition has allowed many patients to survive indefinitely with only a foot or two of small intestine. In some, the remaining bowel eventually adapts and allows the absorption of adequate calories and protein. This is especially true of patients who retain their iliocecal valve and colon. However, fluid and electrolyte imbalance may persist, necessitating some parenteral fluid and electrolyte support. A gradual switch to overnight tube enteral hydration or constant sipping of an electrolyte solution may allow discontinuation of all parenteral support.

Pulmonary Disease Weight loss in patients with advanced pulmonary disease is due to increased work of breathing and poor food intake. Patients with chronic pulmonary disease who are<90% of their ideal weight have a higher 5-year mortality, independent of pulmonary status. The recommended energy intake for these patients is 1.7 times their resting energy expenditure. The use of a low-carbohydrate formula is beneficial in the weaning of patients from ventilators, but the superiority of such formulas in ambulatory patients with chronic lung disease is not established. In cystic fibrosis, malnutrition may hasten pulmonary deterioration, and enteral tube feeding enhances growth and stabilizes or improves pulmonary function, particularly in young children. Tube feeding is safest when delivered into the jejunum. Postpyloric feeding is no safer than gastric feeding.

HIV Disease Specialized nutrition support can replete body cell mass if the weight loss is due to inadequate oral intake caused by oral or esophageal disease or to inadequate intestinal absorption, which is common in HIV patients with cryptosporidiosis or microsporidiosis infections (Chap. 309). The route of nutrition support has usually been parenteral, but patients respond equally well with an isocaloric semi-elemental oral diet. Patients using the oral supplement experience a better quality of life, and their medical costs are significantly lower. Wasting due to systemic infection and increased cytokine secretion is not redressed by specialized nutrition support. Survival, CD4+ counts, and intestinal function also are not improved by specialized nutrition support.

Pregnancy Severe hyperemesis gravidarum can make any oral or tube enteral nutrition impossible, and profound weight loss and ketosis may harm the developing fetus. The underlying mechanism of the disorder is not understood, but it is cured by abortion or delivery. Temporary parenteral nutrition usually results in a successful outcome, but nausea and vomiting tend to persist, despite bowel rest.

Home Parenteral and Enteral Nutrition Some patients require long-term nutrition support, and for many this can be administered at home. Clinical outcomes of patients with severe intestinal disorders that used either parenteral or enteral nutrition are summarized in Table 76-3. Nutrition support is not usually appropriate in terminally ill patients but is an option if the patient is expected to survive for several months. Such therapy must make sense to the patient, and sufficient help must be available so the treatment can be given at home without undue hazard. Both home therapies are relatively safe, with<5% therapy-related mortality.

THE DESIGN OF INDIVIDUAL REGIMENS

Fluid Requirements These can be estimated by adding the normal daily requirement (120 mL/kg of body weight for infants, 35 mL/kg of body weight for adults) to any abnormal loss. If the patient is on parenteral therapy, any enteral intake should be subtracted from the estimate (Table 76-4). Since abnormal loss of enteric fluid implies significant mineral losses, extra amounts of these nutrients, as well as fluid (Table 76-5), must be added to the parenteral formula.

Energy Requirements These can be determined as outlined in Chaps. 73 and74. In the long run, energy expenditure dictates energy requirements, but in the early phase of nutrition repletion, requirements may not reflect expenditure. For example, malnourished patients are hypometabolic and may expend only 85 kJ/kg (20 kcal/kg) per day, but more calories are needed both for tissue repletion and because the metabolic rate increases with refeeding. Conversely, a highly stressed patient (sepsis, trauma) may expend 165 kJ/kg (40 kcal/kg) per day with a significant proportion of the calories coming from protein breakdown and gluconeogenesis and from catecholamine-induced lipolysis. Oxidation of exogenous glucose plateaus at 100 kJ/kg (25 kcal/kg) per day, and administering additional glucose induces hepatic steatosis. Providing such patients with additional calories as exogenous fat does not suppress endogenous lipolysis. Furthermore, lipid solutions are made from vegetable oil and egg phospholipid and lack apoproteins, which they acquire from endogenous lipoproteins.

Initially, the artificial chylomicron may be taken up by the reticuloendothelial system enhancing its blockade. For all these reasons, modest hypocaloric glucose feeding with minimal parenteral fat is safer in the acutely stressed subject.

Parenteral lipid solutions are available as 10 or 20% isotonic solutions and are infused separately from amino acids and glucose or as a combined "three-in-one solution," obviating the need for an extra pump. Three-in-one solutions are less stable than the glucose and amino acid mix, and destabilized fat particles have the potential to coalesce into larger droplets, becoming fat emboli. For this reason, three-in-one solutions have a shorter storage life and must be mixed by a pharmacist knowledgeable about the correct mixing sequence and safe levels of electrolytes and trace elements. Iron, for example, cannot be added to this solution.

Polyunsaturated vegetable oils are used in most enteral formulas because they are absorbed better than animal fat by a diseased gastrointestinal tract. Fat must supply the essential fatty acid requirement (1 to 4% of energy from linoleic and linolenic acid) (Table 76-6). Larger amounts (30% of energy) are safe in relatively stable patients and avoid the problems of providing large amounts of glucose (e.g., hyperglycemia and hepatic steatosis). Substituting omega-3 polyunsaturated fish oils for polyunsaturated vegetable fats may reduce the catabolic response to burn injury, trauma, and radiation by reducing the synthesis of prostaglandins that enhance the inflammatory response (Table 76-2). Such fats are available in enteral formulas and are currently being tested in parenteral formulas.

Protein or Amino Acid Requirements The recommended dietary protein allowance of 0.8 g/kg per day is adequate for nonstressed patients, such as a starved patient with a high-grade esophageal stricture. Catabolic patients, in contrast, may require up to 1.5 g/kg per day of protein to induce positive nitrogen balance and reconstitute normal body mass. Early studies showed that recombinant human growth hormone (rHGH) increases lean body mass. However, subsequent trials have shown that it is associated with increased mortality in critically ill patients, and it should not be used in this setting.

In a stable patient the adequacy of protein support can be assessed by analyzing protein balance:

where protein loss= [(24-h urine urea nitrogen (g) + 4)´ 6.25]. Over a long period, protein balance is assessed by documenting wound healing, restoration of normal body composition, or resumption of longitudinal growth. In states of disturbed protein utilization (e.g., renal and hepatic failure), azotemia and abnormal plasma amino acid patterns develop. The benefit of special enteral and parenteral solutions that correct these aberrations is only established in hepatic encephalopathy (see "Risks and Benefits of Nutrition Support").

Certain nutrients that can normally be synthesized endogenously become essential in severely ill patients when endogenous production or salvage pathways are impaired. This is true of glutamine, nucleotides, and the products of methionine metabolism (Table 76-2). Glutamine, an important fuel for the enterocyte and lymphocyte, is fairly insoluble and is absent from standard parenteral formulas and present in low concentrations in most enteral formulas. Soluble glutamine-containing dipeptides are under investigation.

Nucleotides and their related metabolic products have beneficial effects on the immune system, growth of the small intestine, lipid metabolism, and hepatic function. Nucleotides can be synthesized de novo in all cells only in small amounts, and the body therefore depends on dietary nucleotides or on salvage pathways that recycle nucleotides from purine and pyrimidine turnover. Nutritionally depleted patients benefit from formulas enriched in nucleotides.

When amino acids are infused systemically, rather than via the more physiologic portal vein, methionine, the only sulfur-containing amino acid in most parenteral solutions, is transaminated in peripheral tissues rather than transulfurated in the liver. As a result, downstream sulfur products such as carnitine, taurine, and glutathione become relatively deficient (Chap. 352). Preliminary studies suggest that the addition of an intermediate compound, *S*-adenosyl methionine, to parenteral solutions results in less cholestasis.

Mineral and Vitamin Requirements Parenteral and enteral mineral and vitamin requirements are summarized in Table 76-6. Electrolyte modifications are necessary if the patient has significant gastrointestinal losses (Table 76-5) or renal failure. Requirements of some minerals and vitamins are higher when administered parenterally for several reasons: (1) many micronutrients delivered into the systemic rather than the portal circulation are not captured by the liver and instead pass directly into the urine; (2) patients with bowel disease may have enteric loss of sodium, potassium, chloride, and bicarbonate and malabsorption of divalent cations, fat-soluble vitamins, and vitamin B_{12}; and (3) nutrients may adhere to the tubing and delivery bags, and exposure to oxygen and light may destroy vitamins (particularly vitamin A).

PARENTERAL NUTRITION

Infusion Technique and Patient Monitoring Partial and short-term total parenteral nutrition can be provided via a peripheral vein if the majority of the energy is supplied by isotonic fat solutions; long-term total parenteral nutrition using glucose as the chief energy source requires administration via a central vein catheter so the hypertonic solution can be rapidly diluted in a high-flow system. The preferred site for central vein infusion is the superior vena cava. Access sites and catheter choices are summarized inTable 76-7. Peripherally inserted central catheters are the most economical option for short-term parenteral nutrition. In one randomized study, the number of catheter-related infections was the same with peripherally and centrally inserted catheters. Tunneled catheters and implanted subcutaneous ports require operating room insertion and are more stable for long-term use. Central catheters should be changed when clinically indicated; routine changes are costly and hazardous. Chlorhexidine solution is a more effective local antiseptic than iodophor or alcohol. Although transparent dressings are helpful in stabilizing catheters and allow easy observation of the skin site, the incidence of catheter-related sepsis is higher than with traditional dry gauze dressings; newer transparent dressings that trap less moisture are under investigation. Catheters made from Silastic material or polyurethane are associated with lower complications than polyvinylchloride catheters. Several types of needleless systems use hub valves, and contamination rates are higher with these devices when used for long-term parenteral nutrition. Appropriate clinical and laboratory monitoring for patients on parenteral nutrition are summarized inTable 76-8.

Complications (See Table 76-9)

Mechanical The insertion of a central venous catheter should be done only by trained personnel under aseptic techniques. Major mechanical complications include pneumothorax; hemothorax from laceration of the subclavian artery or vein; brachial plexus injury; and malpositioning of the catheter in a cerebral vein, the azygos vein, or the right ventricle. The correct catheter position must be confirmed by x-ray before hypertonic nutrient solution is infused. Catheters can subsequently dislocate, develop leaks, or become detached from the hub and embolize into the heart or pulmonary artery. Catheter thrombosis may occur, especially if the catheter is used for withdrawing blood samples, and extension of the thrombosis to the central vein is frequently coincident with infection. Thrombosed catheters can sometimes be unblocked by urokinase treatment. The addition of low-dose heparin (1000 U/L) to limit thrombosis in parenteral catheters is controversial; no randomized, controlled studies demonstrate benefit, and heparin can contribute to loss of bone mineral, which is already a problem with long-term parenteral nutrition.

Metabolic Fluid overload can cause congestive heart failure, particularly in elderly and debilitated patients. Glucose overload can cause an osmotic diuresis or stimulate insulin secretion, which in turn promotes extracellular to intracellular shifts of potassium and phosphorous. Such shifts are most dangerous in cachectic patients with depletion of potassium and phosphorus stores and can cause arrhythmias, cardiopulmonary dysfunction, and neurologic symptoms. To avoid these problems, parenteral nutrition should be started slowly and monitored carefully. Glucose content is increased gradually as the patient demonstrates tolerance of the high glucose load. Late metabolic complications include cholestatic liver disease with bile sludging and gallstone formation. The exact cause of the liver disease is not understood, but lack of enteral stimulation to bile flow and defective sulfur amino acid metabolism and cholesterol solubilization appear to play a role. Cholestasis is less likely to occur if some enteral feeding is maintained. Parenteral nutrition induces hypercalciuria, which can result in negative calcium balance and osteopenia. Hypercalciuria may have several causes, including the high fixed-acid load of infused amino acids and the bisulfite preservative in parenteral solutions. Earlier, protein hydrolysates were used as an amino acid source and were contaminated with aluminum, which blocked bone mineralization. Aluminum is still a contaminant of some additives such as calcium gluconate. Once patients on long-term parenteral nutrition change from catabolic breakdown to sustained anabolism, deficiencies of micronutrients such as essential fatty acids, trace minerals, and vitamins may develop unless they are supplied in adequate amounts (Table 76-6).

Infectious Infection of the access line rarely occurs in the first 72 h, and fever during this period is usually due to infection elsewhere or some other cause. Infection of the access line is likely if the fever defervesces when the infusion of the parenteral formula is tapered.

Positive central line cultures suggest catheter sepsis, especially if no other infectious source is identified and if the organism is *Staphylococcus* or *Candida*. Although removal of the central catheter may allow fungemia to clear spontaneously, antibiotic therapy is recommended for bacterial infections and the more invasive fungi. Catheter sepsis rates are similar in single lumen central lines dedicated to parenteral nutrition whether inserted peripherally via the subclavian vein or tunneled; multiple-lumen catheters are associated with a greater incidence of sepsis. While there is no evidence to support the use of prophylactic antibiotics, recurrent catheter sepsis may be avoided if cuffs are used around the catheter exit site or small amounts of an antibiotic solution are left in the line along with a heparin lock.

ENTERAL NUTRITION

Tube Placement and Patient Monitoring The types of enteral feeding tubes, methods of insertion, their clinical uses, and potential complications are outlined inTable 76-10. The different types of enteral formulas are listed in Table 76-11. Patients on enteral feeding are at risk for many of the same metabolic complications as those receiving parenteral nutrition and should be monitored in the same way (Table 76-8). Since small-bore tubes are easily displaced, tube position should be checked at intervals by aspirating and measuring the pH of the gut fluid (<4 in stomach,>6 in jejunum).

Complications

Aspiration The debilitated patient with poor gastric emptying and impairment of swallowing and cough mechanisms is at risk for aspiration; this is particularly so for those on respirators. Tracheal suctioning induces coughing and gastric regurgitation, and cuffs on endotracheal or tracheostomy tubes seldom provide protection against aspiration. Under these circumstances, it may be safer to use a large-bore feeding tube to allow for temporary removal of gastric contents during tracheal suction or to use jejunal feeding. Constant gastric infusion of an enteral formula is better tolerated in sick patients than intermittent bolus feeding. A continuous infusion is best achieved with a pump, especially when using fine-bore tubes that have a greater potential to clog. If long-term feeding is anticipated, endoscopic, radiologic, or surgical placement of a gastric tube is preferred by most patients. For long-term ambulatory patients, a gastrostomy tube can be converted to a gastric "button," an access device that is flush with the skin. A nasojejunal tube reduces the risk of aspiration. However, fluoroscopically guided placement of fine-bore tubes through the pylorus is time-consuming, and such tubes frequently pull back into the stomach. A percutaneous combined gastric-suction and jejunal-feeding tube is more reliable. This can be placed radiologically, endoscopically, or surgically.

Diarrhea Enteral feeding often causes diarrhea, especially if bowel function is compromised by bowel disease or drugs. The diarrhea may be controlled by the use of continuous feeding of fiber-containing formulas or by adding an anticholinergic medication to the formula. Diarrhea associated with enteral feeding does not necessarily imply inadequate absorption of nutrients, other than water and electrolytes. Furthermore, since luminal nutrients exert trophic effects on the gut mucosa and enhance the enteric immunologic barrier, it is often appropriate to persist with tube feeding, despite the diarrhea, even when this necessitates supplemental parenteral fluid support.

THE SCOPE AND COST OF NUTRITION SUPPORT

As many as 25% of patients entering tertiary care hospitals have central catheters placed, and 20 to 30% of these catheters are used for parenteral nutrition. The incidence of catheter-related infection reflects the severity of the underlying medical condition and varies from 2 to 30 per thousand catheter days, depending on the type of patients involved. In critically ill patients, catheter sepsis is associated with a 35% mortality rate and a high cost per survivor. Most catheter-related complications derive from faulty insertion and management of the catheter rather than defects in the device. In large tertiary care hospitals, the insertion and management of these lines by specially trained teams can reduce complications by 80%, impacting significantly on outcome and costs. A growing shift from parenteral to enteral nutrition also promises significant cost savings. Home parenteral nutrition costs approximately half as much as similar treatment in the hospital, and home enteral nutrition costs much less.

(Bibliography omitted in Palm version)

Back to Table of Contents

77. OBESITY - *Jeffrey S. Flier*

In a world where food supplies are intermittent, the ability to store energy in excess of what is required for immediate use is essential for survival. Fat cells, residing within widely distributed adipose tissue depots, are adapted to store excess energy efficiently as triglyceride and, when needed, to release stored energy as free fatty acids for use at other sites. This physiologic system, orchestrated through endocrine and neural pathways, permits humans to survive starvation for as long as several months. However, in the presence of nutritional abundance and a sedentary lifestyle, and influenced importantly by genetic endowment, this system increases adipose energy stores and produces adverse health consequences.

DEFINITION AND MEASUREMENT

Obesity is a state of excess adipose tissue mass. Although often viewed as equivalent to increased body weight, this need not be the case -- lean but very muscular individuals may be overweight by arbitrary standards without having increased adiposity. Body weights are distributed continuously in populations, so that a medically meaningful distinction between lean and obese is somewhat arbitrary. Obesity is therefore more effectively defined by assessing its linkage to morbidity or mortality.

Although not a direct measure of adiposity, the most widely used method to gauge obesity is the *body mass index* (BMI), which is equal to weight/height^2(in kg/m^2) (Fig. 77-1). Other approaches to quantifying obesity include anthropometry (skin-fold thickness), densitometry (underwater weighing), computed tomography (CT) or magnetic resonance imaging (MRI), and electrical impedance. Using data from the Metropolitan Life Tables, BMIs for the midpoint of all heights and frames among both men and women range from 19 to 26 kg/m^2; at a similar BMI, women have more body fat than men. Based on unequivocal data of substantial morbidity, a BMI of 30 is most commonly used as a threshold for obesity in both men and women. Large-scale epidemiologic studies suggest that all-cause, metabolic, and cardiovascular morbidity begin to rise (albeit at a slow rate) when BMIs are ³ 25, suggesting that the cut-off for obesity should be lowered. Some authorities use the term *overweight* (rather than obese) to describe individuals with BMIs between 25 or 27 and 30. A BMI between 25 and 30 should be viewed as medically significant and worthy of therapeutic intervention, especially in the presence of risk factors that are influenced by adiposity, such as hypertension and glucose intolerance.

The distribution of adipose tissue in different anatomic depots also has substantial implications for morbidity. Specifically, intraabdominal and abdominal subcutaneous fat have more significance than subcutaneous fat present in the buttocks and lower extremities. This distinction is most easily made by determining the waist-to-hip ratio, with a ratio >0.9 in women and>1.0 in men being abnormal. Many of the most important complications of obesity, such as insulin resistance, diabetes, hypertension, and hyperlipidemia, and hyperandrogenism in women, are linked more strongly to intraabdominal and/or upper body fat than to overall adiposity. The mechanism underlying this association is unknown but may relate to the fact that intraabdominal adipocytes are more lipolytically active than those from other depots. Release of free fatty acids into the portal circulation has adverse metabolic actions, especially on the liver.

PREVALENCE

Recent data from the National Health and Nutrition Examination Surveys (NHANES) show that the percent of the American adult population with obesity (BMI> 30) has increased from 14.5% (between 1976 and 1980) to 22.5% (between 1998 and 1994). As many as 50% of U.S. adults³20 years of age were overweight (defined as BMI> 25) between the years of 1998 and 1991. Because substantial health risks exist in many individuals with BMI between 25 and 30, the increasing prevalence of medically significant obesity raises great concern. Obesity is more common among women and in the poor; the prevalence in children is also rising at a worrisome rate.

PHYSIOLOGIC REGULATION OF ENERGY BALANCE

Substantial evidence suggests that body weight is regulated by both endocrine and neural components that ultimately influence the effector arms of energy intake and expenditure. This complex regulatory system is necessary because even small imbalances between energy intake and expenditure will ultimately have large effects on body weight. For example, a 0.3% positive imbalance over 30 years would result in a 9-kg (20-lb) weight gain. Alterations in stable weight by forced overfeeding or food deprivation induce physiologic changes that resist these perturbations: with weight loss, appetite increases and energy expenditure falls; with overfeeding, appetite falls and energy expenditure increases. This latter compensatory mechanism frequently fails, however, permitting obesity to develop when food is abundant and physical activity is limited. A major regulator of these adaptative responses is the adipocyte-derived hormone leptin, which acts through brain circuits (predominantly in the hypothalamus) to influence appetite, energy expenditure, and neuroendocrine function (see below).

Appetite is influenced by many factors that are integrated by the brain, most importantly within the hypothalamus (Fig. 77-2). Signals that impinge on the hypothalamic center include neural afferents, hormones, and metabolites. Vagal inputs are particularly important, bringing information from viscera, such as gut distention. Hormonal signals include leptin, insulin, cortisol, and gut peptides such as cholecystokinin, which signals to the brain through the vagus nerve. Metabolites, including glucose, can influence appetite, as seen by the effect of hypoglycemia to induce hunger; however, glucose is not normally a major regulator of appetite. These diverse hormonal, metabolic, and neural signals act by influencing the expression and release of various hypothalamic peptides [e.g., neuropeptide Y (NPY), Agouti-related peptide (AgRP),a melanocyte-stimulating hormone (MSH), and melanin concentrating hormone (MCH)] that are integrated with serotonergic, catecholaminergic, and opioid signaling pathways (see below). Psychological and cultural factors also appear to play a role in the final expression of appetite. Apart from rare syndromes involving leptin, its receptor, and the melanocortin system (see below), the defects in this complex appetite control network that account for common causes of obesity are not well understood.

Energy expenditure includes the following components: (1) resting or basal metabolic rate; (2) the energy cost of metabolizing and storing food; (3) the thermic effect of exercise; and (4) adaptive thermogenesis, which varies in response to chronic caloric intake (rising with increased intake). Basal metabolic rate accounts for about 70% of daily energy expenditure, whereas active physical activity contributes 5 to 10%. Thus, a significant component of daily energy consumption is fixed.

Adaptive thermogenesis occurs in *brown adipose tissue* (BAT), which plays an important role in energy metabolism in many mammals. In contrast to white adipose tissue, which is used to store energy in the form of lipids, BAT expends stored energy as heat. A mitochondrial *uncoupling protein* (UCP-1) in BAT dissipates the hydrogen ion gradient in the oxidative respiration chain and releases energy as heat. The metabolic activity of BAT is increased by a central action of leptin, acting through the sympathetic nervous system, which heavily innervates this tissue. In rodents, BAT deficiency causes obesity and diabetes; stimulation of BAT with a specific adrenergic agonist (b$_3$agonist) protects against diabetes and obesity. Although BAT exists in humans (especially neonates), its physiologic role is not yet established. Homologues of UCP-1 may mediate uncoupled mitochondrial respiration in other tissues.

THE ADIPOCYTE AND ADIPOSE TISSUE

Adipose tissue is composed of the lipid-storing adipose cell and a stromal/vascular compartment in which preadipocytes reside. Adipose mass increases by enlargement of adipose cells through lipid deposition, as well as by an increase in the number of adipocytes. The process by which adipose cells are derived from a mesenchymal preadipocyte involves an orchestrated series of differentiation steps mediated by a cascade of specific transcription factors. One of the key transcription factors is *peroxisome proliferator-activated receptor* g(PPARg), a nuclear receptor that binds the thiazoladinedione class of insulin-sensitizing drugs used in the treatment of type 2 diabetes (Chap. 333).

Although the adipocyte has generally been regarded as a storage depot for fat, it is also an endocrine cell that releases numerous molecules in a regulated fashion (Fig. 77-3). These include the energy balance-regulating hormone leptin, cytokines such as tumor necrosis factor (TNF)a, complement factors such as factor D (also known as adipsin), prothrombotic agents such as plasminogen activator inhibitor I, and a component of the blood pressure regulating system, angiotensinogen. These factors, and others not yet identified, play a role in the physiology of lipid homeostasis, insulin sensitivity, blood pressure control, and coagulation and are likely to contribute to obesity-related pathologies.

ETIOLOGY OF OBESITY

Though the molecular pathways regulating energy balance are beginning to be illuminated, the causes of obesity remain elusive. In part, this reflects the fact that obesity is a heterogeneous group of disorders. At one level, the pathophysiology of obesity seems simple: a chronic excess of nutrient intake relative to the level of energy expenditure. However, due to the complexity of the neuroendocrine and metabolic systems that regulate energy intake, storage, and expenditure, it has been difficult to quantitate all the relevant parameters (e.g., food intake and energy expenditure) over time in human subjects.

Role of Genes vs. Environment Obesity is commonly seen in families. Inheritance is usually not Mendelian, however, and it is difficult to distinguish the role of genes and environmental factors. Adoptees usually resemble their biologic rather than adoptive parents with respect to obesity, providing strong support for genetic influences. Likewise, identical twins have very similarBMIswhether reared together or apart, and their BMIs are much more strongly correlated than those of dizygotic twins. These genetic effects appear to relate to both energy intake and expenditure.

Whatever the role of genes, it is clear that the environment plays a key role in obesity, as evidenced by the fact that famine prevents obesity in even the most obesity-prone individual. In addition, the recent increase in the prevalence of obesity in the United States is too rapid to be due to changes in the gene pool. Cultural factors are also important -- these relate to both availability and composition of the diet and to changes in the level of physical activity. In industrial societies, obesity is more common among poor women, whereas in underdeveloped countries, wealthier women are more often obese. In children, obesity correlates to some degree with time spent watching television. High-fat diets may promote obesity, as may diets rich in simple (as opposed to complex) carbohydrates.

Specific Genetic Syndromes Obesity in rodents has been known for many years to be caused by a number of distinct mutations distributed through the genome. Most of these single-gene mutations cause both hyperphagia and diminished energy expenditure, suggesting a link between these two parameters of energy homeostasis. Identification of the *ob* gene mutation in genetically obese (ob/ob) mice represents a major breakthrough in the field. The ob/ob mouse develops severe obesity, insulin resistance, and hyperphagia, as well as efficient metabolism (e.g., it gets fat even when given the same number of calories as lean littermates). The product of the *ob* gene is the peptide leptin, a name derived from the Greek root *leptos*, meaning thin. Leptin is secreted by adipose cells and acts through the hypothalamus. Its level of production provides an index of adipose energy stores (Fig. 77-4). High leptin levels decrease food intake and increase energy expenditure. Another mouse mutant, db/db, which is resistant to leptin, has a mutation in the leptin receptor and develops a similar syndrome. The *ob* gene is present in humans and expressed in fat. Several families with morbid, early-onset obesity due to inactivating mutations in either leptin or the leptin receptor have been described, thus demonstrating the biologic relevance of leptin in humans. The obesity in these individuals begins shortly after birth, is severe, and is accompanied by neuroendocrine abnormalities. The most prominent of these is hypogonadotropic hypogonadism, which is reversed by leptin replacement. Central hypothyroidism and growth retardation are seen in the mouse model, but their occurrence in leptin-deficient humans is less clear. To date, there is no evidence to suggest that mutations or polymorphisms in the leptin or leptin receptor genes play a prominent role in common forms of obesity.

Mutations in several other genes cause severe obesity in humans (Table 77-1), each of these syndromes is rare. Mutations in the gene encoding proopiomelanocortin (POMC) cause severe obesity through failure to synthesizea-MSH, a key neuropeptide that inhibits appetite in the hypothalamus. The absence of POMC also causes secondary adrenal insufficiency due to absence of adrenocorticotropic hormone (ACTH), as well as pale skin and red hair due to absence of MSH. Proenzyme convertase 1 (PC-1) mutations are thought to cause obesity by preventing synthesis of a-MSH from its precursor peptide, POMC. a-MSH binds to the type 4 melanocortin receptor (MC4R), a key hypothalamic receptor that inhibits eating; mutations of this receptor also cause obesity. These three genetic defects, although rare, define a pathway through which leptin (by stimulating POMC and increasing MSH) restricts food intake and limits weight (Fig. 77-5).

In addition to these human obesity genes, studies in rodents reveal several other molecular candidates for hypothalamic mediators of human obesity or leanness. The *tub* gene encodes a hypothalamic peptide of unknown function; mutation of this gene causes late-onset obesity. The *fat* gene encodes carboxypeptidase E, a peptide-processing enzyme; mutation of this gene is thought to cause obesity by disrupting production of one or more neuropeptides.AgRP is coexpressed withNPY in arcuate nucleus neurons. AgRP antagonizesa-MSHaction at MC4 receptors, and its overexpression induces obesity. A putative activating mutation in the gene encoding PPARg, the adipocyte transcription factor required for adipogenesis, has been linked to obesity in a group of German subjects.

A number of complex human syndromes with defined inheritance are associated with obesity (Table 77-2). Although specific genes are undefined at present, their identification will likely enhance our understanding of more common forms of human obesity. In the Prader-Willi syndrome, obesity coexists with short stature, mental retardation, hypogonadotropic hypogonadism, hypotonia, small hands and feet, fish-shaped mouth, and hyperphagia. Most patients have a chromosome 15 deletion (Chap. 66). Laurence-Moon-Biedl syndrome involves obesity, mental retardation, retinitis pigmentosa, polydactyly, and hypogonadotropic hypogonadism.

Other Specific Syndromes Associated with Obesity

Cushing's Syndrome Although obese patients commonly have central obesity, hypertension, and glucose intolerance, they lack other specific stigmata of Cushing's syndrome (Chap. 331). Nonetheless, a potential diagnosis of Cushing's syndrome is often entertained. Cortisol production and urinary metabolites (17OH steroids) may be increased in simple obesity. Unlike in Cushing's syndrome, however, cortisol levels in blood and urine in the basal state and in response to CRH orACTH are normal; the overnight 1-mg dexamethasone suppression test is normal in 90%, with the remainder being normal on a standard 2-day low-dose dexamethasone suppression test.

Hypothyroidism The possibility of hypothyroidism should be considered when evaluating obesity, but it is an uncommon cause of obesity; hypothyroidism is easily ruled out by measuring thyroid stimulating hormone (TSH). Much of the weight gain that occurs in hypothyroidism is due to myxedema (Chap. 330).

Insulinoma Patients with insulinoma often gain weight as a result of overeating to avoid hypoglycemia symptoms (Chap. 334). The increased substrate plus high insulin levels promotes energy storage in fat. This can be marked in some individuals but is modest in most.

Craniopharyngioma and Other Disorders Involving the Hypothalamus Whether through tumors, trauma, or inflammation, hypothalamic dysfunction of systems controlling satiety, hunger, and energy expenditure can cause varying degrees of obesity (Chap. 328). It is uncommon to identify a discrete anatomic basis for these disorders. Subtle hypothalamic dysfunction is probably a more common cause of obesity than can be documented using currently available techniques. Growth hormone (GH), which exerts lipolytic activity, is diminished in obesity and increases with weight loss. Despite low growth hormone levels, insulin-like growth factor (IGF) I (somatomedin) production is normal, suggesting that GH suppression is a compensatory response to increased nutritional supply.

Pathogenesis of Common Obesity Obesity can result from increased energy intake, decreased energy expenditure, or a combination of the two. Thus, identifying the etiology of obesity should involve measurements of both parameters. However, it is nearly impossible to perform direct and accurate measurements of energy intake in free-living individuals. Obese people, in particular, appear to underreport intake. Measurements of chronic energy expenditure have only recently become available using doubly-labeled water or metabolic chamber/rooms. In subjects at stable weight and body composition, energy intake equals expenditure. Consequently, these techniques allow determination of energy intake in free-living individuals. The level of energy expenditure differs in established obesity, during periods of weight gain or loss, and in the pre- or postobese state. Studies that fail to take note of this phenomenon are not easily interpreted.

There is increased interest in the concept of a body weight "set point." This idea is supported by physiologic mechanisms centered around a sensing system in adipose tissue that reflects fat stores, and a receptor, or "adipostat," that is in the hypothalamic centers. When fat stores are depleted, the adipostat signal is low, and the hypothalamus responds by stimulating hunger and decreasing energy expenditure to conserve energy. Conversely, when fat stores are abundant, the signal is increased, and the hypothalamus responds by decreasing hunger and increasing energy expenditure. The recent discovery of the *ob* gene, and its product leptin, provides a molecular basis for this physiologic concept (see above).

What Is the Status of Food Intake in Obesity (Do the Obese Eat More Than the Lean?) This question has stimulated much debate, due in part to the methodologic difficulties inherent in determining food intake. Many obese individuals believe that they eat small quantities of food, and this claim has often been supported by the results of food intake questionnaires. However, it is now established that average energy expenditure increases as people get more obese, due primarily to the fact that metabolically active lean tissue mass increases with obesity. Given the laws of thermodynamics, the obese person must therefore eat more than the average lean person to maintain their increased weight. It may be the case, however, that a subset of individuals who are predisposed to obesity have the capacity to become obese initially without an absolute increase in caloric consumption.

What Is the State of Energy Expenditure in Obesity? The average total daily energy expenditure is higher in obese than lean individuals when measured at stable weight. However, energy expenditure falls as weight is lost, due in part to loss of lean body mass and to decreased sympathetic nerve activity. When reduced to near-normal weight and maintained there for a while, (some) obese individuals have lower energy expenditure than (some) lean individuals. There is also a tendency for those who develop obesity as infants or children to have lower resting energy expenditure rates than those who remain lean.

The physiologic basis for variable rates of energy expenditure (at a given body weight and level of energy intake) is essentially unknown. A mutation in the humanb$_3$adrenergic receptor may be associated with increased risk of obesity and/or insulin resistance in certain (but not all) populations. Homologues of theBATuncoupling protein, namedUCP-2 and UCP-3, have been identified in both rodents and humans. UCP-2 is expressed widely, whereas UCP-3 is primarily expressed in skeletal muscle. These proteins may play a role in disordered energy balance.

One newly described component of thermogenesis, called *nonexercise activity thermogenesis* (NEAT), has been linked to obesity. It is the thermogenesis that accompanies physical activities other than volitional exercise, such as the activities of daily living, fidgeting, spontaneous muscle contraction, and maintaining posture. NEAT accounts for about two-thirds of the increased daily energy expenditure induced by overfeeding. The wide variation in fat storage seen in overfed individuals is predicted by the degree to which NEAT is induced. The molecular basis for NEAT and its regulation are unknown.

Leptin in Typical Obesity The vast majority of obese people have increased leptin levels but do not have mutations of either leptin or its receptor. They appear, therefore, to have a form of functional "leptin resistance." Data suggesting that some individuals produce less leptin per unit fat mass than others or have a form of relative leptin deficiency that predisposes to obesity are at present contradictory and unsettled. The mechanism for leptin resistance, and whether it can be overcome by raising leptin levels, is not yet established. Some data suggest that leptin may not effectively cross the blood-brain barrier as levels rise. It is also possible that leptin signaling inhibitors are involved in the leptin-resistant state.

PATHOLOGIC CONSEQUENCES OF OBESITY

Obesity has major adverse effects on health. Morbidly obese individuals (>200% ideal body weight) have as much as a twelvefold increase in mortality. Morality rates rise as obesity increases, particularly when obesity is associated with increased intraabdominal fat (see above). It is also apparent that the degree to which obesity affects particular organ systems is influenced by susceptibility genes that vary in the population.

Insulin Resistance and Type 2 Diabetes Mellitus Hyperinsulinemia and insulin resistance are pervasive features of obesity, increasing with weight gain and diminishing with weight loss. Insulin resistance is more strongly linked to intraabdominal fat than to fat in other depots. The molecular link between obesity and insulin resistance has been sought for many years, with the major factors under investigation being: (1) insulin itself, by inducing receptor downregulation; (2) free fatty acids, known to be increased and capable of impairing insulin action; and (3) the cytokineTNF-a, which is produced by adipocytes, overexpressed in obese adipocytes, and capable of inhibiting insulin action. Despite insulin resistance, most obese individuals do not develop diabetes, suggesting that the onset of diabetes requires an interaction between obesity-induced insulin resistance and other factors that predispose to diabetes, such as impaired insulin secretion (Chap. 333). Obesity, however, is a major risk factor for diabetes, and as many as 80% of patients with type 2 diabetes mellitus are obese. Weight loss, even of modest degree, is associated with increased insulin sensitivity and often improves glucose control in diabetes.

Reproductive Disorders Disorders that affect the reproductive axis are associated with obesity in both men and women. Male hypogonadism is associated with increased adipose tissue, often distributed in a pattern more typical of females. In men>160% ideal body weight, plasma testosterone and sex hormone-binding globulin (SHBG) are often reduced, and estrogen levels (derived from conversion of adrenal androgens in adipose tissue) are increased (Chap. 335). Gynecomastia may be seen. However, masculinization, libido, potency, and spermatogenesis are preserved in most of these individuals. Free testosterone may be decreased in morbidly obese men whose weight exceeds 200% ideal body weight.

Obesity has long been associated with menstrual abnormalities in women, particularly in women with upper body obesity (Chaps. 52 and336). Common findings are increased androgen production, decreasedSHBG, and increased peripheral conversion of androgen to estrogen. Most obese women with oligomenorrhea have the polycystic ovarian syndrome (PCOS), with its associated anovulation and ovarian hyperandrogenism; 40% of women with PCOS are obese. Interestingly, most nonobese women with PCOS are also insulin-resistant, suggesting that insulin resistance, hyperinsulinemia, or the combination of the two are causative or contribute to the ovarian pathophysiology in PCOS in both obese and lean individuals. In obese women with PCOS, weight loss or treatment with insulin-sensitizing drugs often restores normal menses, along with a fall in estrone levels and normalized gonadotropin secretion. The increased conversion of androstenedione to estrogen, which occurs to a greater degree in women with lower body obesity, may contribute to the increased incidence of uterine cancer in postmenopausal women with obesity.

Cardiovascular Disease The Framingham Study revealed that obesity was an independent risk factor for the 26-year incidence of cardiovascular disease in men and women [including coronary disease, stroke, and congestive heart failure (CHF)]. The waist/hip ratio may be the best predictor of these risks. When the additional effects of hypertension and glucose intolerance associated with obesity are included, the adverse impact of obesity is even more evident. The effect of obesity on cardiovascular mortality in women may be seen atBMIs as low as 25. Obesity, especially abdominal obesity, is associated with an atherogenic lipid profile, with increased low-density lipoprotein (LDL) cholesterol, very low density lipoprotein and triglyceride, and decreased high-density lipoprotein cholesterol (Chap. 344). Obesity is also associated with hypertension. Measurement of blood pressure in the obese requires use of a larger cuff size to avoid artifactual increases. Obesity-induced hypertension is associated with increased peripheral resistance and cardiac output, increased sympathetic nervous system tone, increased salt sensitivity, and insulin-mediated salt retention; it is often responsive to modest weight loss.

Pulmonary Disease Obesity may be associated with a number of pulmonary abnormalities. These include reduced chest wall compliance, increased work of breathing, increased minute ventilation due to increased metabolic rate, and decreased total lung capacity and functional residual capacity (Chap. 250). Severe obesity may be associated with obstructive sleep apnea and the "obesity hypoventilation syndrome" (Chap. 263). Sleep apnea can be obstructive (most common), central, or mixed. Weight loss (10 to 20 kg) can bring substantial improvement, as can major weight loss following gastric bypass or restrictive surgery. Continuous positive airway pressure has been used with some success.

Gallstones Obesity is associated with enhanced biliary secretion of cholesterol, supersaturation of bile, and a higher incidence of gallstones, particularly cholesterol gallstones (Chap. 302). A person 50% above ideal body weight has about a sixfold increased incidence of symptomatic gallstones. Paradoxically, fasting increases supersaturation of bile by decreasing the phospholipid component. Fasting-induced cholecystitis is a complication of extreme diets.

Cancer Obesity in males is associated with higher mortality from cancer of the colon, rectum, and prostate; obesity in females is associated with higher mortality from cancer of the gallbladder, bile ducts, breasts, endometrium, cervix, and ovaries. Some of the latter may be due to increased rates of conversion of androstenedione to estrone in adipose tissue of obese individuals.

Bone, Joint, and Cutaneous Disease Obesity is associated with an increased risk of osteoarthritis, no doubt partly due to the trauma of added weight bearing. The prevalence of gout may also be increased (Chap. 322). Among the skin problems associated with obesity is acanthosis nigricans, manifested by darkening and thickening of the skin folds on the neck, elbows, and dorsal interphalangeal spaces. Acanthosis reflects the severity of underlying insulin resistance and diminishes with weight loss. Friability of skin may be increased, especially in skin folds, enhancing the risk of fungal and yeast infections. Finally, venous stasis is increased in the obese.

TREATMENT

Obesity is a chronic medical condition. Successful treatment, defined as the sustained attainment of normal body weight without producing unacceptable treatment-induced morbidity, is rarely achieved in clinical practice. Many approaches produce short-term weight loss, and this has clear benefits for associated morbidities such as hypertension and diabetes. Despite the fact that sustained weight loss is uncommon, enormous resources are expended in pursuit of this goal.

Treatment goals should be guided by the health risks of obesity in any given individual (Fig. 77-6). The clinician should always consider the possibility that an individual has an identified cause of obesity, such as hypothyroidism, hypercortisolism, male hypogonadism, insulinoma, or central nervous system disease that affects hypothalamic function. Although they are infrequent causes of obesity, specific therapy may be available.

Behavior Modification The principles of behavior modification provide the underpinnings for many current programs of weight reduction. Typically, the patient is requested to monitor and record the circumstances related to eating, and rewards are designed to modify maladaptive behaviors. Patients may benefit from counseling offered in a stable group setting for extended periods of time, including after weight loss.

Diet Reduced caloric intake is the cornerstone of obesity treatment. The fundamental goal is the sustained reduction of energy intake below that of energy expenditure. The difficulty in achieving this goal has led to a wide array of suggested diets that vary in recommended calorie content (from total fasting to mild reductions), as well as specific food content and form (e.g., liquid vs. solid). There is no scientific evidence to validate the utility of specific "fad diets." The main diet regimens in use follow several general facts relevant to food intake and weight loss. First, a deficit of 7500 kcal will produce a weight loss of approximately 1 kg. Therefore, eating 100 kcal/d less for a year should cause a 5-kg weight loss, and a deficit of 1000 kcal/d should cause a loss of approximately 1 kg per week. The rate of weight loss on a given caloric intake is related to the rate of energy expenditure. Because obese individuals have a higher metabolic rate than lean individuals, and because men have a higher metabolic rate than women (due to their greater lean body mass), the rate of weight loss is greater among the more obese and among men (relative to women). With chronic caloric restriction, metabolic rate diminishes, but because of reduced lean body mass (along with much greater loss in fat mass) and possibly because of other adaptations. This fall in metabolic rate with food restriction slows the rate of weight loss on a constant diet. With total starvation or diets restricted to<600 kcal/d, initial weight loss over the first week results predominantly from natriuresis and the loss of fluids.

Very low energy diets (e.g., 400 to 600 kcal/d) are widely used. The liquid protein diets popularized in the 1970s were proved to be unsafe, causing >60 deaths. Life-threatening arrhythmias were documented in the clinical research setting, a consequence of both low-quality protein and deficiencies of vitamins, minerals, and trace elements. These types of diets have now been substantially modified. A very low energy diet consisting of 45 to 70 g high-quality protein, 30 to 50 g carbohydrate, and approximately 2 g fat per day, as well as supplements of vitamins, minerals, and trace elements, appears to be safe in selected patients under medical supervision. Patients should not be started on such diets unless they are >130% of their ideal body weight. Contraindications include pregnancy, cancer, recent myocardial infarction, cerebrovascular disease, hepatic disease, or untreated psychiatric disease. When used in patients with diabetes who are receiving insulin or oral agent therapy, close supervision is required and diabetic treatment may need to be adjusted. Whenever possible, exercise regimens and behavioral modification approaches should be used in conjunction with the diet.

Advantages of very low calorie diets are the greater rate of weight loss compared to less restrictive diets, as well as the possible beneficial effect of hunger suppression brought about by the production of ketones. In patients on such diets, blood pressure, blood glucose, cholesterol, and triglyceride levels fall, and pulmonary function and exercise tolerance improve. Sleep apnea may improve within a few weeks. Complications of these very low energy diets are usually minor and include fatigue, constipation or diarrhea, dry skin, hair loss, menstrual irregularities, orthostatic dizziness, and difficulty concentrating. Cholelithiasis and pancreatitis may occur when such diets are interrupted by binge eating; gallstones have been shown to develop in as many as 25% of patients while on the diet.

Low-calorie diets,>800 kcal/d, are applicable to most patients and have fewer restrictions than the very low calorie diets. Considerable controversy has attended the question of which diet composition is most appropriate for promoting weight loss. Though commonly recommended, benefits resulting from very low fat diets are modest at best. Nonetheless, the health effects of low-fat diets -- apart from curbing obesity -- may be important. A diet rich in fruits, vegetables, and whole grains may promote weight loss and is preferable to low-fat diets in which large amounts of simple carbohydrates are substituted for fats. The latter may actually promote obesity. Some have advocated diets with protein replacement of simple carbohydrates in an effort to minimize insulin production. The efficacy of this strategy, aside from overall calorie reduction, is unknown.

Exercise Exercise is an important component of the overall approach to treating obesity. Increased energy expenditure is the most obvious mechanism for an effect of exercise. The impact of an exercise regimen as a sole therapy of obesity has been difficult to document. On the other hand, exercise appears to be a valuable means to sustain diet therapy (Fig. 77-7). Even if exercise had no such salutary effect, it would be valuable in the obese individual for its effects on cardiovascular tone and blood pressure. Because many obese individuals have not engaged in exercise on a regular basis and may have cardiovascular risk factors, it should be introduced gradually and under medical supervision, especially in the most obese individuals.

Drugs Unfortunately, drug treatment of obesity is rarely efficacious. Despite short-term benefits, medication-induced weight loss is often associated with rebound weight gain after the cessation of drug use, side effects from the medications, and the potential for drug abuse. Given the need for effective therapies, many possible compounds have been evaluated. On the basis of placebo-controlled trials, the U.S. Food and Drug Administration (FDA) approved several amphetamine-like agents for short-term use. Phentermine is an amphetamine-like drug with low addictive potential that has shown modest efficacy (10 vs. 4.4 kg of weight loss over a 24-week period in well-controlled study). This class of drugs is thought to act centrally by reducing appetite. Effects on energy expenditure are less clear. Over-the-counter drugs, such as phenylpropanolamine HCl, have similar efficacy to prescription appetite suppressants in short-term studies. Drugs that promote serotonin release or inhibit serotonin reuptake, such as fenfluramine, also have modest efficacy. When fenfluramine was administered together with phentermine, as "fen-phen," the combination was widely used based on controlled trials that demonstrated modest but definite efficacy. However, the risk of primary pulmonary hypertension was increased up to 20-fold in association with this treatment. The FDA withdrew approval of the fen-phen combination in 1997 when reports suggested an association with right- and left-sided valvular heart disease. The histopathologic features of the valvular disease are similar to those seen in carcinoid syndrome and are thought to result from fenfluramine. Though the true incidence and long-term effects of these valvular lesions are currently unknown, the occurrence of this complication has been verified in multiple studies.

Sibutramine is a novel central reuptake inhibitor of both norepinephrine and serotonin. Using a once-daily dose over 24 weeks, it produced a 7% weight loss in a double-blind, placebo-controlled trial. It lowered cholesterol and triglyceride levels and exhibited similar clinical efficacy to fenfluramine. Sibutramine increases pulse and blood pressure in some patients, and long-term safety is not established. Orlistat is an inhibitor of intestinal lipase that causes modest weight loss due to drug-induced fat malabsorption. A randomized, double-blind trial over 2 years revealed modest weight loss (8.7 kg for 120 mg orlistat versus 5.8 kg from diet alone) during the first year and better maintenance of weight loss in a second year compared to the placebo-treated group (3.2 kg regained versus 5.6 kg regained for placebo).LDLcholesterol and insulin levels were also reduced. In patients with obesity and type 2 diabetes mellitus, the antidiabetic medication metformin tends to decrease body weight. The mechanism appears to involve inhibition of appetite. Thyroid hormone has little place in the treatment of obesity, as the vast majority of obese individuals are euthyroid. It promotes loss of lean body mass and raises the risk of complications from the hyperthyroid state.

b_3-Adrenergic receptor agonists may provide a new treatment approach for obesity. Drugs of this class are in clinical trials. In animals,$b_3$agonists promote leanness by stimulating thermogenesis inBAT; they also stimulate lipolysis in white adipose tissue. These drugs also reduce insulin resistance and lower blood glucose in animal models by a mechanism that is not yet defined. Recombinant human leptin is also in clinical trials. In the rare cases of leptin deficiency caused by mutations of the leptin gene, the administration of recombinant leptin is highly effective. Preliminary reports suggest that the response to leptin is limited or absent in common causes of obesity (which are associated with hyperleptinemia and leptin resistance). New drugs are also being developed based on insights into central pathways that regulate body weight. These include antagonists forNPYreceptors (subtypes Y1, Y5) and agonists for melanocortin 4 receptors.

Surgery Morbid obesity, commonly defined as either 45 kg (100 lb) or 100% above ideal body weight, is estimated to increase mortality by as much as twelvefold in men between 25 and 34 years of age and sixfold between 35 and 45 years of age. Deaths from cardiovascular disease, diabetes, and accidents have been documented. In response to ineffective treatment using diet, exercise, and available drugs, surgical approaches have been tried. The potential benefits of surgery include major weight loss and improvement in hypertension, diabetes, sleep apnea,CHF, angina, hyperlipidemia, and venous disease. Many different approaches have been used, often without adequate assessment of efficacy and complications. Jejunoileal bypass surgery has largely been abandoned because of complications, which have included electrolyte disturbances, nephrolithiasis, gallstones, gastric ulcers, arthritis, and hepatic dysfunction, with cirrhosis occurring in as many as 7% of patients. Two procedures in common use today are the vertical-banded gastroplasty and the Roux-en-Y gastric bypass (Fig. 77-8).

Following the National Institutes of Health Consensus Conference on Gastrointestinal Surgery for Severe Obesity in 1991, it was recommended that suitable patients be selected using the following criteria: (1) the presence of 45 kg (100 lb) or 100% above ideal body weight, or one or more severe medical conditions related to refractory obesity; (2) repeated failures of other therapeutic approaches; (3) at eligible weight for 3 to 5 years; (4) capability of tolerating surgery; (5) absence of alcoholism, other addictions, or major psychopathology; and (6) prior clearance by a psychiatrist. It is recommended that an appropriately experienced surgeon work together with nutritionists and other support personnel; evaluation and follow-up programs should be monitored closely.

ACKNOWLEDGEMENT

The author acknowledges the contributions of Dr. George A. Bray, who wrote this chapter in the 14th edition.

(Bibliography omitted in Palm version)

Back to Table of Contents

78. EATING DISORDERS - *B. Timothy Walsh*

Anorexia nervosa and bulimia nervosa are characterized by severe disturbances of eating behavior. The salient feature of *anorexia nervosa* is a refusal to maintain a minimally normal body weight. *Bulimia nervosa* is characterized by recurrent episodes of binge eating followed by abnormal compensatory behaviors, such as self-induced vomiting. Anorexia nervosa and bulimia nervosa are closely related. Both occur primarily among previously healthy young women who become overly concerned with body shape and weight. Many patients with bulimia nervosa have past histories of anorexia nervosa, and many patients with anorexia nervosa engage in binge eating and purging behavior. In the current diagnostic system, the critical distinction between anorexia nervosa and bulimia nervosa depends on body weight: patients with anorexia nervosa are, by definition, significantly underweight, whereas the weights of patients with bulimia nervosa are in the normal range or above.

Another syndrome of disturbed eating behavior has been described recently: *Binge eating disorder* is characterized by repeated episodes of binge eating, similar to those of bulimia nervosa, in the absence of inappropriate compensatory behavior. Patients with binge eating disorder are typically middle-aged men or women with significant obesity. They have an increased frequency of anxiety and depression compared to similarly obese patients without binge eating disorder. It is not known whether patients with binge eating disorder are at increased risk for medical complications or what treatments are most useful.

EPIDEMIOLOGY

In women, the full syndrome of anorexia nervosa occurs with a lifetime prevalence of approximately 0.5%; bulimia nervosa occurs with a lifetime prevalence of 1 to 3%. Variants of these eating disorders with only some features of anorexia nervosa or bulimia nervosa are much more common and occur in 5 to 10% of young women. Both anorexia nervosa and bulimia nervosa also occur in males but at frequencies approximately one-tenth of those in females.

Anorexia nervosa and bulimia nervosa are more prevalent in cultures where food is plentiful and in which being thin is associated with attractiveness. These disorders are more frequent among young women who place a premium on thinness, such as ballet dancers and models. The incidence of anorexia nervosa appears to have increased in recent decades. The frequency of bulimia nervosa increased dramatically in the early 1970s and 1980s but may have declined somewhat in recent years.

DIAGNOSIS

The diagnosis of eating disorders is based on the presence of characteristic behavioral, psychological, and physical attributes (Tables 78-1 and78-2). Widely accepted diagnostic criteria are provided by the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV).

ANOREXIA NERVOSA

For anorexia nervosa, these criteria include weight <85% of the expected weight for age and height, which is roughly equivalent to a body mass index (BMI) of 18.5 kg/m^{2}for adult women. This weight criterion is somewhat arbitrary, so that a patient who meets all other diagnostic criteria but weighs between 85 and 90% of expected would still merit the diagnosis of anorexia nervosa. Despite being underweight, patients with anorexia nervosa are irrationally afraid of gaining weight, often out of a concern that weight gain will get "out of control." They also exhibit a distortion of body image (criterion 3,Table 78-1), which may express itself in several ways. For example, despite being emaciated, patients with anorexia nervosa may believe that their body as a whole, or some part of their body, is too fat and experience additional weight loss as a highly rewarding achievement. The current diagnostic criteria require that women with anorexia nervosa not have spontaneous menses, but occasional patients with the characteristics and complications of anorexia nervosa describe regular menses. Two mutually exclusive subtypes of anorexia nervosa are specified inDSM-IV. Patients whose weight loss is maintained primarily by caloric restriction, perhaps augmented by excessive exercise, are considered to have the "restricting" subtype of anorexia nervosa. The "binge eating/purging" subtype is characterized by self-induced vomiting or laxative abuse. Patients with the binge/purge subtype are more prone to develop electrolyte imbalances, are more emotionally labile, and are more likely to have other problems with impulse control, such as drug abuse.

The diagnosis of anorexia nervosa can usually be made confidently on the basis of history when significant weight loss is accomplished by restrictive dieting and excessive exercise and is accompanied by a marked reluctance to gain weight. Patients with anorexia nervosa often deny that they have a serious problem and may be brought to medical attention by concerned family or friends. In atypical presentations, other causes of significant weight loss in previously healthy young people should be considered, including inflammatory bowel disease, gastric outlet obstruction, central nervous system (CNS) tumors, and neoplasm (Chap. 43).

BULIMIA NERVOSA

The critical diagnostic features of bulimia nervosa are repeated episodes of binge eating followed by inappropriate and abnormal behaviors aimed at avoiding weight gain. During binges, patients with this disorder tend to consume large amounts of sweet foods with a high fat content, such as dessert items. The most frequent compensatory behaviors are self-induced vomiting and laxative abuse, but a wide variety of techniques have been described, including the omission of insulin injections by diabetics. Typically, patients with bulimia nervosa are ashamed of their behavior and endeavor to keep their disorder hidden from family and friends. Like patients with anorexia nervosa, those with bulimia nervosa place an unusual emphasis on weight and shape as a basis for their self-esteem.

As in anorexia nervosa, there are two mutually exclusive subtypes of bulimia nervosa. Patients with the "purging" subtype utilize compensatory behaviors that directly rid the body of calories or fluids (e.g., self-induced vomiting, laxative or diuretic abuse), whereas those with the "nonpurging" subtype attempt to compensate for binges by fasting or by excessive exercise. Patients with the nonpurging subtype tend to be heavier and are less prone to fluid and electrolyte disturbances.

The diagnosis of bulimia nervosa requires a candid history from the patient detailing recurrent, large eating binges followed by the purposeful use of inappropriate mechanisms to avoid weight gain. Most patients with bulimia nervosa who present for treatment are distressed by their inability to control their eating behavior but are able to provide such details if queried in a supportive and nonjudgmental fashion.

ETIOLOGY

The fundamental etiology of the eating disorders is unknown but is believed to involve a combination of psychological, biologic, and cultural risk factors. Many of these risk factors, such as sexual or physical abuse and a family history of mood disturbance or substance abuse, are best viewed as nonspecific risk factors that increase vulnerability to a range of psychiatric disorders. Other factors appear to be more specific to the development of an eating disorder.

Patients who develop anorexia nervosa are inclined to be more obsessional and perfectionist than their peers. The disorder often begins as a diet not distinguishable at the outset from those undertaken by many adolescents and young women. As weight loss progresses, the fear of gaining weight grows; dieting becomes stricter; and psychological, behavioral, and medical aberrations increase. The fact that most cases are reported from countries where food is plentiful and where thinness, especially among women, is highly valued suggests that cultural factors play a significant role in the development of anorexia nervosa. However, it is notable that the clinical syndrome was well described over a century ago, when the cultural pressures were quite different.

Numerous physiologic disturbances, including abnormalities in a variety of neurotransmitter systems, have been described in anorexia nervosa (see below). It is difficult to distinguish neurochemical, metabolic, and hormonal changes that may have a role in the initiation or perpetuation of the syndrome from those that are secondary to the disorder. The resolution of most of these abnormalities with weight restoration argues against their having a critical etiologic role.

Bulimia nervosa typically begins during or following an episode of dieting, often in association with depressed mood. Patients who develop bulimia nervosa describe a higher-than-expected prevalence of childhood and parental obesity, suggesting that a predisposition towards obesity may increase vulnerability to this eating disorder. The marked increase in the number of cases of bulimia nervosa during the past 25 years and the rarity of bulimia nervosa in underdeveloped countries suggest that cultural factors are important. Several biologic abnormalities in patients with bulimia nervosa may perpetuate this disorder once it has begun. These include abnormalities ofCNSserotonergic function, which is involved in the regulation of eating behavior, and disruption of peripheral satiety mechanisms, including the release of cholecystokinin (CCK) from the small intestine.

Genetic factors probably contribute to the risk of development of eating disorders, as the incidence of these disorders is greater in families with one affected member and the concordance in monozygotic twins is greater than in dizygotic twins. However, specific genes have not been identified, and the range of the estimates of heritability is large.

CLINICAL FEATURES (Table 78-3)

ANOREXIA NERVOSA

Anorexia nervosa typically begins in mid to late adolescence, sometimes in association with a stressful life event such as leaving home for school. The disorder occasionally develops in early puberty, before menarche, but seldom begins after age 40. Despite being underweight, patients with anorexia nervosa rarely complain of hunger or fatigue and often exercise extensively. Further weight loss is viewed by the patient as a fulfilling accomplishment, while weight gain is seen as a personal failure. Patients tend to become socially withdrawn and increasingly committed to work or study, dieting, and exercise. As weight loss progresses, thoughts of food dominate mental life and idiosyncratic rules develop around eating. Patients with anorexia nervosa may obsessively collect cookbooks and recipes and be drawn to food-related occupations. Despite the denial of hunger, one-quarter to one-half of patients with anorexia nervosa engage in eating binges.

Physical Features Patients with anorexia nervosa typically have few physical complaints but may note cold intolerance and constipation. Some women who develop anorexia nervosa after menarche report that their menses ceased before significant weight loss occurred. Weight and height should be measured to allow calculation ofBMI(kg/m2). Vital signs may reveal bradycardia, hypotension, and hypothermia. Soft, downy hair growth (lanugo) sometimes occurs, especially on the back, and alopecia may be seen. Salivary gland enlargement, which is associated with starvation as well as with binge eating and vomiting, may make the face appear surprisingly full in contrast to the marked general wasting. Acrocyanosis of the digits is common, and peripheral edema can be seen in the absence of hypoalbuminemia, particularly when the patient begins to regain weight. Some patients who consume large amounts of vegetables containing vitamin A develop a yellow tint to the skin (*hypercarotenemia*), which is especially notable on the palms.

Laboratory Abnormalities Mild normochromic, normocytic anemia is frequent, as is mild to moderate leukopenia, with a disproportionate reduction of polymorphonuclear leukocytes. Dehydration may result in slightly increased levels of blood urea nitrogen and creatinine. Serum liver enzyme levels may increase, especially during the early phases of refeeding. The level of serum proteins is usually normal. Blood sugar is often low and serum cholesterol may be moderately elevated. Gastrointestinal motility is diminished, leading to reduced gastric emptying and constipation. A range of electrolyte disturbances may develop, reflecting the degree to which the patient restricts or overconsumes fluids and whether the patient engages in purging behavior. Hypokalemic alkalosis suggests self-induced vomiting or the use of diuretics. Hyponatremia is common and may result from excess fluid intake and disturbances in the secretion of antidiuretic hormone.

Endocrine Abnormalities The regulation of virtually every endocrine system is altered in anorexia nervosa, but the most striking changes occur in the reproductive system. Amenorrhea is hypothalamic in origin and reflects diminished production of gonadotropin-releasing hormone (GnRH). When exogenous GnRH is administered in a physiologic pulsatile manner, pituitary responses of luteinizing hormone (LH) and follicle stimulating hormone (FSH) are normalized, indicating the absence of a primary pituitary abnormality. The resulting gonadotropin deficiency causes low plasma estrogen in women and reduced testosterone in men. The hypothalamic GnRH pulse generator is exquisitely sensitive, particularly in women, to body weight, stress, and exercise, each of which may contribute to *hypothalamic amenorrhea* in anorexia nervosa (Chap. 336). Although the mechanisms underlying these effects are unknown, the decreased adipose tissue associated with weight loss leads to a marked reduction in leptin, a hormone that plays a permissive role in GnRH production (Chap. 77). In many patients, weight gain to a specific threshold triggers restoration of the GnRH pulse generator, initially recapitulating the pubertal pattern of nocturnal gonadotropin secretion before returning to the normal adult pattern.

Serum cortisol and 24-h urine free cortisol levels are generally elevated but without characteristic clinical signs of cortisol excess. Thyroid function tests resemble the pattern seen in euthyroid sick syndrome (Chap. 330). Thyroxine (T_4) and free T_4levels are usually in the low-normal range, triiodothyronine (T_3) levels are reduced, and reverse T_3(rT_3) is elevated. The level of thyroid stimulating hormone (TSH) is normally or partially suppressed. Growth hormone is increased, but insulin-like growth factor 1 (IGF-1), which is produced mainly by the liver, is reduced, as it is in other conditions of starvation. Diminished bone density is routinely observed in anorexia nervosa and reflects the effects of multiple nutritional deficiencies, reduced gonadal steroids, and increased cortisol. The degree of bone density reduction is proportional to the length of the illness, and patients are at risk for the development of symptomatic fractures. The occurrence of anorexia nervosa during adolescence may lead to the premature cessation of linear bone growth and a failure to achieve expected adult height.

Cardiac Abnormalities Cardiac output is reduced, and congestive heart failure occasionally occurs during rapid refeeding. The electrocardiogram usually shows sinus bradycardia, reduced QRS voltage, and nonspecific ST-T-wave abnormalities. Some patients develop a prolonged QT_cinterval, which may predispose to serious arrhythmias.

BULIMIA NERVOSA

The typical patient presenting for treatment of bulimia nervosa is a woman of normal weight in her mid-twenties who reports binge eating and purging 5 to 10 times a week for 5 to 10 years. The disorder usually begins in late adolescence or early adulthood during or following a diet. The self-imposed caloric restriction leads to increased hunger and to overeating. In an attempt to avoid weight gain, the patient induces vomiting, takes laxatives or diuretics, or engages in some other form of compensatory behavior. Initially, patients may experience a sense of satisfaction that appealing food can be eaten without weight gain. However, as the disorder progresses, patients perceive diminished control over eating. Binges increase in size and frequency and are provoked by a variety of stimuli, such as transient depression, anxiety, or a sense that too much food has been consumed in a normal meal. Between binges, patients attempt to restrict caloric intake, which increases hunger and sets the stage for the next binge.

Although vomiting may be triggered initially by manual stimulation of the gag reflex, most patients with bulimia nervosa develop the ability to induce vomiting at will. Rarely, patients resort to the regular use of syrup of ipecac. Laxatives and diuretics are frequently taken in impressive quantities, such as 30 or 60 laxative pills on a single occasion. The resulting fluid loss produces dehydration and a feeling of emptiness but has little impact on caloric balance.

The physical abnormalities associated with bulimia nervosa primarily result from the purging behavior. Painless bilateral salivary gland hypertrophy (sialadenosis) may be noted. A scar or callus on the dorsum of the hand may develop due to repeated trauma from the teeth among patients who manually stimulate the gag reflex. Recurrent vomiting and the exposure of the lingual surfaces of the teeth to stomach acid leads to loss of dental enamel and eventually to chipping and erosion of the front teeth. Laboratory abnormalities are surprisingly infrequent, but hypokalemia, hypochloremia, and hyponatremia are observed occasionally. Repeated vomiting may lead to alkalosis, whereas repeated laxative abuse may produce a mild metabolic acidosis. Serum amylase may be mildly elevated due to an increase in the salivary isoenzyme.

Serious physical complications resulting from bulimia nervosa are rare. Oligomenorrhea and amenorrhea are more frequent than in women without eating disorders. Arrhythmias occasionally occur secondary to electrolyte disturbances. Tearing of the esophagus and rupture of the stomach, which constitute life-threatening events, have been reported. Some patients who have chronically abused laxatives or diuretics develop transient peripheral edema when this behavior ceases, presumably due to high levels of aldosterone resulting from persistent fluid and electrolyte depletion.

PROGNOSIS

The course and outcome of anorexia nervosa are highly variable. One-quarter to one-half of patients eventually recover fully, with few psychological or physical sequelae. However, many patients have persistent difficulties with weight maintenance, depression, and eating disturbances, including bulimia nervosa. The development of obesity following anorexia nervosa is rare. The long-term mortality of anorexia nervosa is among the highest associated with any psychiatric disorder. Approximately 5% of patients die per decade of follow-up, primarily due to the physical effects of chronic starvation or by suicide.

Virtually all of the physiologic abnormalities associated with anorexia nervosa are observed in other forms of starvation and markedly improve or disappear with weight gain. A worrisome exception is the reduction in bone mass, which may not recover fully, particularly when anorexia nervosa occurs during adolescence when peak bone mass is normally achieved.

The prognosis of bulimia nervosa is much more favorable. Mortality is low, and full recovery occurs in approximately 50% of patients within 10 years. Approximately 25% of patients have persistent symptoms of bulimia nervosa over many years. Few patients progress from bulimia nervosa to anorexia nervosa.

TREATMENT

Anorexia Nervosa Because of the profound physiological and psychological effects of starvation, there is a broad consensus that weight restoration to 90% of predicted weight is the primary goal in the treatment of anorexia nervosa. Unfortunately, because most patients resist this goal, its accomplishment is often accompanied by frustration for the patient, the family, and the physician. In attempting to engage the patient in treatment, it may be useful for the physician to elicit the patient's physical concerns (e.g., about osteoporosis, weakness, or fertility) and, if possible, educate the patient regarding the importance of normalizing nutritional status in order to address those concerns. The physician should attempt to reassure the patient that weight gain will not be permitted to get "out of control" but simultaneously emphasize that weight restoration is medically and psychologically imperative.

The intensity of the initial treatment, including the need for hospitalization, is determined by the patient's current weight, the rapidity of recent weight loss, and the severity of medical and psychological complications (Fig. 78-1). Hospitalization should be strongly considered for patients weighing <75% of expected, even if the results of routine blood studies are within normal limits. Acute medical problems, such as severe electrolyte imbalances, should be identified and addressed. Nutritional restoration can almost always be successfully accomplished by oral feeding, and parenteral methods are rarely required. For severely underweight patients, sufficient calories (approximately 1500 to 1800 kcal/d) should be provided initially in divided meals as food or liquid supplements to maintain weight and to permit stabilization of fluid and electrolyte balance. Calories can then be gradually increased to achieve a weight gain of 1 to 2 kg (2 to 4 lb) per week, typically requiring an intake of 3000 to 4000 kcal/d. Meals must be supervised, ideally by personnel who are firm regarding the necessity of food consumption, empathic regarding the challenges entailed, and reassuring regarding the patient's eventual recovery. Patients have great psychological difficulty complying with the need for increased caloric consumption, and the assistance of psychiatrists or psychologists experienced in the treatment of anorexia nervosa is usually necessary.

Psychiatric treatment focuses primarily on two issues. First, patients require much emotional support during the period of weight gain. Second, patients must learn to base their self-esteem, not on the achievement of an inappropriately low weight, but on the development of satisfying personal relationships and the attainment of reasonable academic and occupational goals. For younger patients, the active involvement of the family in treatment is crucial.

Less severely affected patients may be treated in a partial hospitalization program where medical and psychiatric supervision is available and several meals can be monitored each day. Outpatient treatment may suffice for mildly ill patients. Weight must be monitored at frequent intervals, and explicit goals agreed on for weight gain, with the understanding that more intensive treatment will be required if the level of care initially employed is not successful.

Medical complications occasionally occur during refeeding. Most patients transiently retain excess fluid, occasionally resulting in peripheral edema. Fluid retention occurs during recovery from other forms of malnutrition and generally does not require specific treatment in the absence of cardiac, renal, or hepatic dysfunction. Congestive heart failure and acute gastric dilatation have been described when refeeding has been rapid. Transient modest elevations in serum levels of liver enzymes occasionally occur. Low levels of magnesium and phosphate should be repleted. Multivitamins should be given, and it is important to ensure adequate intake of vitamin D (400 IU/d) and calcium (1500 mg/d) to minimize bone loss.

No psychotropic medications are of established value in the treatment of anorexia nervosa; tricyclic antidepressants are contraindicated when there is prolongation of the QT_cinterval. The alterations of cortisol and thyroid hormone metabolism do not require specific treatment and are corrected by weight gain. Estrogen treatment appears to have minimal impact on bone density in underweight patients but may be helpful to relieve symptoms of estrogen deficiency.

Bulimia Nervosa Bulimia nervosa can usually be treated on an outpatient basis. Cognitive behavioral therapy (CBT) is a short-term (4 to 6 months) psychological treatment that focuses on the intense concern with shape and weight, the persistent dieting, and the binge eating and purging that characterize this disorder. Patients are directed to monitor the circumstances, thoughts, and emotions associated with binge/purge episodes, to eat regularly, and to challenge their assumptions linking weight to self-esteem. CBT produces symptomatic remission in 25 to 50% of patients.

Numerous double-bind, placebo-controlled trials have documented that antidepressant medications are useful in the treatment of bulimia nervosa but are probably somewhat less effective thanCBT. Although efficacy has been established for virtually all chemical classes of antidepressants, only the selective serotonin reuptake inhibitor fluoxetine (Prozac) has been approved for use in bulimia nervosa by the U.S. Food and Drug Administration. Antidepressant medications are helpful even for patients with bulimia nervosa who are not depressed, and the dose of fluoxetine recommended for bulimia nervosa (60 mg/d) is higher than that typically used to treat depression. These observations suggest that different mechanisms may underlie the utility of these medications in bulimia nervosa and in depression.

A substantial minority of patients with bulimia nervosa do not respond adequately toCBT, antidepressant medication, or their combination. More intensive forms of treatment, including hospitalization, may be required for such patients.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART SIX -ONCOLOGY AND HEMATOLOGY

SECTION 1 -NEOPLASTIC DISORDERS

79. APPROACH TO THE PATIENT WITH CANCER - *Dan L. Longo*

The application of current treatment techniques (surgery, radiation therapy, chemotherapy, and biological therapy) results in the cure of>50% of patients diagnosed with cancer. Nevertheless, patients experience the diagnosis of cancer as one of the most traumatic and revolutionary events that has ever happened to them. Independent of prognosis, the diagnosis brings with it a change in a person's self-image and in his or her role in the home and workplace. The prognosis of a person who has just been found to have pancreatic cancer is the same as the prognosis of the person with aortic stenosis who develops the first symptoms of congestive heart failure (median survival, about 8 months). However, the patient with heart disease may remain functional and maintain a view of him- or herself as a fully intact person with just a malfunctioning part, a diseased organ ("a bum ticker"). By contrast, the patient with pancreatic cancer has a completely altered self-image and is viewed differently by family and anyone who knows the diagnosis. He or she is being attacked and invaded by a disease that could be anywhere in the body. Every ache or pain takes on desperate significance. Cancer is an exception to the coordinated interaction among cells and organs. In general, the cells of a multicellular organism are programmed for collaboration. Many diseases occur because the specialized cells fail to perform their assigned task. Cancer takes this malfunction one step further. Not only is there a failure of the cancer cell to maintain its specialized function, but it also strikes out on its own; the cancer cell competes to survive using natural mutability and natural selection to seek advantage over normal cells in a recapitulation of evolution. One consequence of the traitorous behavior of cancer cells is that the patient feels betrayed by his or her body. The cancer patient feels that he or she, and not just a body part, is diseased.

THE MAGNITUDE OF THE PROBLEM

There is no nationwide cancer registry; therefore, the incidence of cancer is estimated on the basis of the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) database, which tabulates cancer incidence and death figures from nine sites, accounting for about 10% of the U.S. population, and from population data from the Bureau of the Census. In 2000, 1.22 million new cases of invasive cancer (619,700 men, 600,400 women) were diagnosed and 552,200 people (284,100 men, 268,100 women) died from cancer. The percent distribution of new cancer cases and cancer deaths by site for men and women are shown inTable 79-1. Cancer incidence has been declining by about 2% each year since 1992.

The most significant risk factor for cancer overall is age; two-thirds of all cases were in people over age 65. Cancer incidence increases as the third, fourth, or fifth power of age in different sites. For the interval between birth and age 39, 1 in 62 men and 1 in 52 women will develop cancer; for the interval between ages 40 and 59, 1 in 12 men and 1 in 11 women will develop cancer; and for the interval between ages 60 and 79, 1 in 3 men and 1 in 4 women will develop cancer.

Cancer is the second leading cause of death behind heart disease. Deaths from heart disease have declined 45% in the United States since 1950 and continue to decline. After a 70-year period of increases, cancer deaths began to decline in 1997 (Fig. 79-1). The five leading causes of cancer deaths are shown for various populations inTable 79-2. Along with the decrease in incidence has come an increase in survival for cancer patients. The 5-year survival for white patients was 39% in 1960-1963 and 61% in 1989-1995. Cancers are more often deadly in blacks; the 5-year survival was 48% for the 1989-1995 interval. Incidence and mortality vary among racial and ethnic groups (Table 79-3). The basis for these differences is unclear.

PATIENT MANAGEMENT

Important information is obtained from every portion of the routine history and physical examination. The duration of symptoms may reveal the chronicity of disease. The past medical history may alert the physician to the presence of underlying diseases that may affect the choice of therapy or the side effects of treatment. The social history may reveal occupational exposure to carcinogens or habits, such as smoking or alcohol consumption, that may influence the course of disease and its treatment. The family history may suggest an underlying familial cancer predisposition and point out the need to begin surveillance or other preventive therapy for unaffected siblings of the patient. The review of systems may suggest early symptoms of metastatic disease or a paraneoplastic syndrome.

DIAGNOSIS

The diagnosis of cancer relies most heavily on invasive tissue biopsy and should never be made without obtaining tissue; no noninvasive diagnostic test is sufficient to define a disease process as cancer. Although in rare clinical settings (e.g., thyroid nodules) fine-needle aspiration is an acceptable diagnostic procedure, the diagnosis generally depends on obtaining adequate tissue to permit careful evaluation of the histology of the tumor, its grade, and its invasiveness and to yield further molecular diagnostic information, such as the expression of cell-surface markers or intracellular proteins that typify a particular cancer, or the presence of a molecular marker, such as the t(8;14) translocation of Burkitt's lymphoma. Increasing evidence links the expression of certain genes with the prognosis and response to therapy (Chaps. 81 and82).

Occasionally a patient will present with a metastatic disease process that is defined as cancer on biopsy but has no apparent primary site of disease. Efforts should be made to define the primary site based on age, sex, sites of involvement, histology and tumor markers, and personal and family history. Particular attention should be focused on ruling out the most treatable causes (Chap. 99).

Once the diagnosis of cancer is made, the management of the patient is best undertaken as a multidisciplinary collaboration among the primary care physician, medical oncologists, surgical oncologists, radiation oncologists, oncology nurse specialists, pharmacists, social workers, rehabilitation medicine specialists, and a number of other consulting professionals working closely with each other and with the patient and family.

DEFINING THE EXTENT OF DISEASE AND THE PROGNOSIS

The first priority in patient management after the diagnosis of cancer is established and shared with the patient is to determine the extent of disease. The curability of a tumor usually is inversely proportional to the tumor burden. Ideally, the tumor will be diagnosed before symptoms develop or as a consequence of screening efforts (Chap. 80). A very high proportion of such patients can be cured. However, most patients with cancer present with symptoms related to the cancer, caused either by mass effects of the tumor or by alterations associated with the production of cytokines or hormones by the tumor.

For most cancers, the extent of disease is evaluated by a variety of noninvasive and invasive diagnostic tests and procedures. This process is called *staging*. There are two types. *Clinical staging* is based on physical examination, radiographs, isotopic scans, computed tomography, and other imaging procedures; *pathologic staging* takes into account information obtained during a surgical procedure, which might include intraoperative palpation, resection of regional lymph nodes and/or tissue adjacent to the tumor, and inspection and biopsy of organs commonly involved in disease spread. Pathologic staging includes histologic examination of all tissues removed during the surgical procedure. Surgical procedures performed may include a simple lymph node biopsy or more extensive procedures such as thoracotomy, mediastinoscopy, or laparotomy. Surgical staging may occur in a separate procedure or may be done at the time of definitive surgical resection of the primary tumor.

Knowledge of the predilection of particular tumors for spread to adjacent or distant organs helps direct the staging evaluation.

Information obtained from staging is used to define the extent of disease either as localized, as exhibiting spread outside of the organ of origin to regional but not distant sites, or as metastatic to distant sites. The most widely used system of staging is the TNM (tumor, node, metastasis) system codified by the International Union Against Cancer and the American Joint Committee on Cancer (AJCC).

[1] The AJCC *Manual for Staging Cancer*, 5th edition, can be obtained from the AJCC at 55 East Erie Street, Chicago, Il, 60611.

The TNM classification is an anatomically based system that categorizes the tumor on the basis of the size of the primary tumor lesion (T1-4, where a higher number indicates a tumor of larger size), the presence of nodal involvement (usually N0 and N1 for the absence and presence, respectively, of involved nodes, although some tumors have more elaborate systems of nodal grading), and the presence of metastatic disease (M0 and M1 for the absence and presence, respectively, of metastases). The various permutations of T, N, and M scores are then broken into stages, usually designated by the roman numerals I through IV. Tumor burden increases and curability decreases with increasing stage. Other anatomic staging systems are used for some tumors, e.g., the Dukes classification for colorectal cancers, the International Federation of Gynecologists and Obstetricians (FIGO) classification for gynecologic cancers, and the Ann Arbor classification for Hodgkin's disease.

Certain tumors cannot be grouped appropriately on the basis of anatomic considerations. For example, hematopoietic tumors such as leukemia, myeloma, and lymphoma are often disseminated at presentation and do not spread in the fashion typical of solid tumors. For these tumors, other prognostic factors have been identified (Chaps. 111,112, and113).

In addition to tumor burden, a second major determinant of treatment outcome is the physiologic reserve of the patient. Patients who are bedridden before developing cancer are likely to fare worse, stage for stage, than fully active patients. Physiologic reserve is a determinant of how a patient is likely to cope with the physiologic stresses imposed by the cancer and its treatment. This factor is difficult to assess directly. Instead, surrogate markers for physiologic reserve are used, such as the patient's age or Karnofsky performance status (Table 79-4). Older patients and those with a Karnofsky performance status <70 have a poor prognosis unless the poor performance is a reversible consequence of the tumor.

Increasingly, biologic features of the tumor are being related to prognosis. The expression of particular oncogenes, drug-resistance genes, apoptosis-related genes, and genes involved in metastasis are being found to influence response to therapy and prognosis. The presence of selected cytogenetic abnormalities may influence survival. Tumors with higher growth fractions, as assessed by expression of proliferation-related markers such as proliferating cell nuclear antigen (PCNA), behave more aggressively than tumors with lower growth fractions. Information obtained from studying the tumor itself will increasingly be used to influence treatment decisions.

MAKING A TREATMENT PLAN

From information on the extent of disease and the prognosis and in conjunction with the patient's wishes, it is determined whether the treatment approach should be curative or palliative in intent. Cooperation among the various professionals involved in cancer treatment is of the utmost importance in treatment planning. For some cancers, chemotherapy or chemotherapy plus radiation therapy delivered before the use of definitive surgical treatment (so-called neoadjuvant therapy) may improve the outcome, as seems to be the case for locally advanced breast cancer and head and neck cancers. In certain settings in which combined modality therapy is intended, coordination among the medical oncologist, radiation oncologist, and surgeon is crucial to achieving optimal results. Sometimes the chemotherapy and radiation therapy need to be delivered sequentially, and other times concurrently. Surgical procedures may precede or follow other treatment approaches. It is best for the treatment plan either to follow a standard protocol precisely or else to be part of an ongoing clinical research protocol evaluating new treatments. Ad hoc modifications of standard protocols are likely to compromise treatment results.

The choice of treatment approaches was formerly dominated by the local culture in both the university and the practice settings. However, it is now possible to gain access electronically to standard treatment protocols and to every approved clinical research study in North America through a personal computer interface with the Internet.

2The National Cancer Institute maintains a database called PDQ (Physician Data Query) that is accessible on the Internet under the name CancerNet at wwwicic.nci.nih.gov/health.htm. Information can be obtained through a facsimile machine using CancerFax by dialing 301-402-5874. Patient information is also provided by the National Cancer Institute in at least three formats: on the Internet via CancerNet at wwwicic.nci.nih.gov/patient.htm, through the CancerFax number listed above, or by calling 1-800-4-CANCER. The quality control for the information provided through these services is rigorous.

The skilled physician also has much to offer the patient for whom curative therapy is no longer an option. Often a combination of guilt and frustration over the inability to cure the patient and the pressure of a busy schedule greatly limit the time a physician spends with a patient who is receiving only palliative care. Resist these forces. In addition to the medicines administered to alleviate symptoms (see below), it is important to remember the comfort that is provided by holding the patient's hand, continuing regular examinations, and taking time to talk.

MANAGEMENT OF DISEASE AND TREATMENT COMPLICATIONS

Because cancer therapies are toxic (Chap. 84), patient management involves addressing complications of both the disease and its treatment as well as the complex psychosocial problems associated with cancer. In the short term during a course of curative therapy, the patient's functional status may decline. Treatment-induced toxicity is less acceptable if the goal of therapy is palliation. The most common side effects of treatment are nausea and vomiting (see below), febrile neutropenia (Chap. 85), and myelosuppression (Chap. 104). Therapeutic tools are now available to minimize the acute toxicity of cancer treatment.

New symptoms developing in the course of cancer treatment should always be assumed to be reversible until proven otherwise. The fatalistic attribution of anorexia, weight loss, and jaundice to recurrent or progressive tumor could result in a patient dying from a reversible intercurrent cholecystitis. Intestinal obstruction may be due to reversible adhesions rather than progressive tumor. Systemic infections, sometimes with unusual pathogens, may be a consequence of the immunosuppression associated with cancer therapy. Some drugs used to treat cancer or its complications (e.g., nausea) may produce central nervous system symptoms that look like metastatic disease or may mimic paraneoplastic syndromes such as the syndrome of inappropriate antidiuretic hormone. A definitive diagnosis should be pursued and may even require a repeat biopsy.

A critical component of cancer management is assessing the response to treatment. In addition to a careful physical examination in which all sites of disease are physically measured and recorded in a flow chart by date, response assessment usually requires periodic repeating of imaging tests that were abnormal at the time of staging. If imaging tests have become normal, repeat biopsy of previously involved tissue is performed to document complete response by pathologic criteria. Biopsies are not usually required if there is macroscopic residual disease. A *complete response* is defined as disappearance of all evidence of disease, and a *partial response* as>50% reduction in the sum of the products of the perpendicular diameters of all measureable lesions. *Progressive disease* is defined as the appearance of any new lesion or an increase of

>25% in the sum of the products of the perpendicular diameters of all measurable lesions. Tumor shrinkage or growth that does not meet any of these criteria is considered *stable disease*. Some sites of involvement (e.g., bone) or patterns of involvement (e.g., lymphangitic lung or diffuse pulmonary infiltrates) are considered unmeasurable. No response is complete without biopsy documentation of their resolution but partial responses may exclude their assessment unless clear objective (though unmeasurable) progression has occurred.

Tumor markers may be useful in patient management in certain tumors. Response to therapy may be difficult to gauge with certainty. However, some tumors produce or elicit the production of markers that can be measured in the serum or urine and, in a particular patient, rising and falling levels of the marker are usually associated with increasing or decreasing tumor burden, respectively. Some clinically useful tumor markers are shown in Table 79-5. Tumor markers are not in themselves specific enough to permit a diagnosis of malignancy to be made, but once a malignancy has been diagnosed and shown to be associated with elevated levels of a tumor marker, the marker can be used to assess response to treatment.

The recognition and treatment of depression are important components of management. The incidence of depression in cancer patients is ~25% overall and may be greater in patients with greater debility. This diagnosis is likely in a patient with a depressed mood (dysphoria) and/or a loss of interest in pleasure (anhedonia) for at least 2 weeks. In addition, three or more of the following symptoms are usually present: appetite change, sleep problems, psychomotor retardation or agitation, fatigue, feelings of guilt or worthlessness, inability to concentrate, and suicidal ideation. Patients with these symptoms should receive therapy. Medical therapy with a serotonin reuptake inhibitor such as fluoxetine (10 to 20 mg/d), sertraline (50 to 150 mg/d), or paroxetine (10 to 20 mg/d) or a tricyclic antidepressant such as amitriptyline (50 to 100 mg/d) or desipramine (75 to 150 mg/d) should be tried, allowing 4 to 6 weeks for response. Effective therapy should be continued at least 6 months after resolution of symptoms. If therapy is unsuccessful, other classes of antidepressants may be used. In addition to medication, psychosocial interventions such as support groups, psychotherapy, and guided imagery may be of benefit.

Many patients opt for unproven or unsound approaches to treatment when it appears that conventional medicine is unlikely to be curative. Those seeking such alternatives are often well educated and may be early in the course of their disease. Unsound approaches are usually hawked on the basis of unsubstantiated anecdotes and not only cannot help the patient but may be harmful. Physicians should strive to keep communications open and nonjudgmental, so that patients are more likely to discuss with the physician what they are actually doing. The appearance of unexpected toxicity may be an indication that a supplemental therapy is being taken.

[3] Information about unsound methods may be obtained from the National Council Against Health Fraud, Box 1276, Loma Linda, CA 92354, or from the Center for Medical Consumers and Health Care Information, 237 Thompson Street, New York, NY 10012.

LONG-TERM FOLLOW-UP/LATE COMPLICATIONS

At the completion of treatment, sites originally involved with tumor are reassessed, usually by radiography or imaging techniques, and any persistent abnormality is biopsied. If disease persists, the multidisciplinary team discusses a new salvage treatment plan. If the patient has been rendered disease-free by the original treatment, the patient is followed regularly for disease recurrence. The optimal guidelines for follow-up care are not known. For many years, a routine practice has been to follow the patient monthly for 6 to 12 months, then every other month for a year, every 3 months for a year, every 4 months for a year, every 6 months for a year, and then annually. At each visit, a battery of laboratory and radiographic and imaging tests were obtained on the assumption that it is best to detect recurrent disease before it becomes symptomatic. However, where follow-up procedures have been examined, this assumption has been found to be untrue. Studies of breast cancer, melanoma, lung cancer, colon cancer, and lymphoma have all failed to support the notion that asymptomatic relapses are more readily cured by salvage therapy than symptomatic relapses. In view of the enormous cost of a full battery of diagnostic tests and their manifest lack of impact on survival, new guidelines are emerging for less frequent follow-up visits during which the history and physical examination are the major investigations performed.

As time passes, the likelihood of recurrence of the primary cancer diminishes. For many types of cancer, survival for 5 years without recurrence is tantamount to cure. However, important medical problems can occur in patients treated for cancer and must be examined (Chap 103). Some problems emerge as a consequence of the disease and some as a consequence of the treatment. An understanding of these disease- and treatment-related problems may help in their detection and management.

Despite these concerns, most patients who are cured of cancer return to normal lives.

SUPPORTIVE CARE

In many ways, the success of cancer therapy depends on the success of the supportive care. Failure to control the symptoms of cancer and its treatment may lead patients to abandon curative therapy. Of equal importance, supportive care is a major determinant of quality of life. Even when life cannot be prolonged, the physician must strive to preserve its quality. Quality-of-life measurements have become common end-points of clinical research studies. Furthermore, palliative care has been shown to be cost-effective when approached in an organized fashion. A credo for oncology could be to cure sometimes, to extend life often, and to comfort always.

Pain Pain occurs with variable frequency in the cancer patient: 25 to 50% of patients present with pain at diagnosis, 33% have pain associated with treatment, and 75% have pain with progressive disease. The pain may have several causes. In about 70% of cases, pain is caused by the tumor itself -- by invasion of bone, nerves, blood vessels, or mucous membranes or obstruction of a hollow viscus or duct. In about 20% of cases, pain is related to a surgical or invasive medical procedure, to radiation injury (mucositis, enteritis, or plexus or spinal cord injury), or to chemotherapy injury (mucositis, peripheral neuropathy, phlebitis, steroid-induced aseptic necrosis of the femoral head). In 10% of cases, pain is unrelated to cancer or its treatment.

Assessment of pain requires the methodical investigation of the history of the pain, its location, character, temporal features, provocative and palliative factors, and intensity (Chap. 12); a review of the oncologic history and past medical history as well as personal and social history; and a thorough physical examination. The patient should be given a 10-division visual analogue scale on which to indicate the severity of the pain. The clinical condition is often dynamic, making it necessary to reassess the patient frequently. Pain therapy should not be withheld while the cause of pain is being sought.

A variety of tools are available with which to address cancer pain. About 85% of patients will have pain relief from pharmacologic intervention. However, other modalities, including antitumor therapy (such as surgical relief of obstruction, radiation therapy, and strontium-89 or samarium-153 treatment for bone pain), neurostimulatory techniques, regional analgesia, or neuroablative procedures are effective in an additional 12% or so. Thus, very few patients will have inadequate pain relief if appropriate measures are taken.

The World Health Organization (WHO) has devised a simple and effective method for the rational titration of oral analgesia, called the *WHO ladder*. The ladder has the following three steps. (1) For mild to moderate pain, one begins with acetaminophen (650 mg every 4 h or 975 mg every 6 h), aspirin (650 mg every 4 h or 975 mg every 6 h), or a nonsteroidal anti-inflammatory agent (NSAID; e.g., ketoprofen, 25 to 60 mg every 6 h) with or without an adjuvant such as a glucocorticoid (dexamethasone) or an antidepressant (amitriptyline). (2) When pain persists or increases, an opioid such as codeine or hydrocodone (30 mg every 3 to 4 h is roughly equivalent to 10 mg of intravenous morphine) should be added (not substituted); fixed combinations such as oxycodone/acetaminophen (Percocet) or oxycodone/aspirin (Percodan) are worth testing. (3) Pain that is persistent or that is moderate to severe at the outset should be treated by increasing the potency of the opioid or using higher dosages (e.g., morphine, 15 to 30 mg every 3 to 4 h, or controlled-release morphine, 90 to 120 mg bid), and fixed opioid/NSAID combinations should be abandoned. Adjuvants may be used at all steps. The critical features of this approach are that the treatment is oral, should be given around the clock with supplemental doses as needed to control pain, and is tailored to the individual patient. Transmucosal fentanyl (in lollipop form) may aid in control of breakthrough pain. Records of pain control should be a prominent component of the medical record. When opioids are used, the patient should be placed on a prophylactic regimen to prevent constipation.

Nausea Emesis in the cancer patient is usually caused by chemotherapy (Chap. 84). Its severity can be predicted from the drugs used to treat the cancer. Three forms of emesis are recognized on the basis of their timing with regard to the noxious insult. *Acute emesis*, the most common variety, occurs within 24 h of treatment. *Delayed emesis* occurs 1 to 7 days after treatment; it is rare, but, when present, usually follows cisplatin administration. *Anticipatory emesis* occurs before administration of chemotherapy and represents a conditioned response to visual and olfactory stimuli previously associated with chemotherapy delivery.

Acute emesis is the best understood form. Stimuli that activate signals in the chemoreceptor trigger zone in the medulla, the cerebral cortex, and peripherally in the intestinal tract lead to stimulation of the vomiting center in the medulla, the motor center responsible for coordinating the secretory and muscle contraction activity that leads to emesis. Diverse receptor types participate in the process, including dopamine, serotonin, histamine, opioid, and acetylcholine receptors. The serotonin receptor antagonists ondansetron and granisetron are the most effective drugs against highly emetogenic agents, but they are expensive.

As with the analgesia ladder, emesis therapy should be tailored to the situation. For mildly and moderately emetogenic agents, prochlorperazine, 5 to 10 mg orally or 25 mg rectally, is effective. Its efficacy may be enhanced by administering the drug before the chemotherapy is delivered. Dexamethasone, 10 to 20 mg intravenously, is also effective and may enhance the efficacy of prochlorperazine. For highly emetogenic agents such as cisplatin, mechlorethamine, dacarbazine, and streptozocin, combinations of agents work best and administration should begin 6 to 24 h before treatment. Ondansetron, 8 mg orally every 6 h the day before therapy and intravenously on the day of therapy, plus dexamethasone, 20 mg intravenously before treatment, is an effective regimen. Like pain, emesis is easier to prevent than to alleviate.

Delayed emesis may be related to bowel inflammation from the therapy and can be controlled with oral dexamethasone and oral metoclopramide, a dopamine receptor antagonist that also blocks serotonin receptors at high dosages. The best strategy for preventing anticipatory emesis is to control emesis in the early cycles of therapy to prevent the conditioning from taking place. If this is unsuccessful, prophylactic antiemetics the day before treatment may help. Experimental studies are evaluating behavior modification.

Effusions Fluid may accumulate abnormally in the pleural cavity, pericardium, or peritoneum. Asymptomatic malignant effusions may not require treatment. Symptomatic effusions occurring in tumors responsive to systemic therapy usually do not require local treatment but respond to the treatment for the underlying tumor. Symptomatic effusions occurring in tumors unresponsive to systemic therapy may require local treatment in patients with a life expectancy of at least 6 months.

Pleural effusions due to tumors may or may not contain malignant cells. Lung cancer, breast cancer, and lymphomas account for about 75% of malignant pleural effusions. Their exudative nature is usually gauged by an effusion/serum protein ratio of 0.5 or an effusion/serum lactate dehydrogenase ratio of 0.6. When the condition is symptomatic, thoracentesis is usually performed first. In most cases, symptomatic improvement occurs for <1 month. Chest tube drainage is required if symptoms recur within 2 weeks. Fluid is aspirated until the flow rate is <100 mL in 24 h. Then either 60 units of bleomycin or 1 g of doxycycline is infused into the chest tube in 50 mL of 5% dextrose in water; the tube is clamped; the patient is rotated on four sides, spending 15 min in each position; and, after 1 to 2 h, the tube is again attached to suction for another 24 h. The tube is then disconnected from suction and allowed to drain by gravity. If<100 mL drains over the next 24 h, the chest tube is pulled, and a radiograph taken 24 h later. If the chest tube continues to drain fluid at an unacceptably high rate, sclerosis can be repeated. Bleomycin may be somewhat more effective than doxycycline but is very expensive. Doxycycline is usually the drug of first choice. If neither doxycycline nor bleomycin is effective, talc can be used.

Symptomatic pericardial effusions are usually treated by creating a pericardial window or by stripping the pericardium. If the patient's condition does not permit a surgical procedure, sclerosis can be attempted with doxycycline and/or bleomycin.

Malignant ascites is usually treated with repeated paracentesis of small volumes of fluid. If the underlying malignancy is unresponsive to systemic therapy, peritoneovenous shunts may be inserted. Despite the fear of disseminating tumor cells into the circulation, widespread metastases are an unusual complication. The major complications are occlusion, leakage, and fluid overload. Patients with severe liver disease may develop disseminated intravascular coagulation.

Nutrition Cancer and its treatment may lead to a decrease in nutrient intake of sufficient magnitude to cause weight loss and alteration of intermediary metabolism. The prevalence of this problem is difficult to estimate because of variations in the definition of cancer cachexia, but most patients with advanced cancer experience weight loss and decreased appetite. A variety of both tumor-derived factors (e.g., bombesin, adrenocorticotropic hormone) and host-derived factors (e.g., tumor necrosis factor, interleukins 1 and 6, growth hormone) contribute to the altered metabolism, and a vicious cycle is established in which protein catabolism, glucose intolerance, and lipolysis cannot be reversed by the provision of calories.

It remains controversial how to assess nutritional status and when and how to intervene. Efforts to make the assessment objective have included the use of a prognostic nutritional index based on albumin levels, triceps skin fold thickness, transferrin levels, and delayed-type hypersensitivity skin testing. However, a simpler approach has been to define the threshold for nutritional intervention as>10% unexplained body weight loss, serum transferrin level <1500 mg/L (150 mg/dL), and serum albumin <34 g/L (3.4 g/dL).

The decision is important, because it appears that cancer therapy is substantially more toxic and less effective in the face of malnutrition. Nevertheless, it remains unclear whether nutritional intervention can alter the natural history. Unless some pathology is affecting the absorptive function of the gastrointestinal tract, enteral nutrition provided orally or by tube feeding is preferred over parenteral supplementation. However, the risks associated with the tube may outweigh the benefits. Megestrol acetate, a progestational agent, has been advocated as a pharmacologic intervention to improve nutritional status. Research in this area may provide more tools in the future as cytokine-mediated mechanisms are further elucidated.

Psychosocial Support The psychosocial needs of patients vary with their situation. Patients undergoing treatment experience fear, anxiety, and depression. Self-image is often seriously compromised by deforming surgery and loss of hair. Women who receive cosmetic advice that enables them to look better also feel better. Loss of control over how one spends time can contribute to the sense of vulnerability. Juggling the demands of work and family with the demands of treatment may create enormous stresses. Sexual dysfunction is highly prevalent and needs to be discussed openly with the patient. An empathetic health care team is sensitive to the individual patient's needs and permits negotiation where such flexibility will not adversely affect the course of treatment.

Cancer survivors have other sets of difficulties. Patients may have fears associated with the termination of a treatment they associate with their continued survival. Adjustments are required to physical losses and handicaps, real and perceived. Patients may be preoccupied with minor physical problems. They perceive a decline in their job mobility and view themselves as less desirable workers. They may be victims of job and/or insurance discrimination. Patients may experience difficulty reentering their normal past life. They may feel guilty for having survived and may carry a sense of vulnerability to colds and other illnesses. Perhaps the most pervasive and threatening concern is the ever-present fear of relapse (the Damocles syndrome).

Patients in whom therapy has been unsuccessful have other problems related to the end of life.

Death and Dying The most common causes of death in patients with cancer are infection (leading to circulatory failure), respiratory failure, hepatic failure, and renal failure. Intestinal blockage may lead to inanition and starvation. Central nervous system disease may lead to seizures, coma, and central hypoventilation. About 70% of patients develop dyspnea preterminally. However, many months usually pass between the diagnosis of cancer and the occurrence of these complications, and during this period the patient is severely affected by the possibility of death. The path of unsuccessful cancer treatment usually occurs in three phases. First, there is optimism at the hope of cure; when the tumor recurs, there is the acknowledgment of an incurable disease, and the goal of palliative therapy is embraced in the hope of being able to live with disease; finally, at the disclosure of imminent death, another adjustment in outlook takes place. The patient imagines the worst in preparation for the end of life and may go through stages of adjustment to the diagnosis. These stages include denial, isolation, anger, bargaining, depression, acceptance, and hope. Of course, patients do not all progress through all the stages or proceed through them in the same order or at the same rate. Nevertheless, developing an understanding of how the patient has been affected by the diagnosis and is coping with it is an important goal of patient management.

It is best to speak frankly with the patient and the family regarding the likely course of disease. These discussions can be difficult for the physician as well as for the patient and family. The critical features of the interaction are to reassure the patient and family that everything that can be done to provide comfort will be done. They will not be abandoned. Many patients prefer to be cared for in their homes or in a hospice setting rather than a hospital. The American College of Physicians has published a book called *Home Care Guide for Cancer: How to Care for Family and Friends at Home* that teaches an approach to successful problem-solving in home care. With appropriate planning, it should be possible to provide the patient with the necessary medical care as well as the psychological and spiritual support that will prevent the isolation and depersonalization that can attend in-hospital death.

The care of dying patients may take a toll on the physician. A "burnout" syndrome has been described that is characterized by fatigue, disengagement from patients and colleagues, and a loss of self-fulfillment. Efforts at stress reduction, maintenance of a balanced life, and setting realistic goals may combat this disorder.

End-of-Life Decisions Unfortunately, a smooth transition in treatment goals from curative to palliative may not be possible in all cases because of the occurrence of serious treatment-related complications or rapid disease progression. Vigorous and invasive medical support for a reversible disease or treatment complication is assumed to be justified. However, if the reversibility of the condition is in doubt, the patient's wishes determine the level of medical care. These wishes should be elicited before the terminal phase of illness and reviewed periodically. This information can guide the physician should the patient be unable to speak for him- or herself. The family cannot be expected to make such decisions without guidance from the patient and support from the physician when surrogate decisions are required. Advance directives such as a living will or a durable power of attorney for health care provide guidance for the health care team and the family regarding the patient's wishes and may protect the patient's assets from depletion on expensive but unwanted care.

Only about 15%of the population has implemented an advance directive. Physicians should take the initiative to speak with patients and family members about advance directives.

[4] Information about advance directives can be obtained from the American Association of Retired Persons, 601 E Street, NW, Washington, DC 20049, 202-434-2277 or Choice in Dying, 250 West 57th Street, New York, NY 10107, 212-366-5540.

(Bibliography omitted in Palm version)

Back to Table of Contents

80. PREVENTION AND EARLY DETECTION OF CANCER - *Otis W. Brawley*, *Barnett S. Kramer*

The prevention and control of cancer is a burgeoning field because of advances in understanding the biology of carcinogenesis. The field has expanded beyond the identification and avoidance of carcinogens to include studies of specific interventions to lower cancer risk, as well as screening for early detection of cancer.

Central to the prevention and control of cancer is the concept that carcinogenesis is not an event but a process, a series of discrete cellular changes that result in progressively more autonomous cellular processes. *Primary prevention* concerns the identification and manipulation of the genetic, biologic, and environmental factors in the causal pathway. Smoking cessation, diet modification, and chemoprevention are primary prevention activities. *Secondary prevention* concerns the identification of asymptomatic neoplastic lesions combined with effective therapy. Screening is a form of secondary prevention. Screening may also be a form of primary prevention of invasive cancer; screening Pap smears are used to identify and treat preinvasive lesions of the cervix.

EDUCATION AND HEALTHFUL HABITS

Public education on the avoidance of identified risk factors for cancer and encouraging healthy habits were among early efforts in cancer prevention and control. Many educational messages have come to the public through commercials in the print and electronic media and through school health courses. The physician is a potentially powerful messenger in this education campaign about the hazards of smoking, the benefits of a healthful diet, and sun avoidance.

Smoking Cessation Tobacco use through cigarettes and other means is the most avoidable risk factor for cardiovascular disease and cancer. Lung cancer mortality rates correlate with the number of cigarettes smoked per day as well as the degree of inhalation of cigarette smoke. Those who stop smoking have a lower lung cancer mortality rate than those who continue smoking, despite the persistence for years of some carcinogen-induced genetic mutations. In addition to lung cancer, cigarette smoking is a causative agent in cancers of the larynx, oropharynx, esophagus, bladder, and pancreas. Smoking cessation and avoidance have the potential to save and extend more lives than any other public health activity. About 400,000 Americans die prematurely every year because of cigarette smoking. A smoker has a one in three lifetime risk of dying prematurely of a cancer or cardiovascular or pulmonary disease caused by cigarette smoking. Indeed, more human lives are lost due to cardiovascular disease caused by smoking than from smoking-related cancer. The risk of tobacco smoke is not necessarily limited to the smoker. Epidemiologic studies suggest that environmental tobacco smoke may cause lung cancer and other pulmonary diseases in nonsmokers.

Nonsmoking persons should be encouraged not to start smoking, and persons who smoke should be encouraged to stop. Tobacco prevention is a pediatric issue. Over 80% of American smokers begin smoking before the age of 18. Nearly 20% of Americans aged 12 to 18 have smoked a cigarette in the past month. Counseling of adolescents and young adults is critical to prevent smoking. A physician's simple advice to not start smoking or to quit smoking can be of benefit. The U.S. Agency for Health Care Research and Quality recommends that physicians query patients on tobacco use on every office visit, record the answer with the vital signs, and ask smokers if they would like assistance in quitting.

Current approaches to smoking cessation recognize that smoking is an addiction (Chap. 390). The smoker who is quitting goes through a process with identifiable stages that include contemplation of quitting, an action phase in which the smoker quits, and a maintenance phase. Smokers who quit completely are more likely to be successful than those who gradually reduce the number of cigarettes smoked or change to cigarettes lower in tar or nicotine. More than 90% of the Americans who have successfully quit smoking did so on their own without participation in an organized cessation program, but cessation programs are helpful for some smokers. The Community Intervention Trial for Smoking Cessation (COMMIT) was a community-based 4-year program. One community of each of 11 matched community pairs was randomly assigned to intervention. The intervention included public education through the media and community-wide events, health care providers, worksites and other organizations, and cessation resources. COMMIT demonstrated that light smokers can benefit from simple cessation messages and cessation programs. The quit rate (fraction of the subjects followed who achieved and maintained cessation at the end of the trial) was 30.6% in the intervention communities and 27.5% in the control communities. This finding is statistically significant, but modest. The control communities enjoyed a substantial decrease in smoking through study participation. The COMMIT interventions were not successful for heavy smokers (>25 cigarettes per day). Heavy smokers need an intensive, broad-based cessation program that includes counseling, behavioral strategies, and pharmacologic adjuncts such as nicotine gum and nicotine patches.

Cigar and pipe smoking carry the same risks as tobacco smoke including lung cancer. Smokeless tobacco is the fastest growing part of the tobacco industry and represents a significant health risk. Chewing tobacco is a carcinogen linked to dental caries, gingivitis, oral leukoplakia, and oral cancer. The systemic effects of smokeless tobacco may increase risks for other cancers. Nitrosamines found in smokeless tobacco cause lung cancer in laboratory animals.

Diet Modification Dietary modification may have significant potential for lowering cancer risk in western culture. Studies of international dietary patterns and animal studies suggest that diets high in fat increase the risk for cancers of the breast, colon, prostate, and endometrium. These cancers have their highest incidence and mortalities in western countries, where fat comprises an average of 40 to 45% of the total calories consumed. In populations at low risk for these cancers, fat accounts for<20% of dietary calories.

Nonetheless, dietary fat has not been accepted by all as important in the etiology of cancers. Case-control and cohort epidemiologic studies give conflicting results. In addition, diet is a highly complex exposure to many nutrients and chemicals. Low-fat diets may render some protection through anticarcinogens found in vegetables, fruits, legumes, nuts, and grains. Substances found in these foods that may be protective include phenols, sulfur-containing compounds, flavones, and fiber.

In observational studies, dietary fiber appears protective against colonic polyps and invasive cancer of the colon. The mechanisms involved are complex and speculative. They involve binding of oxidized bile acids, a decrease in bowel transit time, and generation of soluble fiber products, such as butyrate, that may have differentiating properties. High-fiber diets may also protect against breast and prostate cancer by absorbing and inactivating dietary estrogenic and androgenic cancer promoters. Protective effects of fiber have not been proved in a prospective clinical trial.

The U.S. National Institutes of Health (NIH) Women's Health Initiative, launched in 1994, is a long-term clinical trial enrolling more than 100,000 women aged 45 to 69. It studies the potential cancer-preventing effects of a low-fat diet and vitamin supplementation. It must be stressed that the scientific evidence does not currently establish the anticarcinogenic value of vitamin, mineral, or nutritional supplements in amounts greater than that provided by a good diet.

The Polyp Prevention Trial studied 2000 elderly persons randomly assigned to a low-fat, high-fiber diet or a routine diet followed for 4 years. No significant differences in polyp formation were noted.

A simple way to decrease dietary fat and increase fiber is to consume at least 5 to 9 servings of fruits and vegetables a day. Such a diet may lower the risk of cardiac disease as well as cancer.

Sun Avoidance Nonmelanoma skin cancers (basal cell and squamous cell) are induced by cumulative exposure to ultraviolet radiation. Intermittent acute sun exposure and sun damage have been linked to melanoma. Sunburns, especially in childhood and adolescence, are associated with an increased risk of melanoma in adulthood. Reduction of sun exposure through use of protective clothing and changes in the pattern of outdoor activities can reduce skin cancer risk. Sunscreens decrease the risk of actinic keratoses, the precursor to squamous cell skin cancer, but melanoma risk may be increased. Sunscreens prevent burning and may encourage more prolonged exposure to the sun; yet they may not filter out wavelengths of energy that cause melanoma.

Educational interventions to help people assess their risk of developing skin cancer accurately have some impact. Self-examination for skin pigment characteristics associated with melanoma, such as freckling, may be useful in identifying people at high risk. People who recognize themselves as being at risk tend to be more compliant with sun-avoidance recommendations. Possible risk factors for melanoma include a propensity to sunburn, a large number of benign melanocytic nevi, and atypical nevi.

CANCER CHEMOPREVENTION

Chemoprevention of cancer is a relatively new concept. It involves the use of specific natural or synthetic chemical agents to reverse, suppress, or prevent carcinogenesis before the development of invasive malignancy. While the concept that pharmacologic agents can prevent a cancer is relatively new, the idea that a compound can prevent chronic disease is not. Clinicians routinely prevent heart disease, kidney disease, and stroke by treating hypertension with pharmacologic agents. Lipid-lowering drugs are used to prevent coronary artery disease.

Improved understanding of the biology of cancer makes chemoprevention a real possibility. Cancer develops through an accumulation of genetic changes that are potential points of intervention to prevent cancer. The initial genetic changes are termed *initiation*. The alteration can be inherited or acquired through the action of physical, infectious, or chemical carcinogens. Like most human diseases, cancer arises through an interaction between genetics and environmental exposures (Table 80-1). Influences that cause the initiated cell to progress through the carcinogenic process and to change phenotypically are termed *promoters*. Promoters include hormones such as androgens, linked to prostate cancer, and estrogen, linked to breast and endometrial cancer. The distinction between an initiator and a promoter is sometimes arbitrary; some components of cigarette smoke are "complete carcinogens," acting as both initiators and promoters. Cancer can be prevented or controlled through interference with the factors that cause initiation, promotion, or progression. Compounds of interest in chemoprevention often have antimutagenic, antioxidant, or antiproliferative activity.

Before a chemoprevention strategy can become standard practice, evidence of benefit must be gathered from clinical trials. These trials are usually large, long-term, randomized, placebo-controlled, and double-blinded. They often allow for the study of drugs for prevention of multiple cancers and the study of end-points beyond cancer, such as other chronic diseases. Several large clinical trials have been completed, and a number are continuing in the twenty-first century. Only tamoxifen has been approved by the U.S. Food and Drug Administration for prevention; it lowers risk of breast cancer in high-risk women.

Multiple Cancer Site Prevention Trials The Physicians' Health Trial involves 22,071 American male physicians. Participants were randomly assigned to receiveb-carotene, aspirin, and/or placebo in a 2´ 2 factorial design. All major medical events were recorded. In 1988, the aspirin arm was unblinded after the trial demonstrated that aspirin therapy causes a significant reduction in cardiovascular mortality. The b-carotene arm of the study stopped in 1998, and data analysis is proceeding.

The Women's Health Study, launched in 1992, is a 10-year trial involving 44,000 female nurses. Subjects are randomly assigned to b-carotene,a-tocopherol, aspirin, and/or placebo in a factorial design yielding eight different treatment groups. The end-points are total epithelial cancers, breast cancer, lung cancer, colon cancer, and vascular disease.

The Women's Health Initiative uses a partial factorial design that places women in 22 intervention groups. Participants can receive calcium and vitamin D supplementation, hormone replacement therapy, and counseling to increase exercise and cease smoking. Prevention of a number of cancers, cardiovascular disease, osteoporosis, and other diseases will be assessed.

Prevention of Hormonally Driven Cancers Hormonal manipulation is being tested in the primary prevention of breast and prostate cancer. Tamoxifen is an antiestrogen with partial estrogen agonistic activity in some tissues, such as endometrium and bone. One of its actions is to upregulate transforming growth factorb, which decreases breast cell proliferation. In randomized placebo-controlled trials to assess tamoxifen as an adjuvant in breast cancer treatment, this drug reduced the number of new breast cancers in the uninvolved breast by more than a third. In a randomized placebo-controlled trial involving >13,000 women at high risk, tamoxifen decreased the risk of developing cancer by 49% compared to placebo. Tamoxifen also reduced the risk of bone fractures; a small increase in risk of endometrial cancer, stroke, pulmonary emboli, and deep vein thrombosis was noted. A trial to compare tamoxifen with another selective estrogen receptor modulator, raloxifene, is ongoing.

Finasteride is a 5a-reductase inhibitor. It inhibits the conversion of testosterone to dihydrotestosterone, a more potent stimulator of prostate cell proliferation than testosterone. In an F344 rat model of carcinogen-induced prostate cancer, finasteride decreased the incidence of cancers. Finasteride is being tested as a preventive agent for prostate cancer in a 10-year study involving 18,000 men age 55 and older.

Chemoprevention of Cancers of the Upper Aerodigestive Tract Smoking causes diffuse epithelial injury in the head, neck, esophagus, and lung. Patients cured of squamous cell cancers of the lung, esophagus, head, and neck are at risk (as high as 5% per year) of developing a second cancer of the upper aerodigestive tract. Cessation of cigarette smoking does not markedly decrease the cured cancer patient's risk of second malignancy, even though it does lower the cancer risk in those who have never developed a malignancy. Smoking cessation may halt the early stages of the carcinogenic process (such as metaplasia), but it may have no effect on late stages of carcinogenesis. This "field carcinogenesis" hypothesis for cancer of the upper aerodigestive tract has made "cured" patients an important population for chemoprevention of second malignancies. A randomized, placebo-controlled clinical trial has demonstrated that adjuvant isoretinoin (13-*cis*-retinoic acid) can reduce the incidence of second primary tumors in patients treated with local therapy for head and neck cancer. However, overall survival was not improved due to mortality from recurrences of the primary tumor.

Oral leukoplakia, a premalignant lesion commonly found in smokers, has been used as an intermediate marker allowing the demonstration of chemopreventive activity in smaller, shorter-duration, randomized, placebo-controlled trials. Response was associated with upregulation of retinoic acid receptor b. Therapy with isoretinoin causes regression of oral leukoplakia. However, the lesions recur when the agent is withdrawn, suggesting the need for chronic administration of retinoids. Premalignant lesions in the oropharyngeal area have also responded to b-carotene, retinol,a-tocopherol (vitamin E), and selenium. Further study to improve the definition of the activity of these drugs is ongoing. The ability of isoretinoin to prevent second malignancies in patients cured of early-stage non-small cell lung cancer is also being assessed.

Several large-scale trials have assessed agents in the chemoprevention of lung cancer in patients at high risk. In the Alpha-Tocopherol/Beta-Carotene (ATBC) Lung Cancer Prevention Trial, participants were male smokers, aged 50 to 69 at entry. At entry, participants had smoked an average of one pack of cigarettes per day for 35.9 years. Participants received a-tocopherol,b-carotene, and/or placebo in a randomized, 2´ 2 factorial design. After a median follow-up of 6.1 years, lung cancer incidence and mortality were statistically significantly *increased* in those receiving b-carotene.a-Tocopherol had no significant impact on lung cancer mortality, and no evidence suggested interaction between the two drugs. Patients receiving a-tocopherol had a higher incidence of hemorrhagic stroke.

The Beta-Carotene and Retinol Efficacy Trial (CARET) involved 17,000 American smokers and workers with asbestos exposure. Entrants were randomly assigned to one of four arms and received b-carotene, retinol, and/or placebo in a 2´2 factorial design. This trial demonstrated harm from b-carotene: a lung cancer rate of 5 per 1000 subjects per year for those taking placebo and of 6 per 1000 subjects per year for those taking b-carotene. The difference was statistically significant.

TheseATBC andCARETresults demonstrate the importance of testing chemoprevention hypotheses before implementing them widely, because the results stand in contrast to a number of observational epidemiologic studies. In the ATBC trial, participants takinga-tocopherol had a one-third reduction in the incidence of prostate cancer, compared to those not takinga-tocopherol. Assessment of these findings continues. The Physicians' Health Trial showed neither an increased nor a decreased risk of lung cancer in those using b-carotene; fewer of its participants were smokers than those in the ATBC and CARET studies.

Chemoprevention of Colon Cancer Many of the current colon cancer prevention trials are based on the premise that most colorectal cancers develop from adenomatous polyps. These trials use adenoma recurrence or disappearance as a surrogate end-point to assess colon cancer prevention. Early clinical trial results suggest that nonsteroidal anti-inflammatory drugs (NSAIDs), such as piroxicam, sulindac, and aspirin, may prevent adenoma formation or cause regression of adenomatous polyps. The mechanism of action of NSAIDs is unknown, but they are presumed to work through the cyclooxygenase pathway. In the Physicians' Health Trial, aspirin had no effect on colon cancer incidence, although the 6-year assessment period may not have been long enough to evaluate definitively this end-point.

Epidemiologic studies suggest that diets high in calcium lower colon cancer risk. Calcium binds bile and fatty acids, which cause hyperproliferation of colonic epithelium. It is hypothesized that this effect reduces intraluminal exposure to these compounds. Early data from randomized studies suggest that calcium supplementation decreases the risk of adenomatous polyp recurrence by about 20%, even though it does not decrease the proliferative rate of the colonic epithelium. Epithelial proliferative rate may not be an adequate surrogate marker in colon cancer prevention trials.

Cyclooxygenase II inhibitors may be even more effective at colon cancer prevention.

Vaccines and Cancer Prevention A number of infectious agents have been linked to the development of cancer, leading to interest in developing vaccines to protect against these agents. The hepatitis B vaccine is quite effective in preventing hepatitis and hepatomas due to chronic hepatitis B infection. Public health officials are encouraging widespread administration of this vaccine, especially in Asia, where the disease is epidemic. In the future, human papilloma virus (HPV) vaccines could be developed to prevent cervical cancer, and *Helicobacter pylori* vaccines may be developed to prevent gastric cancer.

CANCER SCREENING

Screening is a means of detecting disease early in asymptomatic individuals with the goal of decreasing morbidity and mortality. Screening for cancer is intuitively appealing and has attracted great public interest as technology has generated a number of diagnostic tests and procedures that are safe, quick, and inexpensive. While screening can potentially save lives, and has been shown clearly to do so in the case of breast, cervical, and colon cancer, it is also subject to a number of biases, which can suggest a benefit when actually there is none. Bias can even mask net harm. Early detection does not in itself confer benefit. To be of value, screening must detect disease earlier, and treatment of earlier disease must yield a better outcome than treatment at the onset of symptoms. Cause-specific mortality, rather than survival after diagnosis, is the preferred end point (see below).

Because screening is done on asymptomatic, healthy persons, it should offer substantial likelihood of benefit. A critical approach to screening is necessary to ensure that benefit results. Screening tests and their appropriate use should be carefully evaluated before their use is widely encouraged in screening programs as a matter of public policy.

Screening examinations, tests, or procedures are usually not diagnostic of cancer but instead indicate that a cancer may be present. The diagnosis is then made following a workup that includes a biopsy and pathologic confirmation.

A number of genes have been identified that predispose for a disease, and many more will be identified in the near future. Testing for these genes can define a high-risk population. The ability to predict the development of a particular cancer may some day present therapeutic options as well as ethical dilemmas. It may eventually allow for early intervention to prevent a cancer or limit its severity. People at high risk will be ideal candidates for chemoprevention and screening; however, the efficacy of these interventions in the high-risk population should be investigated. Currently, persons at high risk for a particular cancer can engage in intensive screening. While this course is clinically prudent, it is not known if it saves lives in these populations.

The Accuracy of Screening A screening test's accuracy or ability to discriminate disease is described by four indices: sensitivity, specificity, positive predictive value, and negative predictive value (Table 80-2). *Sensitivity* is the proportion of persons with the disease who test positive in the screen (i.e., the ability of the test to detect disease when it is present). *Specificity* is the proportion of persons who do not have the disease and test negative in the screening test (i.e., the ability of a test to tell that the disease is not present). The *positive predictive value* is the proportion of persons who test positive who actually have the disease. Similarly, *negative predictive value* is the proportion of who test negative and do not have the disease. The sensitivity and specificity of a test are relatively independent of the underlying prevalence (or risk) of the disease in the population screened, but the predictive values depend strongly on the prevalence of the disease (Table 80-3).

Screening is most beneficial, efficient, and economical when the target disease is common in the population being screened. To be valuable, the screening test should have a high specificity; sensitivity need not be very high, as demonstrated inTable 80-3.

Potential Biases of Screening Tests The common biases of screening are lead time, length, and selection. These biases can make a screening test seem beneficial when actually it is not (or even causes net harm). Whether beneficial or not, screening can create the false impression of an epidemic by increasing the number of cancers diagnosed. It can also give the appearance of a shift in stage, thus improving survival statistics without reducing mortality (i.e., the number of deaths from a given cancer relative to the number of people at risk for the cancer). In such a case, the *apparent* duration of survival increases without lives being saved or life expectancy changed.

Lead-time bias occurs when a test does not influence the natural history of the disease; the patient is merely diagnosed at an earlier date. When lead-time bias occurs, survival *appears* increased, but life is not really prolonged. The screening test only prolongs the time the subject is aware of the disease and spends as a patient.

Length bias occurs when slow-growing, less aggressive cancers are detected during screening. Cancers diagnosed owing to the onset of symptoms between scheduled screenings are on average more aggressive, and treatment outcomes are not as favorable. An extreme form of length bias is termed *overdiagnosis*, the detection of "pseudodisease." The reservoir of some undetected slow-growing tumors is large. Many of these tumors fulfill the histologic criteria of cancer but will never become clinically significant or cause death. This problem is compounded by the fact that the most common cancers appear most frequently at ages when competing causes of death are more frequent.

Selection bias must be considered in assessing the results of any screening effort. The population most likely to seek screening may differ from the general population to which the screening test might be applied. The individuals screened may have volunteered because of a particular risk factor not found in the general population, such as a strong family history. In general, volunteers for studies may be more health conscious and thus likely to have a better prognosis or lower mortality rate, irrespective of the screening result. This is termed the *healthy volunteer effect*.

Potential Drawbacks of Screening Risks associated with screening include harm caused by the screening intervention itself, harm due to the further investigation of persons with positive test results (both true and false positives), and harm from the treatment of persons with a true-positive result, even if life is extended by treatment. The diagnosis and treatment of cancers that would never have caused medical problems can lead to the harm of unnecessary treatment and give patients the anxiety of a cancer diagnosis. The psychosocial impact of cancer screening, whether the result is positive or negative, can also be substantial when applied to the entire population.

Assessment of Screening Tests Good clinical trial design can offset some biases of screening and demonstrate the relative risks and benefits of a screening test. A randomized, controlled screening trial with cause-specific mortality as the end-point provides the strongest support for a screening intervention. In a randomized trial, two like populations are randomly established. One is given the medical standard of care (which may be no screening at all), and the other receives the screening intervention being assessed. The two populations are compared over time. Efficacy for the population studied is established when the group receiving the screening test has a better cause-specific mortality rate than the control group. Studies showing a reduction in the incidence of advanced-stage disease, an improved survival, or a stage shift are weaker evidence of benefit. These latter criteria are necessary but not sufficient to establish the value of a screening test.

Although a randomized, controlled screening trial provides the strongest evidence to support the usefulness of a screening test, it is not perfect. Unless the trial is population-based, it does not remove the issue of generalizability to the target population. Screening trials generally involve thousands of persons and last for years. Less definitive study designs are therefore often used to estimate the effectiveness of screening practices. After a randomized controlled clinical trial, in descending order of strength, evidence may be derived from:

· The findings of internally controlled trials using intervention allocation methods other than randomization (e.g., allocation determined by birth date, date of clinic visit);

· The findings of cohort or case-control analytic observational studies;

· The results of multiple time series studies with or without the intervention;

· The opinions of respected authorities based on clinical experience, descriptive studies, or consensus reports of experts (the weakest evidence because even experts can be misled by the biases described above).

Screening for Specific Cancers Widespread screening for breast, cervical, and colon cancer is beneficial for certain age groups. Special surveillance of those at high risk for a specific cancer because of a family history or a genetic risk factor may be prudent, but few studies have been carried out to assess the impact of this practice on mortality in specific high-risk populations. A number of organizations have considered whether or not to endorse routine use of certain screening tests. Because these groups have not used the same criteria to judge whether a screening test should be endorsed, they have arrived at different recommendations. The screening guidelines of the U.S. Preventive Services Task Force, the Canadian Task Force on Preventive Health Care, and the American Cancer Society are often quoted and show a range of recommendations (Table 80-4).

Breast Cancer Breast self-examination, clinical breast examination by a care giver, and mammography have been advocated as useful screening tools. Only screening mammography alone and screening mammography with clinical examination have been evaluated in randomized controlled trials. A number of well-designed trials have demonstrated that annual or biennial screening with mammography or mammography plus clinical breast examination in women over the age of 50 saves lives. In these trials, the breast cancer mortality rate is decreased by about a third. Experts disagree on whether average-risk women aged 40 to 49 should receive regular screening (Table 80-4). The statistical significance of the screening effect in women aged 40 to 49 depends on the statistical test used. An analysis of eight large randomized trials showed no benefit from mammographic screening for women aged 40 to 49 when assessed 5 to

7 years after trial entry. However, a small benefit emerged 10 to 12 years after study entry. What proportion of this possible benefit is due to screening after these women turned 50 is not known. In randomized screening studies of women aged 50 to 69, the decline in mortality begins about 5 years after initiation of screening. Nearly half of women aged 40 to 49 years screened annually will have false-positive mammograms necessitating further evaluation, often including biopsy. The risk of false-positive testing should be discussed with the patient.

While no study has shown breast self-examination to decrease mortality, it is recommended as prudent by many organizations. A substantial fraction of breast cancers are first detected by the patient, even with widespread mammographic screening.

Cervical Cancer Screening with Papanicolaou smears decreases cervical cancer mortality. The cervical cancer mortality rate has fallen significantly since the widespread use of the Pap smear, although this trend actually began earlier. Most screening guidelines recommend regular Pap testing for all women who are or have been sexually active or have reached the age of 18. With the onset of sexual activity comes the risk of sexual transmission ofHPV, the most common etiologic factor for cervical cancer. The recommended interval for Pap screening varies from 1 to 3 years. An upper age limit at which screening ceases to be effective is not known.

Colorectal Cancer Fecal occult blood testing, digital rectal examination, rigid and flexible sigmoidoscopy, radiographic barium contrast studies, and colonoscopy have been considered for colorectal cancer screening. Annual fecal occult blood testing using hydrated specimens could reduce colorectal cancer mortality by a third. The sensitivity for fecal occult blood is increased if specimens are rehydrated before testing, but at the cost of lower specificity. The false-positive rate for rehydrated fecal occult blood testing is high; 1 to 5% of persons tested have a positive result. About 2 to 10% of those with occult blood in the stool have cancer, and 20 to 30% have adenomas. The high false-positive rate of fecal occult blood testing dramatically increases the number of colonoscopies performed.

Two case-control studies suggest that regular screening of people over 50 with sigmoidoscopy decreases mortality. These types of studies are prone to selection biases. A quarter to a third of polyps can be discovered with the rigid sigmoidoscope; half are found with a 35-cm flexible scope, and two-thirds to three-quarters are found with a 60-cm scope. Diagnosis of polyposis by sigmoidoscopy should lead to evaluation of the entire colon with colonoscopy and/or barium enema. The most efficient interval for screening sigmoidoscopy is unknown. Case-control studies suggest that testing at intervals of up to 9 years may confer benefit. Most authorities feel that full colonoscopy is too cumbersome and invasive for widespread use as a screening tool in standard-risk populations. It may be suitable for subjects at extremely high risk, such as members of families with a genetic predisposition to colorectal cancer. Colonoscopy is accepted in screening persons with inflammatory bowel disease. Data are not available on digital rectal examination or barium enema as colon cancer screening tools, but both are insensitive.

Lung Cancer Screening chest radiographs and sputum cytology have been evaluated as methods for lung cancer screening. No reduction in lung cancer mortality has been found in these studies, although all the controlled trials performed have had low statistical power. Even screening of high-risk subjects (smokers) has not been proved to be beneficial. Spiral computed tomography (CT) can diagnose lung cancers at early stages; however, false-positive rates are high. Ongoing studies are evaluating spiral CT screening.

Ovarian Cancer Adnexal palpation, transvaginal ultrasound, and serum CA-125 determination have been considered for ovarian cancer screening. Adnexal palpation is too insensitive to detect ovarian cancer at an early enough stage to affect mortality substantially. Neither transvaginal ultrasound nor CA-125 screening has been tested in a completed randomized prospective trial. Ovarian cancer screening can lead to an invasive diagnostic workup, which may include laparotomy. In a clinical study, 0.6% of 900 adult women had a serum CA-125 level >35 U/mL. Thus, if 100,000 adult women were screened, 600 would be identified as having a high CA-125. The prevalence of ovarian cancer in the female adult population is approximately 20 per 100,000. Thus, the screening test would identify 600 women who would undergo further evaluation to identify 20 cases of ovarian cancer. Some of these 600 would only be inconvenienced by an ultrasound examination. Others would undergo an exploratory laparotomy. A large proportion of the 20 women identified as having ovarian cancer would have advanced, incurable disease and thus not benefit from screening. AnNIHconsensus conference in 1994 concluded that routine screening for ovarian cancer was not indicated for standard-risk women or those with a single affected family member, but that it might be worthwhile in families with genetic ovarian cancer syndromes.

Prostate Cancer The most common prostate cancer screening modalities are digital rectal examination and assays for serum prostate-specific antigen (PSA). Newer serum tests, such as measurement of the ratio of bound to free serum PSA, have yet to be fully evaluated. An emphasis on PSA screening has caused prostate cancer to become the most common non-skin cancer diagnosed in American males. Screening for this disease is very prone to lead-time bias, length bias, and overdiagnosis, and substantial debate rages among experts on whether it is effective. Some experts are concerned that prostate cancer screening, more than screening for other cancers, may cause net harm. Prostate cancer screening clearly detects many asymptomatic cancers, but the ability to distinguish tumors that are lethal but still curable from those that pose little or no threat to health is limited. Men over age 50 have a very high prevalence of indolent, clinically insignificant prostate cancers. No well-designed trial has demonstrated the true benefit of prostate cancer screening and treatment, but trials are in progress.

The effectiveness of radical prostatectomy, radiation therapy, and other treatments for low-stage prostate cancer is also under study in randomized trials. Definitive treatment of cancers detected by screening may cause morbidity for some men, such as impotence and urinary incontinence, and carries a low but finite risk of death. Pending the completion of ongoing randomized trials comparing usual care to prostate screening and comparing definitive therapy to "watchful waiting," organizations have provided conflicting recommendations on prostate cancer screening (Table 80-4). After a thorough review of the literature, the American Cancer Society and the American Urologic Association changed their guidelines from a recommendation for screening to a recommendation that men be offered screening after being informed of the potential risks and benefits. A man should have a life expectancy of at least 10 years to be eligible for screening.

Endometrial Cancer Transvaginal ultrasound and endometrial sampling have been advocated as screening tests for endometrial cancer. Benefit from routine screening has not been shown. Transvaginal ultrasound and endometrial sampling are indicated for workup of vaginal bleeding in postmenopausal women but are not considered as screening tests in symptomatic women.

Skin Cancer Visual examination of all skin surfaces by the patient or by a health care provider is used in screening for basal and squamous cell cancers and melanoma. No prospective randomized study has been performed to look for a mortality decrease. Observational epidemiologic evidence from Scotland and Australia suggests that screening programs have caused a stage shift in melanomas diagnosed. Screening may reinforce sun avoidance and other skin cancer prevention behaviors.

(Bibliography omitted in Palm version)

Back to Table of Contents

81. CANCER GENETICS - *Francis S. Collins*, *Jeffrey M. Trent*

THE CLONAL NATURE OF CANCER

Nearly all cancers originate from a single cell. While multiple cumulative events are invariably required to move a cell from normal to the transformed phenotype (see below andChap. 82), the origin of tumors from a single clone of cells is a critical discriminating feature between neoplasia and hyperplasia.

CANCER IS A GENETIC DISEASE

Cancer arises because of alterations in DNA that result in unrestrained cellular proliferation. Most of these alterations involve actual sequence changes in the DNA (i.e., mutation). They may arise as a consequence of random replication errors, exposure to carcinogens (e.g., radiation), or faulty DNA repair processes.

While virtually all cancer is genetic, most cancer is not inherited. Certain individuals with cancer have inherited a germline mutation that predisposes them to the cancer, but even in that situation additional somatic mutations are required for a tumor to develop. In a truly sporadic cancer, *all* of the mutations responsible for the malignant phenotype arise somatically. Such a cancer is caused by genetic alterations but has no hereditary implications.

RNA AND RNA TUMOR VIRUSES

Many malignancies in animals are transmissible, and the etiologic agent is frequently a retrovirus, which possesses a single-stranded RNA genome. During the life cycle of the virus, the single-stranded RNA is converted to double-stranded DNA and is inserted at random into the host chromosome. On rare occasions the virus can be remobilized, carrying along with it an adjacent segment of host DNA. Should this host DNA contain a growth-promoting gene, then the retrovirus is potentially transforming. Although efforts to identify retroviruses in human malignancies have mostly been fruitless, retroviruses are implicated in at least one human malignancy. Human T cell lymphotropic virus (HTLV) type I causes adult T cell lymphoma/leukemia, particularly in Japan and the Caribbean (Chap. 191). Unlike animal retroviruses that induce neoplasia, HTLV-I does not contain a growth-promoting transforming oncogene. The tax protein, a 40-kDa molecule encoded in the pX region of the viral genome, induces the activation of a number of genes (including some promoting growth) through interactions with *rel* family and CREB (cyclic AMP response element binding protein) family transcription factors.

DNA tumor viruses are more commonly involved in human malignancy. Human papilloma viruses (especially types 16 and 18) cause cervical cancer (Chap. 188), and both hepatitis B and hepatitis C viruses have been implicated in hepatocellular carcinoma (Chap. 297) . In addition, the Epstein-Barr virus, a herpesvirus that causes a mild illness in children but infectious mononucleosis in nonimmune adolescents and adults, causes Burkitt's lymphoma in Africa, nasopharyngeal carcinoma in Asia, and lymphomas in the setting of immune deficiency (Chap. 184).

GENERAL CLASSES OF CANCER GENES

In 1914 Boveri hypothesized that cells become malignant either because of overactivation of a gene that promotes cell division or because of loss of function of a gene that normally restrains growth. This hypothesis is largely correct, although defects in DNA repair genes are also involved. Genes that promote normal cell growth are referred to as *protooncogenes*, and activation of such genes by point mutation, amplification, or dysregulation converts them to *oncogenes*.

Genes that normally restrain growth are called *tumor suppressors* (use of the alternative designation of anti-oncogenes is to be discouraged), and unregulated cell growth arises if their function is lost. The diploid nature of mammalian cells allows certain predictions about the consequences of somatic mutations of tumor suppressor genes. Loss of one allele is unlikely to have significant consequences in most instances, as the remaining normal allele is usually sufficient for normal function. Thus, most cells of an individual with an inherited loss of function of one tumor suppressor allele are functionally normal. Only the rare cell that loses or develops a mutation in the remaining normal copy will exhibit uncontrolled growth. This model correctly predicts that the inheritance pattern of cancer in a family with a tumor suppressor gene mutation will be expressed as an autosomal dominant trait, though the cellular mechanism is recessive.

The third category of genes that contribute to malignancy is the DNA repair genes. Every cell division involves the copying of 6 billion base pairs (bp) of DNA. DNA polymerase has a finite error rate, and many environmental influences can damage DNA. As a consequence, repair systems are essential to protect the integrity of the genome. When the repair systems themselves are faulty, either on the basis of inherited or acquired mutation, the rate of accumulation of mutations throughout the genome rises as cell divisions occur. To the extent that these mutations involve oncogenes and tumor suppressor genes, the likelihood of developing malignancy increases.

MENDELIAN CANCER SYNDROMES

Roughly 100 syndromes of familial cancer have been reported, though many are rare. The majority are inherited as autosomal dominant traits, although some of those associated with DNA repair abnormalities (xeroderma pigmentosum, Fanconi anemia, ataxia telangiectasia) are autosomal recessives. Most of the genes responsible for the dominantly inherited cancer syndromes are tumor suppressor genes (Table 81-1). The hallmarks of a tumor suppressor gene are as follows: (1) the germline mutation that affects one allele generally causes a loss of function; (2) tumors also show loss of the second normal allele as a result of a somatic mutation; and (3) often the *normal* function of the gene is to suppress unrestrained cellular growth or to promote differentiation.

The retinoblastoma gene (*RB*) is a paradigm of such a tumor suppressor gene. In a pedigree showing dominant inheritance of susceptibility to retinoblastoma, a loss-of-function germline mutation occurs in one allele of the *RB* gene on chromosome 13. Analysis of the DNA from the tumors invariably shows that the wild-type allele has also been lost by one of several possible mechanisms (Fig. 81-1). However, not all retinoblastoma tumors arise in the context of a strong family history. Sporadic retinoblastoma, which is usually unilateral and on average occurs at a slightly older age than familial retinoblastoma, is usually a consequence of somatic mutation in both alleles of the *RB* gene without any germline predisposition.

Another tumor suppressor gene is the p53 gene on chromosome 17p, which is frequently altered in solid tumors. p53 is somewhat unusual for a tumor suppressor gene in that missense mutations that produce a dominant negative protein product may also be growth-promoting, so that not all alterations obliterate function. Mutations in p53 are found in nearly half of human tumors. Germline mutations in p53 have dramatic consequences, resulting in a phenotype known as the *Li-Fraumeni syndrome*, where affected individuals may develop a variety of sarcomas, brain tumors, and leukemia. Figure 81-2 illustrates a typical pedigree of this devastating disorder.

In many instances the discovery of genes responsible for familial cancer syndromes has provided insight into the normal control of cell growth. For instance, in type I neurofibromatosis -- one of the more common dominant disorders of humans -- positional cloning efforts uncovered a previously unknown gene on chromosome 17q that, when mutated, produces a clinical phenotype of cafe au lait spots, neurofibromas, Lisch nodules of the iris, and a predisposition to neurofibrosarcoma and glioma. The responsible gene, which (like many of the genes inTable 81-1) has a close homologue in yeast, is neurofibromin (*NF1*), a critical participant in the regulation of the protooncogene *ras*. As shown inFig. 81-3, the NF1 protein is a GTPase-activating protein (GAP) that normally acts to convert *ras* from its active, growth-promoting, GTP-bound form to its inactive, GDP-bound form. Loss of both copies of *NF1* (one copy by inheritance, one by somatic mutation) thus renders a cell vulnerable to overgrowth, since *ras* is left in the "on" position.

While most autosomal dominant inherited cancer syndromes are due to mutations in tumor suppressor genes, there are a few interesting exceptions. Multiple endocrine neoplasia type II -- a dominant disorder characterized by pituitary adenomas, medullary carcinoma of the thyroid, and (in some pedigrees) pheochromocytoma -- is due to gain-of-function mutations in the protooncogene *ret* on chromosome 10. Interestingly, loss-of-function mutations in *ret* cause a totally different phenotype, Hirschsprung's disease (aganglionic megacolon) (Chaps. 339 and289).

Dominantly inherited colon cancer is sometimes associated with familial polyposis, which is usually due to mutations in the adenomatous polyposis coli (*APC*) tumor suppressor gene on chromosome 5 (Table 81-1). However, in most colon cancer families affected individuals do not have familial polyposis, but instead the cancer arises from normal-appearing epithelium. Hereditary nonpolyposis colon cancer (HNPCC, or Lynch's syndrome) is commonly defined as the occurrence of colon cancer in at least three individuals over at least two generations and with at least one individual diagnosed under the age of 50. As many as 1 in 200 individuals in the general population may have HNPCC, although this number is somewhat controversial. Most HNPCC is due to mutations in one of four DNA mismatch repair genes (Table 81-2). All four of these genes are components of a repair system that is normally responsible for correcting errors in freshly copied DNA. Tumors in patients with HNPCC are characterized by profound genomic instability, especially for short repeated sequences called *microsatellites*. Figure 81-4 shows an example of the instability in allele sizes for dinucleotide repeats in the cancers in HNPCC. The unstable phenotype [sometimes referred to as the "mutator" phenotype, or the "RER$_{+}$" (replication error) phenotype] probably requires loss of both copies of the particular mismatch repair gene (one inherited, one somatic), so that the mechanism is similar to that typical of a tumor suppressor gene.

MORE COMPLEX INHERITED FORMS OF CANCER

While the Mendelian forms of cancer described above have taught us much about mechanisms of cellular growth control, most forms of cancer do not follow such simple patterns of inheritance. In many instances (e.g., lung cancer), a strong environmental contribution is at work, but even in such circumstances some individuals may be genetically more susceptible to developing cancer given the appropriate exposure.

In the case of breast and ovarian cancer, circumstantial evidence indicates that a subset of affected individuals (5 to 10%) might be accounted for by dominantly inherited high-penetrance susceptibility genes; two of these genes have been identified by positional cloning. *BRCA1*, located on chromosome 17, is capable when mutated of producing a high risk (up to 85% lifelong) of breast cancer and also of ovarian cancer (50% lifelong risk). Roughly 1 in 500 women carries a germline *BRCA1* mutation, often giving rise to a strong family history. Men with *BRCA1* mutations may have a modestly increased risk of prostate cancer. An array of mutational heterogeneity has been described for *BRCA1* (Fig. 81-5), as is often the case for genetic disorders. An exception is the Ashkenazi Jewish population, where 1 in 100 individuals carries a particular 2-bp deletion (denoted 185delAG) of *BRCA1*, apparently as a consequence of descent from a common ancestor.

Mutations in another gene on chromosome 13, *BRCA2*, also confer a high risk of breast cancer (and a somewhat lower risk of ovarian cancer); men with *BRCA2* mutations are prone to develop breast cancer. The frequency of *BRCA2* mutations is estimated to be about half that of *BRCA1*. About 1% of Ashkenazi Jews again have a common mutation: 6174delT.

What then of the 90 to 95% of breast cancers that arise in individuals without germline alterations in *BRCA1* or *BRCA2*? Hereditary factors may still contribute to a significant fraction of these, but those factors must be weaker and therefore more difficult to discern.

GENETIC TESTING AND COUNSELING FOR CANCER SUSCEPTIBILITY

The discovery of genes like *RB*, p53, *NF1*, *ret*, theHNPCCmismatch repair genes, *BRCA1*, and *BRCA2* raises the possibility of DNA analysis to predict risk of cancer. There are many complexities associated with such testing. First, one must know the sensitivity and specificity of the test; the mutational heterogeneity for each of these genes constitutes a considerable technical challenge, as it is often necessary to sample every nucleotide of the coding region, the splice junctions, and the promoter to identify most mutations. False-positive results -- i.e., sequence alterations that turn out to be benign polymorphic variants (allelic variations) rather than disease-causing mutations -- can present a thorny problem. Unless proven interventions are available and the test is sensitive, specific, and relatively inexpensive, it will be inappropriate to offer such tests to the general population; the number of false-positive tests will exceed the number of true positives and a great deal of anxiety and expense will be incurred evaluating persons who are not at an increased risk. Generally, therefore, such testing is not considered except for individuals of higher-than-normal risk, usually on the basis of their family history. In deciding whether to offer such testing, it is critical to determine whether evidence exists for effective interventions to reduce the risk of cancer in those found to be at high risk. If such interventions do not exist (as is the case for Li-Fraumeni syndrome), then the value of the information is limited, and the major negative psychological consequences of this information must be seriously considered.

For conditions such as colon and breast cancer, prophylactic measures exist (total colectomy and bilateral mastectomy, respectively), but these prophylactic measures are more radical and potentially disfiguring than the surgical procedures that would be used to treat the patient if the malignancies actually occurred (segmental bowel resection and lumpectomy, respectively). Other potential negative consequences of a positive genetic test include insurance and employment discrimination. One can still argue, of course, that a close relative of an individual known to carry a mutation in a cancer-causing gene is already sensitized to his or her personal risk of cancer, and that a test establishing that an individual at risk does *not* harbor the mutation can be quite useful. Testing should never be undertaken, however, without a full consideration of how the individual will handle a positive as well as a negative result.

Despite these caveats, genetic testing for some cancer syndromes already appears to have greater benefits than risks, and in those situations it is reasonable to offer testing to individuals at high risk. This would include conditions such as multiple endocrine neoplasia type 2 (Chap. 339) and von Hippel-Lindau disease (Chap. 370). More in the gray zone, although potentially applicable to much larger numbers of individuals, are tests for *BRCA1*, *BRCA2*, and theHNPCCgenes. More research is urgently needed in those situations to determine the effectiveness of various interventions (life-style, diet, surveillance, or surgery). Until those answers are available, such testing should be offered only as part of a research protocol. As more susceptibility genes are identified, better answers become available about the effectiveness of interventions, and health insurance discrimination is legislatively prohibited, genetic testing will move into the mainstream of medicine. Every physician of the future will need to have the skills of a genetic counselor.

ACQUIRED MUTATIONS IN CANCER

The identification of mutations in the germline of patients with heritable cancers means that the alteration is present in every cell of the body. However, in most cancers a normal cell becomes a malignant cell by a series of mutations that arise not in the germline but in somatic cells. Usually mutations must occur in several genes to give rise to neoplasia. The underlying questions are "how many mutations cause a cancer?" and "what specific genes are affected?" rather than whether or not mutational events cause cancer.

While answers to these questions are not available for every human malignancy, advances in molecular and cellular biology and epidemiologic analyses of human and experimental cancers are providing insights in cancer causation.Table 81-3 summarizes evidence from several lines of investigation pointing to a mutational basis for cancer causation. One particularly striking feature is the fact that the overall incidence of cancer increases as the fourth to sixth power of age for most malignancies (Fig. 81-6*A*). For some tumors, the shape of the age-incidence curve suggests heterogeneity in molecular mechanisms. For example, Hodgkin's disease has a bimodal age distribution, suggesting that two etiologically (and therefore mutationally) distinct forms of this disease may exist (Fig. 81-6*B*).

MULTISTEP BASIS OF CANCER

From 5 to 10 accumulated mutations are thought to be necessary for a cell to move from the normal to the fully malignant phenotype. At each step the mutated cell may gain a slight growth advantage, so that it is increased in its representation relative to its neighbors. Figure 81-7, a representation of a lineage diagram hypothesized by Peter Nowell, illustrates how a single cell, afflicted with progressive alterations in tumor suppressor genes and protooncogenes, can develop into a clonal malignancy.

We are beginning to understand the precise nature of the genetic alterations responsible for some malignancies and to get a sense of the order in which they occur. Perhaps the best studied example is colon cancer, where an analysis of DNA from tissues extending from normal epithelium through adenoma to carcinoma have identified some of the genes mutated along the way (Fig. 81-8). However, the order of mutational events is far from uniform, and the diagram inFig. 81-8 should be considered a generalization and not a defined pathway. Similar data are being accumulated for other malignancies.

MECHANISMS OF SOMATIC MUTATION OF ONCOGENES IN MALIGNANCY

Cellular protooncogenes, their necessity and importance in normal cell growth, and their responsibility for transformation-associated change after removal of normal growth controls are discussed in Chap. 82. Mechanisms that upregulate (or activate) cellular protooncogenes can be grouped into three broad areas: point mutations, DNA amplification, and chromosome rearrangements.

Point Mutation One protooncogene that is commonly activated in solid tumors by point mutation is a member of the *ras* family of oncogenes; these were initially cloned from human bladder carcinoma cells and are critical regulators of normal and aberrant cell growth (Fig. 81-3). Mutations in one of the *ras* genes (H-*ras*, K-*ras*, or N-*ras*) are present in up to 85% of pancreatic cancers and 15% of all human cancers. In studies of K-*ras* (particularly in lung and colon cancer), the mutational spectrum of this gene has been identified. Remarkably, and in contrast to the diversity of mutations observed in the *BRCA1* gene (Fig. 81-5), most of these activated genes contain point mutations in codons 12 or 61 (which convey resistance to the inactivating action of GAP). The specificity of this pattern of mutation means that it has potential value in diagnostic or prognostic studies of cancer. For K-*ras*, mutations may be a useful prognostic marker in lung cancer, but for most other cancers (including pancreas and colon cancer) no prognostic utility has been demonstrated. This is in part because *ras* mutations occur early in colon cancer (Fig. 81-8), being common in precancerous lesions of the bowel.

DNA Amplification The second mechanism for activation of oncogene overexpression is DNA sequence amplification. This increase in DNA sequence copy number may cause cytologically recognizable chromosome alterations referred to as *homogeneous staining regions* (HSRs), if integrated within chromosomes, or *double minutes* (dmins), if extrachromosomal in nature (Fig. 81-9).

The recognition of DNA amplification was greatly facilitated by the development of a procedure based on dual-color fluorescence in situ hybridization (FISH) called *comparative genomic hybridization* (CGH). DNA from tumor and normal cells is labeled with different fluorescent reporter molecules and then hybridized to normal metaphase chromosomes. Regions of duplications and deletions within tumor DNA are then demonstrated as quantifiable alterations in signal intensity at particular sites. With this technique the entire genome can be surveyed for gains and losses of DNA sequences, thus pinpointing chromosomal regions likely to contain genes important in the development or progression of cancer.

Numerous genes are known to be amplified in human malignancies. Several genes, including N-*myc* were identified because they were present within the amplified DNA sequences of a tumor and had homology to known oncogenes. Because the region amplified often extends to hundreds of thousands of base pairs, more than one oncogene may be amplified in some cancers (particularly sarcomas). Genes simultaneously amplified in many cases include *MDM2*, *GLI*, *CDK4*, *SAS*, and others implicated in cellular growth control. The clinical implications of gene amplification have been explored for some cancers [most notably *ERBB2 (HER-2/neu)* in breast cancer and N-*myc* in neuroblastoma]; demonstration of amplification of a cellular gene is usually a predictor of poor prognosis. Once a patient has been exposed to the selective effects of chemotherapy, gene amplification may lead to drug resistance. Amplification of the dihydrofolate reductase gene may follow clinical exposure to methotrexate, a drug that inhibits the activity of the enzyme.

Chromosomal Alterations in Human Cancer Chromosomal alterations provide important clues to the genetic changes in cancers. To date, most chromosome analyses have been performed on hematopoietic cancers, although solid tumors may also have translocations. The breakpoints of several recurring chromosome abnormalities often occur at the sites of cellular protooncogenes. Translocations are particularly common in lymphoid tumors, probably because these cell types normally rearrange DNA to generate antigen receptors. Indeed, antigen receptor genes are commonly involved in the translocations, implying that an imperfect regulation of receptor gene rearrangement may be involved in the pathogenesis. An example is Burkitt's lymphoma, a B-cell tumor characterized by a reciprocal translocation between chromosomes 8 and 14. Molecular analysis of Burkitt's lymphomas demonstrated that the breakpoints occurred within or near the *myc* locus on chromosome 8 and within the immunoglobulin heavy chain locus on chromosome 14, resulting in the transcriptional activation of *myc*. Enhancer activation by translocation, although not universal, appears to play an important role in malignant progression.

Chromosome rearrangements can lead to the abnormal overexpression of a transcription factor that performs its normal function and turns on growth-related genes. The translocation may create a chimeric transcription factor that has altered function. For example, the t(15;17) of acute promyelocytic leukemia produces a retinoic acid receptor with an abnormal cell distribution that inhibits differentiation. Gene rearrangements most commonly involve transcription factors, but other components of signaling pathways may also be involved.

The first reproducible chromosome abnormality detected in human malignancy was the Philadelphia chromosome in chronic myelogenous leukemia (CML). This cytogenetic abnormality is generated by reciprocal translocation involving the *ABL* oncogene, a tyrosine kinase on chromosome 9 being placed in proximity to the *BCR* (breakpoint cluster region) on chromosome 22. Figure 81-10 illustrates the generation of the translocation and its protein product. The consequence of expression of the *BCR-ABL* gene product is the activation of signal transduction pathways, leading to cell growth independent of normal external growth factor signals.

In addition to transcription factors and signal transduction molecules, translocations may involve the overexpression of cell cycle regulatory proteins, such as cyclins, and of proteins that regulate cell death, such as bcl2. Altering control of expression of cell cycle regulatory proteins can lead to aberrant cell cycle control. The overexpression of bcl-2 can prevent the death of a cell that has endured enough genetic damage to cause its death. If such a cell survives to receive additional genetic damage, a tumor can develop. Table 81-4 lists representative examples of recurring chromosome alterations in malignancy and the associated gene(s) rearranged or dysregulated by the chromosomal change.

Technical obstacles have slowed the identification of recurring chromosome abnormalities in human solid tumors (particularly carcinomas) because of the complexity of chromosome alterations in such tumors, in contrast to the solitary, often reciprocal, nature of chromosome rearrangements in hematopoietic malignancies.

EPIGENETIC REGULATION OF GENE EXPRESSION AND CANCER

The term *epigenetic* refers to mechanisms of gene regulation independent of DNA sequence. The inactivation of the second X chromosome in female cells is an example of an epigenetic mechanism that prevents gene expression from the inactivated chromosome. During embryologic development, entire regions of chromosomes from one parent are silenced and gene expression proceeds from the chromosome of the other parent. For most genes, expression occurs from both parental alleles or randomly from one parent or the other. The preferential expression of a particular gene exclusively from the allele contributed by one parent is called *parental imprinting* and is thought to be regulated by the methylation of the silenced allele.

The role of epigenetic control mechanisms in the development of human cancer is unclear. However, a general decrease in the level of DNA methylation has been noted as a common change in cancers. In addition, the loss of imprinting of the normally silent maternal allele of the insulin-like growth factor II gene at chromosome 11p15.5 has been implicated in some cases of the rare pediatric malignancy Wilms' tumor. The loss of imprinting may result in the overexpression of the growth factor and a predisposition to malignant transformation.

THE FUTURE

The real challenge in oncology is to convert the growing molecular understanding of cancer into clinical advances, particularly the development of new therapies. One can anticipate that in the coming years the molecular analysis of mutations in tumors will allow stratification of malignancies into more precise subgroups than is currently possible by histologic classification, including subgroups with particularly good or bad prognoses or that have a lower or higher likelihood of responding to a particular therapy. Some of this information is already accumulating, but usually only one or two genes are assessed; the promise of the future is to obtain a detailed molecular "fingerprint" of every tumor in order to provide the maximum information about its biology and response to therapy. The application of techniques for assessing global gene expression (cDNA microarrays, serial analysis of gene expression, or SAGE, and others) is leading to novel ways of looking at cancer that are considerably more discriminating than light microscopy, the gold standard of medical practice. The National Cancer Institute in conjunction with the National Center for Biotechnology Information have undertaken the Cancer Genome Anatomy Project (CGAP) (http://www.ncbi.nlm.nih.gov/ncicgap/) to collect data on the differences in gene expression between normal and malignant tissues and make it available on the Internet.

Genetics will also influence cancer prevention and early detection. The ability to identify cancer susceptibility genes presages a new era of cancer prevention, if the potential risks of such testing can be surmounted. Currently, most cancer early detection strategies (such as mammography, stool occult blood testing, or digital rectal examination) are applied to population groups. The ability to identify the individuals at highest risk and to focus preventive medicine efforts accordingly may be both better received by patients and more cost effective. Early detection strategies will be even more effective if we can develop the ability to identify very small numbers of malignant or premalignant cells at a time when the risk of metastasis is still very low.

More importantly, detailed molecular information about the regulation of the cell cycle and the interplay of tumor suppressor genes and proto-oncogenes that control it may lead to new effective therapies, based on pathophysiology rather than empiricism. Whether such strategies will rely on drugs of the traditional types or will be based on more novel strategies such as gene therapy or immunotherapy is hard to predict.

(Bibliography omitted in Palm version)

Back to Table of Contents

82. CELL BIOLOGY OF CANCER - *Robert G. Fenton*, *Dan L. Longo*

Two characteristic features define a cancer: cell growth not regulated by external signals (i.e., autonomous) and the capacity to invade tissues and metastasize to and colonize distant sites (Chap. 83). The first of these features, the uncontrolled growth of abnormal cells, is a property of all neoplasms, or new growths. A neoplasm may be benign or malignant. If invasion, the second cardinal feature of cancer, is present, the neoplasm is malignant. Cancer is a synonym for *malignant neoplasm*. Cancers of epithelial tissues are called *carcinomas*; cancers of nonepithelial (mesenchymal) tissues are called *sarcomas*.

Cancer is a genetic disease, but the level of its expression is the single cell. Although some forms of cancer are heritable, most mutations occur in somatic cells and are caused by intrinsic errors in DNA replication or are induced by carcinogen exposure. A single genetic lesion is usually not sufficient to induce neoplastic transformation of a cell. The malignant phenotype is acquired only after several (5 to 10) mutations lead to derangements in a variety of gene products. Each genetic alteration may cause phenotypic changes typified by the progression in epithelial tissues from hyperplasia to adenoma to dysplasia to carcinoma in situ to invasive carcinoma. Cells have evolved mechanisms to resist neoplastic transformation (see below).

The>200 discrete cell types in the body are not equally susceptible to developing cancer. Some cells, such as cardiac myocytes, sensory receptor cells for light and sound, and lens fibers, persist throughout life without dividing or being replaced. Neoplasia in such tissues is exceedingly rare. Most differentiated tissues undergo constant renewal characterized by cell death and replacement.

In tissues with rapid turnover, such as skin, bone marrow, and gut, an individual cell is on one of two largely mutually exclusive paths: division or differentiation. Cells capable of dividing are undifferentiated (stem cells), whereas terminally differentiated cells are unable to divide. Stem cells produce daughter cells that can either become new stem cells (thus replenishing the stem cell compartment) or undergo terminal differentiation, depending on the circumstances and the environmental signals. Stem cells are distinguished from differentiating cells by different patterns of gene expression. Gene expression is the product of the tissue-specific programming interacting with environmental factors such as cell-to-cell contact; interactions with extracellular matrix; endocrine hormones; paracrine growth and differentiation factors; and stresses such as heat, oxidation, irradiation, and physical distortion or traction.

Cancer is most common in tissues with rapid turnover, especially those exposed to environmental carcinogens and whose proliferation is regulated by hormones. The most common genetic changes involve the activation of proto-oncogenes or the inactivation of tumor suppressor genes (Chap. 81). Although genetic damage is nearly universal in human cancer, cells with neoplastic features can be generated in vitro without genetic damage. Removal and in vitro culture of cells from the epiblast of a murine embryo lead to the uncontrolled proliferation of the cells and the generation of a teratocarcinoma cell line capable of producing tumors when inoculated into animals. The removal of these normal embryonic cells from their normal environment leads to uncontrolled growth. However, if the teratocarcinoma cells are reinjected into an early embryo, under the inductive influence of their normal neighbors they can differentiate into normal organs and tissues appropriate for the location where they are injected.

Thus, environmental factors exert potent effects on the gene expression of target cells. The panoply of signals received by a particular cell leads to the activation of particular sets of transcription factors. The pattern of gene expression determines whether a cell will divide, differentiate, or die.

PRINCIPLES OF CELL CYCLE REGULATION

The mechanism of cell division is substantially the same in all dividing cells and has been conserved throughout evolution. The process assures that the cell accurately duplicates its contents, especially its chromosomes. The cell cycle is divided into four phases. During M phase, the replicated chromosomes are separated and packaged into two new nuclei by mitosis and the cytoplasm is divided between the two daughter cells by cytokinesis. The other three phases of the cell cycle are called *interphase*: G1 (gap 1), a period of growth during which the cell determines its readiness to commit to DNA synthesis; S (DNA synthesis), during which the genetic material is replicated and no re-replication is permitted; and G2 (gap 2), during which the fidelity of DNA replication is determined and errors are corrected.

During S phase, DNA synthesis begins with the unfolding of chromatin and nucleosome complexes rendering DNA accessible for the addition of DNA helicase and single-strand binding proteins that help open the double helix. Replication origins are spaced roughly 100,000 nucleotide pairs apart throughout the genome. DNA polymerase and DNA primase attach to these sites and catalyze the polymerization of the DNA at a rate of about 50 nucleotides per second. DNA polymerased catalyzes leading-strand synthesis, while DNA polymerase a uses DNA primase-generated Okazaki fragments for lagging-strand synthesis. Topoisomerase I nicks DNA, relieving torsional tension of the replicating helix; topoisomerase II introduces double strand breaks to avoid DNA tangles. Topoisomerases are targets of many chemotherapeutic drugs. Once a DNA segment is replicated, chromatin is reassembled, and replication origins are relicensed by binding of specific proteins that prevent re-replication until the next S phase. Although this system for replication is efficient and accurate, occasional mistakes are made, and these are repaired by a variety of mechanisms. In some cancers, the mismatch repair mechanisms are defective and errors increase by 3 to 4 logs, greatly increasing the likelihood of mutations in growth regulatory genes in daughter cells.

DNA polymerase is unable to replicate the end of a DNA chain completely, resulting in loss of DNA with each replication. This problem has been solved by a mechanism that replicates tandem repeats of a six-nucleotide sequence (GGGTTA) to the ends of each chromosome. These repeated sequences are called *telomeres* and are replicated through an RNA-dependent DNA polymerase called *telomerase*. Normal somatic cells do not express telomerase, and the replicative lifetime of such cells is limited to approximately 30 cell divisions due to the progressive loss of telomere repeats; the limit imposed on somatic cell division is called the *Hayflick limit*, at which time replicative senescence occurs. Germ cells do express telomerase and have a long (possibly unlimited) replicative lifetime. The aberrant expression of telomerase in cancer cells is thought to be a component of the neoplastic process, assuring that the cell will be able to undergo many divisions without inducing senescence or genetic catastophe. Inhibition of telomerase activity in cancer cells could have antitumor effects.

The cell cycle transitions between G1 and S and between G2 and M are tightly regulated to ensure that cells are prepared to divide and to minimize errors in the replication process. Checkpoints in G1 and G2 determine whether a cell will enter S or M phase, respectively; these checkpoints are regulated by serine/threonine protein kinases (cyclin-dependent kinases, or cdk) and kinase-associated proteins called *cyclins*. The enzymatic activity of each cdk is determined by its association with a cyclin and its phosphorylation state. There are at least seven cdk family members, and a like number of cyclins, which generate a group of cdk/cyclin complexes with differing substrate specificities and times of action during the cell cycle. Cyclin expression varies with the cell cycle, and the synthesis of these proteins is transcriptionally regulated and their degradation is mediated by ubiquitin conjugation and destruction in proteasomes.

The cyclin B/cdc2 complex (also called *mitosis promoting factor*, or MPF) is the primary regulator of transition from G2 to M phase. It is activated by acdk-activating kinase (CAK) and a phosphatase (cdc25c) that removes inhibitory phosphates. The cdc25C is the target of a DNA damage-induced kinase that inhibits its activity. DNA damage leads to phosphorylation of cdc25C and its transport out of the nucleus, away from cyclin B/cdc2. This prevents entry into M phase until DNA damage is repaired. The regulated movement of molecules into and out of the nucleus is a common control mechanism in signal transduction. Some of the substrates of cyclin B/cdc2 are defined; its phosphorylation of histone H1, nuclear lamins, and microtubule-associated proteins facilitates chromosome condensation, nuclear membrane breakdown, and spindle formation, respectively.

The checkpoint regulating transition from G1 to S is frequently disrupted in cancer. The product of the retinoblastoma tumor suppressor gene, the nuclear phosphoprotein Rb, governs the key transition referred to as the *reaction point*. A second pathway regulated by the p53 tumor suppressor interacts with the Rb pathway to ensure that cell proliferation can safely take place. Rb and p53 are inactivated by products encoded by DNA tumor viruses, including SV40 large T antigen, adenovirus E1A and E1B, and human papillomavirus E6 and E7. The Rb and p53 pathways each include other oncogenes and tumor suppressors that are frequently disrupted in cancer (Table 82-1).

Regulation of passage through the restriction point is complex. In early G1, Rb is hypophosphorylated and in a complex with the E2F transcription factor. This nuclear complex binds to the promoters of genes required for cell cycle progression and inhibits their expression. However, in mid and late G1, Rb becomes phosphorylated (at ~10 sites) by the sequential activity of the cyclin D/cdk4 and cyclin E/cdk2 complexes. Hyperphosphorylated Rb releases E2F, thus relieving transcriptional repression, and heterodimers of E2F and DP1 transcription factor family members activate several genes required for S phase progression, including dihydrofolate reluctase, thymidine kinase, DNA polymerase, and cdc2 (Fig. 82-1). In addition to its role in growth regulation, Rb is required for the in vitro differentiation of muscle cells and adipocytes. Cell cycle control and induction of differentiation are functions of Rb that contribute to tumor suppression.

The activity ofcdk is also regulated by cdk inhibitors (cdki). These low-molecular-weight proteins are divided into two families: the Cip/Kip family encoding p21$_{Cip1/Waf1}$, p27$_{Kip1}$, p57$_{Kip2}$, which inhibit cdk activity broadly, and the Ink4 family encoding p16$_{INK4a}$, p15$_{INK4b}$, p18$_{INK4c}$, and p19$_{INK4d}$, which block cyclin D/cdk4 activity and inhibit Rb phosphorylation and G1/S transition. p21 is induced by p53 in response to DNA damage, causing G1 arrest to permit DNA repair. If DNA damage is too great, a cell suicide pathway is induced to eliminate cells that may be dysfunctional (see below).

Thecdki can be induced by growth inhibitors such as transforming growth factor (TGF)b and can be inhibited by growth factors such as interleukin (IL) 2. Genetic alterations in cdki, especially p16 and p15, occur with high frequency in certain tumors. Alterations at the p16 locus on chromosome 9p21 have been detected in 75% of pancreatic cancers; 40 to 70% of glioblastomas; 50% of esophageal cancers; and about 20% of non-small cell lung cancers, soft tissue sarcomas, and bladder cancers. Mutations in p16 account for half of familial melanomas. Some tumors fail to express cdki because the genes are methylated, an epigenetic mechanism for blocking transcription. Rb pathway regulation is also circumvented by overexpression of cyclin D1 [breast cancer and the t(11;14) in mantle cell lymphoma] and by mutations in cdk4 that abrogate p16 binding.

Whereas Rb, cyclin D,cdk4, and p16 are commonly altered in cancer, E2F overexpression or p21 mutations have not yet been seen. Additional study may reveal why some components of the system are susceptible to alterations and other components are not.

p53, the "guardian of the genome," is a transcription factor that is not usually called upon to act in the course of normal replication. Levels of p53 are normally kept low by its association with mdm2, which binds p53 and shuttles it out of the nucleus for degradation. However, with DNA damage, p53 is phosphorylated by the ataxia telangiectasia gene product ATM, it dissociates from mouse double minute 2 (mdm2), and its destruction is slowed, leading to increased levels. Also, p53 influences transcription to either halt cell cycle progression (e.g., through induction of p21 expression to inhibitcdkactivity) to permit repair of the DNA or, if the damage is too great, to initiate cell suicide (*apoptosis*). p53-induced genes involved in apoptosis include death receptors (DR5) and death-inducing members of the Bcl-2 family. p53 also induces expression of mdm2, thus down-regulating its own activity.

Inducers of p53 include hypoxia, DNA damage, ribonucleotide depletion, and telomere shortening. Dysregulated activity of oncogenes such as *myc*, which promote aberrant G1/S transition, induces p53-mediated apoptosis. A second product of the Ink4a locus is p14$_{ARF}$, encoded by an alternative reading frame (ARF) from p16. Levels of ARF are upregulated by *myc* and E2F. ARF binds to mdm2/p53 complexes and rescues p53 from the inhibitory effects of mdm2 with subsequent activation of p53-induced genes. This oncogene checkpoint leads to the death of renegade cells that attempt to traverse the restriction point without the right signals.

Mutation in p53 is the most common genetic alteration found in human cancer (>50%) and is the causative lesion in Li-Fraumeni familial cancer sydrome. In tumors, usually one p53 allele on chromosome 17p is deleted and the other is mutated. The mutations often involve the region between codons 120 and 290, the portion of the gene specifying the site of p53 involved in transcription. Some environmental agents cause mutations at specific sites. In 81% of hepatomas in persons from developing countries, codon 249 is mutated due to exposure to the carcinogen, aflatoxin. Codon 249 mutations occur in only 11% of hepatomas in persons from industrialized countries where aflatoxin exposure is low. Inactivation of the p53 pathway compromises cell cycle arrest, inhibits apoptosis induced by DNA damage or oncogene activation, and predisposes cells to chromosome instability.

Regardless of the pathogenesis of the tumor, most have some mechanism(s) to bypass the G1 checkpoint, avoid activation of cell suicide pathways, and propagate cells with damaged DNA. Table 82-1 summarizes some of the changes in cell cycle regulators detected in human cancers.

SIGNALING FROM OUTSIDE THE CELL TO THE NUCLEUS

The behavior of cells in the body is tightly regulated by environmental signals. The ability of a cell to respond to a specific set of signals determines whether the cell will live or die, differentiate, proliferate, or remain quiescent. In normal cells and tissues, coordinated action such as wound healing or the inflammatory response is regulated by signaling pathways that convert extracellular signals into the performance of specialized action in the responding cells. In cancer cells, the process of invading and metastasizing is influenced by signal transduction pathways activated by paracrine and autocrine factors.

The coupling of extracellular signals to cell response varies for different receptor and signaling systems. The binding of a growth factor [e.g., epidermal growth factor (EGF)] to its receptor on the cell surface produces measurable changes in the cell within seconds and elicits a sequence of events that may last for days. Rapid responses are elicited by changes in ion flux, phosphorylation events, lipid metabolism, and production of second messengers. Long-term responses are mediated by the transfer of signaling information from the receptor to the nucleus, where alterations in the pattern of gene expression result in phenotypic change. Signal transduction comprises the mechanisms by which information received at the plasma membrane is imparted to the nuclear transcriptional machinery and other cell functions. Many signal transduction pathways are perturbed in cancer cells. There are three families of cell surface receptors: ion channel-linked receptors. G protein-linked receptors, and enzyme-linked receptors. Although ion channel-linked receptors are a component of growth-related activation in many cell types, they are primarily involved in neurotransmitter signal transduction and are somewhat less important in the pathogenesis of neoplasia than the other two types and will not be discussed further.

G PROTEIN-LINKED RECEPTORS

The G protein-linked receptors traverse the plasma membrane seven times (serpentine receptors). They do not induce covalent modification of their substrates, as do enzyme-linked receptors, but generate second messenger molecules such as cyclic AMP, cyclic GMP, and calcium to activate downstream processes. Upon ligand binding, these receptors activate trimeric G proteins inducing the release of Ga and Gbg subunits, each of which elicits downstream signals. The process is terminated by GTP hydrolysis by Ga subunits. The roles of G protein signaling pathways in human cancer include certain endocrine tumors whose cells of origin depend on cyclic AMP for growth. About half of growth hormone-secreting pituitary tumors encode mutant Ga subunits that are defective in GTPase activity and are constitutively activated even in the absence of ligand. These Ga subunits bind to and stimulate the activity of adenyl cyclase, leading to unregulated synthesis of cyclic AMP. Cyclic AMP binds to the repressor subunit of protein kinase A thus activating the kinase, which enters the nucleus and phosphorylates CREB (cyclic AMP response element binding protein), a transcription factor that activates genes required for proliferation of the cancer cells. Growth stimulatory Ga mutations have also been described in adrenal cortical tumors and endocrine tumors of the ovary. Factors involved in normal cell and tumor cell migration also stimulate cells through G protein-coupled receptors.

ENZYME-LINKED RECEPTORS

There are at least five classes of enzyme-linked receptors: receptor guanylyl cyclases, receptor tyrosine kinases, tyrosine kinase-associated receptors, receptor tyrosine phosphatases, and receptor serine/threonine kinases. The atrial natriuretic peptide receptor is a receptor guanylyl cyclase. Some disease manifestations in cancer may be related to atrial natriuretic peptide activity (such as hyponatremia), but little is known about this receptor class. Receptor phosphatases are not known to be involved in cancer. The other classes of enzyme-linked receptors are better defined and play a more important role in cancer.

Receptor Tyrosine Kinases The receptors for most growth factors are transmembrane tyrosine kinase receptors, including platelet-derived growth factor (PDGF), fibroblast growth factors (FGFs),EGF, heregulin, insulin, insulin-like growth factors (IGF) I and II, nerve growth factor (NGF), stem cell factor, vascular endothelial growth factor, macrophage colony stimulating factors (CSF), and others. Much of what we know about receptor tyrosine kinases and the events that follow their ligation emerged from the study of the proliferation-inducing altered forms of the normal cellular genes (proto-oncogenes) that are the cancer-causing genes (oncogenes) in animal retroviruses. Although downstream events vary with the receptor/ligand combinations, the activation of the receptor follows a typical pattern. Ligand binding induces dimerization or oligomerization of receptor subunits, which activates tyrosine kinase activity and causes autophosphorylation of specific tyrosine residues in the cytoplasmic domain of the receptor. Phosphorylated tyrosine residues on the receptors or on associated adaptor proteins form docking sights for other signal transduction molecules that contain one or more *src-homology region 2*, or SH2, domains, named because the sequence was first identified in the *src* nonreceptor tyrosine kinase. Phosphorylation of tyrosine residues provides a unique amplification signal because of the rapid and specific binding of SH2 domains to p-Tyr, although p-Tyr comprises only 0.05% of total cell phosphoamino acid. These associations via SH2 domains trigger subsequent events (Fig. 82-2). Signaling is terminated by the action of p-Tyr-specific phosphatases.

Protein domain interactions between evolutionarily conserved motifs play critical roles in all forms of signal transduction, ranging from tyrosine kinase pathways, death-inducing molecules, and the association of transcription complexes on gene regulatory regions. The most common docking mechanisms are based on recognition of particular protein sequences; the SH2 domains recognize p-Tyr-containing sequences with specificity conferred by surrounding amino acid residues, the SH3 domains dock with proline-rich sequences, and the pleckstrin homology domains [pleckstrin is a major protein kinase C (PKC) substrate in platelets] lead to associations with phosphatidylinositol lipids phosphorylated in the 3 position by phosphatidylinositol 3-kinase (PI3K; see below). Some molecules that do not have docking domains are brought into association with the receptors through the activity of adaptor proteins that are composed of docking domains only. Thus, the nucleotide exchange factor son of sevenless (SOS; named for its role in *Drosophila* eye development) is brought close to the membrane to activate Ras through its association with the adaptor protein grb2 (identified because it "grabbed" p-Tyr-containing proteins).

Receptor tyrosine kinases activate many signaling pathways including phospholipase C-g, which hydrolyzes phosphoinoside 4,5-bisphosphate (PIP2) into diacylglycerol (DAG) and inositol triphosphate (IP3). DAG together with calcium ion activatesPKC, a family of serine/threonine kinases with different activation requirements, subcellular locations, and substrates in different cell types. PKC is the target of tumor-promoting phorbol esters, and its activation can influence cell proliferation, differentiation, and tumorigenesis. IP3 induces the release of intracellular calcium, which binds to calmodulin, a protein that regulates the activity of many enzymes, including phosphatases. Calcium fluxes within cells can be short or prolonged, and the duration has profound physiologic effects. PI3K is a lipid kinase that generates PI(3,4)P2 and PI(3,4,5)P3, membrane lipids that act as binding sites for proteins containing PH motifs. Such proteins include Akt serine/threonine kinase, which is implicated in activating survival pathways in many cells. Src family tyrosine kinases bind to p-Tyr on activated receptors and amplify signaling information by phosphorylation of distinct protein substrates within the cell. Src activity is required for G1 progression in some cells through its induction of the transcription factor c-myc. Another consequence of receptor tyrosine kinase activation is stimulation of the Ras/MAP (mitogen-activated protein) kinase pathway that leads to activation of a number of transcription factors that regulate proliferation, differentiation, and cell survival. This pathway is frequently abnormal in cancer cells.

Ras is a 21-kDa member of a large family of proteins, including rho, rac, rab, and others, that regulate cytoskeletal changes, vesicular and nuclear transport, and proliferation and that share sequence homology with the Ga subunit of G protein-linked receptors. Ras is attached to the inner cell membrane through an isoprenyl lipid group added after translation by the enzyme farnesyl transferase. If the lipid group is not added, Ras does not localize to the membrane and cannot function normally. In unstimulated cells, Ras is bound to GDP and is inactive. Following receptor tyrosine kinase activation, the guanine nucleotide exchange factorSOS is brought to the membrane by its association with grb2. SOS removes GDP from Ras and adds GTP. GTP-bound Ras then activates a cascade ending withMAPkinase, which migrates to the nucleus and phosphorylates (activates) a number of transcription factors (Fig. 82-3). The kinetics of MAP kinase activity are critical: in PC12 rat pheochromocytoma cells, stimulation withEGFresults in transient stimulation of MAP kinase activity, retention of MAP kinase in the cytoplasm, and cell proliferation; stimulation of PC12 cells with NGF induces sustained activation of MAP kinase, nuclear translocation of MAP kinase, and neuronal differentiation.

Genetic defects leading to increased signaling from receptor tyrosine kinase-linked pathways are important in the etiology and progression of human cancer. About 30% of human cancers (especially pancreatic, lung, and colon adenocarcinomas) have mutated *Ras*. The mutations usually involve codons 12, 13, or 61 and result in a Ras protein that fails to hydrolyze its bound GTP and is thus constitutively active. In the hereditary disorder, neurofibromatosis, a mutation in the gene that encodes neurofibromin, a GTPase activating protein (GAP), inhibits its ability to inactivate Ras by converting the GTP to GDP. Constitutively activated Ras results in the unregulated activity of the signaling pathways downstream of Ras, including the MAP kinase pathway and activation of PI3K. Some epithelial cancers overexpress one or more members of the receptor tyrosine kinase family.EGFreceptors,IGF-I receptors, and HER-2/*neu* are overexpressed in lung, bladder, breast, head and neck, and ovarian cancers. Mutations within the Ret tyrosine kinase receptor lead to constitutive receptor dimerization and kinase activation and are responsible for the dominant inherited cancer syndromes multiple endocrine neoplasia (MEN) type 2A and type 2B and familial medullary thyroid carcinoma. Autocrine and paracrine sources of the relevant growth factors have been noted in some cases.

Tyrosine Kinase-Associated Receptors The receptors for growth hormone, prolactin, erythropoietin, thrombopoietin, most interleukins, granulocyteCSF, granulocyte-macrophage CSF, interferon-a, interferon-g, and many other cytokines are members of the tyrosine kinase-associated receptor family. These single-transmembrane receptors contain ligand-specific subunits and shared signaling subunits. Ligand binding induces the activation of receptor-associated tyrosine kinases. Three families of kinases are known to be associated with this class of receptors: *src* family (*src*, *yes*, *fgr*, *fyn*, *lck*, *lyn*, *hck*, *blk*, and counting), *syk* family (*syk*, ZAP-70), and Janus family (JAK1, JAK2, JAK3, Tyk2). The Janus family kinases have receptor sites that act as docking sites for SH2-containing transcription factors called STATs (signal transducers and activators of transcription). Tyrosine phosphorylation of STATs induces their dimerization by SH2-p-Tyr association followed by translocation to target genes in the nucleus. A unique feature of JAK/STAT signaling is that the pathway from cell membrane to nucleus is traversed by a single dimeric molecule, as opposed to the cascade of kinase and adaptor molecules associated with membrane tyrosine kinases. The *src* family kinases can associate with receptor tyrosine kinases as well as tyrosine kinase-associated receptors, and, not surprisingly, signal transduction through either receptor class leads to the activation of similar signaling cascades. As a consequence of *src* family activation, *myc* is one of the transcription factors activated. The *syk* family usually activates the *src* family member in the receptor complex.

These receptors are often overexpressed on tumors of hematopoietic origin, and, similar to receptor tyrosine kinases, autocrine or paracrine stimulation may contribute to the neoplastic state of the tumor cell.

Serine/Threonine Kinase Receptors These receptors recognizeTGF-b, bone morphogenetic factors, and other activins as ligands. Ligand binding leads to activation of the receptor kinase activity, but downstream events are not well defined. Bone morphogenetic factors are important in bone formation and in determining ventral vs. dorsal orientation in the developing embryo. TGF-b induces transformation of mesenchymal cells but inhibits the proliferation of most cell types through the induction ofcdki, which block G1 progression in an Rb-dependent manner. Activation of TGF-b receptors leads to the phosphorylation of transcription factors Smad2 and Smad3, which then associate with their obligate partner Smad4. This complex, probably a heterotrimer, translocates into the nucleus where specific genes are activated. The direct path to the nucleus is analogous to the JAK/STATsignaling pathway. Smads bind to specific DNA sequences adjacent to other transcription factor sites in the promoters of target genes, leading to cooperative interaction for gene induction. TGF-binduced genes include plasminogen activator inhibitor-1 (PAI-1), collagenase I, and p15$_{INK4b}$.

Many cancers are resistant to growth inhibition byTGF-b, including leukemias, lymphomas, melanomas, and breast and colon cancers. Colon cancers harboring defects in DNA mismatch repair develop inactivating mutations in the extracellular domain of the TGF-b receptor. In pancreatic and colon cancers, missense mutations and loss of heterozygosity at the DPC4 locus on chromosome 18q21 are frequent. This locus has been found to encode Smad4, and its inactivation blocks TGF-bsignaling. Loss of expression or loss of function of TGF-b receptors occurs in several tumor types including colon cancer and lymphomas.

Nuclear Hormone Receptors Steroids, retinoids, thyroid hormone, vitamin D, and other lipid-soluble hormones diffuse through the plasma membrane and bind to members of the nuclear hormone receptor family. Receptors for these ligands are transcription factors that reside in the nucleus. The hydrophobic nature of the ligands obviates the need for machinery to transduce signals from the cell surface to the cell interior. Steroid hormone receptors are bound as heterodimers to promoter/enhancer elements of genes; in the absence of ligand, these complexes act as transcriptional repressors. Upon ligand binding, conformational changes are induced and the active transcription factor binds to coactivating factors and transcription is induced. Coactivators tend to open chromatin structure by adding acetyl groups to histones. Histone acetylation in nucleosomal complexes permits access of promoter regions to RNA polymerase II. Transcriptional repressor complexes associate with histone deacetylases (HDAC; co-repressor complexes), and chromatin remains condensed. Nuclear hormone-induced pathways affect virtually all biologic processes. Retinoic acid receptors (RAR) provide a clear link to cancer. Retinoids bind receptors composed of an RAR subunit (a, b,g) dimerized with a retinoid-X receptor (RXR) and activate genes that influence differentiation in many cell types.

Acute promyelocytic leukemia (APL) is associated with the t(15;17) translocation, which fuses sequences from a novel gene PML (promyelocytic leukemia) to theRARagene, resulting in expression of a PML-RARa fusion protein. PML-RARa binds to and represses RARa-inducible genes required for myeloid differentiation. Repression is mediated byHDACbinding to PML-RARa. The developmental arrest at the promyelocyte stage of differentiation is associated with unregulated proliferation in these cells. Patients with APL can achieve complete remission with pharmacologic doses of all-trans retinoic acid (tretinoin), the ligand for RARa. Tretinoin induces the release of HDAC, permitting coactivator binding. Drugs that inhibit HDAC activity may provide a therapeutic benefit by activating genes required for the differentiation of cancer cells.

Cell-Cell and Cell-Extracellular Matrix (ECM) Communication In addition to information conveyed by soluble mediators, cell surface receptors relay important signals between cells, such as contact inhibition, and cell-ECM signals, such as anchorage-dependent growth. In cancer, these highly organized mechanisms of intercellular interaction often become disrupted or are subsumed for the purpose of metastasizing (Chap. 83). Individual cells no longer respond to signals from their neighbors, actin filaments are highly disorganized, and adherens junctions are lost. Normal patterns of growth and differentiation are disrupted, and the potential for metastasis increases.

E-cadherins are integral membrane glycoproteins that mediate calcium-dependent homophilic adhesion as the major component of adherens junctions between epithelial cells. E-cadherin cytoplasmic domains bind complexes containing a- andb-catenins, which are structurally linked to the cytoskeleton (actin cables and intermediate filaments).b-Catenin that is not sequestered in E-cadherin complexes is rapidly phosphorylated by glycogen synthesis kinase (GSK) 3b in a complex with the APC (adenomatous polyposis coli) gene product (maps to chromosome 5, mutated in familial polyposis) and is degraded by the ubiquitin/proteosome pathway. Degradation ofb-catenin can be blocked by several mechanisms, including mutations that inactivate APC and mutations in serine phosphorylation sites within b-catenin that target it for degradation. Such mutations result in increased free b-catenin, which translocates into the nucleus and binds to members of the T cell factor (TCF) family of transcription factors, influencing the expression of genes such as c-*myc* and cyclin D1 that promote progression through G1. Excess freeb-catenin has been implicated in hereditary and sporadic forms of colon cancer and melanoma. Decreased expression of E-cadherin has been noted in breast, colon, prostate, gastric, and other cancers and is a marker of poor prognosis.

Epithelial cell growth and survival require attachment of cells to components of theECM that compose basement membranes, including collagen, fibronectin, vitronectin, and laminin. The integrin family of transmembrane receptors is composed ofa and b subunits that adhere to the ECM and convey information to cytoplasmic membrane-associated structures called *focal adhesions*. The complexes, whose assembly is mediated by the Rho and Rac GTPases, are sites of attachment of actin cables but are also active in cell signaling through their association with focal adhesion kinase (FAK) and Src tyrosine kinases. Integrin-ECM interactions lead to activation of the Ras/MAPkinase, PI3K, and phospholipase C-g pathways. Detachment of epithelial and endothelial cells from ECM induces their death by a form of programmed cell death called *anoikis* (Greek, "homeless"). This molecular safeguard prevents abnormal spread of cells. Invasive cancers often avoid anoikis by activating Ras or Src, which allow anchorage-independent growth of cells by activation of Akt kinase.

REGULATION OF GENE TRANSCRIPTION

One consequence of signal transduction is the activation of sequence-specific transcription factors that regulate gene expression. Whether a cell proliferates, differentiates, or undergoes apoptosis is regulated by gene products made in response to physiologic stimuli. For some transcription factors, the ligand goes directly to the nucleus where they reside (nuclear hormone receptors). For others, activated kinases enter the nucleus and phosphorylate factors already bound to DNA (MAPkinase and

AP-1 transcription factor). Some transcription factors are activated in the cytoplasm and translocate to the nucleus (STATs). NF-kB is held in the cytoplasm by the negative regulator IkB, which is phosphorylated and degraded as a consequence of signal transduction. NF-kB is then released from IkB, and NF-kB translocates to the nucleus.

Transcription factors recognize short stretches of DNA of a defined nucleotide sequence 6 to 12 base pairs in length. These recognition sites may be upstream or downstream of the transcription start site [the TATA box where the first subunit of the transcription machinery, transcription factor IID (TFIID), binds]. Transcription factors may affect transcription at sites remote from the start site by looping out large intervening DNA sequences.

Transcription factors contain specific amino acid sequences capable of recognizing the DNA sequence and usually form one of several structural motifs: helix-turn-helix, homeodomain, zinc finger, leucine zipper, and helix-loop-helix are all used as DNA-binding motifs or mediate dimerization of factors required for DNA binding. Transcription factors function in one or more of several ways. They can physically bend the DNA to permit the ordered addition of the components of the transcription machinery. Activated transcription factors bind to coactivator proteins in complexes that encode enzymatic activity leading to the acetylation of histones. This alters nucleosomal conformation and increases accessibility of DNA to transcription proteins. Transcription factors can inhibit transcription by blocking binding of a positive transcription factor or preventing the assembly of a transcription complex. They can form complexes with co-repressors and deacetylate histones. Promoters can also be made inaccessible by methylation of cytosine- and guanosine-rich sequences near promoters. The complex interaction between positive and negative transcription factors dictates the level of gene transcription. Individual genes may have³20 sites for transcription factor binding. The pattern of gene expression is determined by which factors are expressed in a given cell type.

Most genes are regulated at multiple levels, though transcription initiation is the dominant control point. The von Hippel-Lindau gene on chromosome 3p (a tumor suppressor gene involved in the pathogenesis of renal cell cancer) appears to act by inhibiting the elongation of an RNA chain after transcription initiation. A message may be spliced alternatively and encode different proteins in different cells. Transport of the message from the nucleus to the cytoplasm may be altered. Messenger RNA turnover may be accelerated. Some proteins, such as apoferritin and thymidylate synthase, regulate the translation of their own messages (and perhaps other messages) by binding to mRNA and preventing initiation of protein synthesis. Thus, there are many levels at which gene expression may be influenced.

Some transcription factors were identified because of their transforming effects when their genes, usually in mutated form, were incorporated into animal retroviral genomes; *myc*, *rel*, *fos*, *jun*, and others are examples of proto-oncogene transcription factors that are overexpressed in certain cancers and that contribute to the malignant phenotype of tumor cells. The mutated oncogenes are often more resistant to protein degradation and have a longer half-life than the normal cellular counterpart. Transcription factors with unusual properties may be generated by chromosome translocation that produces a chimeric protein. Novel genes may be activated that promote proliferation and inhibit apoptosis. Usually the genetic changes lead to inhibition of normal lineage-specific differentiation, resistance to apoptosis, and proliferation.

REGULATION OF CELL DEATH

The homeostasis of adult organisms requires a balance between the generation of new cells and the death of old cells. Some cells die when their telomeres no longer protect the integrity of DNA replication. Some cells die when they have sustained sufficient hypoxic, heat, oxidative, or ultraviolet irradiation damage that cannot be repaired. A cell can be killed if it becomes infected with a virus or other intracellular pathogen that destroys the cell or is recognized by the host's lymphocytes, which kill the infected cell. Multicellular organisms are models of cellular cooperation; some cells die to preserve the rest of the organism.

Genetic damage to growth-regulating genes of stem cells could be catastrophic; however, single genetic events such as activation of *myc* expression or loss of the Rb checkpoint often lead to the death of the cell by apoptosis. Apoptosis is a form of cell death initiated by extracellular or intracellular signals in which enzymes are activated to degrade nuclear DNA by making intranucleosomal cuts, causing the cell to shrink and finally break up. The core apoptosis machinery consists of a family of specialized proteases called *caspases* (they contain a *c*ysteine at their active site and cleave substrates after *asp*artic acid residues). Like coagulation and complement systems, caspases exist as proenzymes with minimal enzymatic activity that can be rapidly induced by activators, and each enzyme acts to activate the next enzyme in the cascade. Key targets include DNA (chromatin degraded into nucleosomal multimers), nuclear lamins (nucleus shrinks and fragments), cytoskeletal regulatory proteins, DNA repair enzymes, and others. The cell shrinks, its chromatin fragments, and the cell breaks apart forming apoptotic bodies.

The latent activity of caspases is tightly regulated to prevent the death of normal cells. Assembly of initiator caspases into active complexes occurs by two main mechanisms. Members of the tumor necrosis factor (TNF) receptor superfamily, including Fas (CD95), and DR4 and DR5 death receptors encode transmembrane proteins whose cytoplasmic domains encode protein association or docking domains, called *death domains* and *death effector domains* (DED). Ligand binding induces dimerization of death domains with recruitment to the membrane of an adaptor signaling protein called *Fas-associated death domain* (FADD). FADD forms a complex with procaspase 8 mediated by DED interactions; caspase 8 is activated by self-cleavage. Caspase 8 then cleaves effector caspase 3 into active heterodimeric subunits, and death ensues (Fig. 82-4). Regulation of this pathway occurs at the level of expression of Fas receptor and ligand and the secretion of death-inducing cytokines, TNF and *t*umor necrosis factor-*r*elated *a*poptosis-*i*nducing *l*igand (TRAIL) (ligand for DR4 and -5).

The second pathway of caspase activation encompasses responses to a greater variety of noxious signals including DNA damage, growth factor deprivation, reactive oxygen damage, and heat stress. The mitochondrion plays a key role in this pathway as the storehouse of protein cofactors required for the activation of caspases. Damage within the cell is detected by the mitochondria by unknown mechanisms. The mitochondria then lose membrane potential and release cytochrome c, which forms a complex with apoptosis-activating factor (Apaf) 1. This complex binds to procaspase 9 via a caspase recruitment domain (CARD), and caspase 9 is activated. Caspase 9 then cleaves effector caspases and induces cell death. The release of cytochrome c from the mitochondria appears to be regulated by bc12.

The *bcl2* gene was discovered as the chromosome 18q contribution to the t(14;18) translocation in follicular lymphoma. The gene did not transform cells but prolonged the life of cells destined to die, greatly increasing a pool of cells available for subsequent genetic mutations. Members of the *bcl2* family fall into two groups: *bcl2* and *bcl*-$X_L$prevent cell death, whereas *bax*, *bad*, *bak*, and others promote cell death. The *bcl2* family members associate as homodimers or heterodimers; the combinatorial effects of the various dimers allow a fine level of control over cell survival. When any of the death promoters exist as homo- or heterodimers, the cell dies by apoptosis. When a death promoter heterodimerizes with either *bcl2* or *bcl*-X_L, cell death is prevented. Thus, the relative amounts of different *bcl2* family members determine whether a cell will survive potentially damaging insults. Furthermore, phosphorylation of *bcl2* family members by cellular kinases can alter the biologic activity of individual members, altering the balance in favor of death or survival.

In addition to its presumed role in the etiology of follicular lymphoma, *bcl2* is expressed in a number of cancers. It prevents the normal p53-mediated destruction of cells with damaged DNA and also appears to prevent the death of cells severely damaged by cancer chemotherapy. In addition, *bcl2* mediates drug resistance and contributes to neoplasia in a novel way, preventing the death that would normally eliminate the damaged cell, rather than promoting aberrant cell growth. Strategies to overcome *bcl2* function might well make available therapies more effective.

The apoptotic machinery is subject to regulation by multiple signal transduction pathways, and many of these are subverted in cancer cells to shift the balance toward survival of the malignant clone. In addition to *bcl2*, two other important links have been established between growth factor signaling and survival pathways. Activation of P13K by tyrosine kinases leads to activation of the serine/threonine kinase Akt. Akt directly promotes cell survival by phosphorylation of Bad and procaspase 9, inhibiting their apoptotic functions. Cancer cells can usurp the activity of Akt; in some cases, cells expressing a mutated Ras oncogene or increased levels of tyrosine kinase receptors (e.g., HER-2/*neu*) have upregulated the Akt pathway. An alternative genetic lesion leading to increased Akt kinase activity results in the loss of the PTEN tumor suppressor, a lipid phosphatase that normally downregulates the P13K pathway by dephosphorylating lipid second messengers. Another important pathway usurped by cancer cells involves NF-kB activation. NF-kB induces expression of the inhibitor of apoptosis (IAP) family of genes whose products inhibit caspase activity; one such, survivin, is expressed in lymphomas and other tumors.

Thus, cancer becomes more adaptive to its host from genetic events that alter apoptosis. Stimulation of proliferation or prevention of death can be complementary targets for cell transformation. Apoptosis pathways are important targets for treatment. Paclitaxel and other microtubule inhibitors induce the phosphorylation and inactivation of *bcl2*. One could make bone marrow cells highly resistant to chemotherapy-induced death by expressing a form of *bcl2* that lacks the loop domain in the BH1 region, as this is the site that is phosphorylated to inhibit *bcl2* function. Tumor cells and normal cells express DR4 and -5 receptors; however, tumor cells fail to express decoy receptors that protect normal cells from the DR4 and -5 ligand, TRAIL. Thus, therapies directed at DR4 or -5 may be tumor selective.

CELL BIOLOGY AND CANCER

For a cancer to arise, mutations must occur that affect a variety of pathways. Often the G1 cell cycle checkpoint is affected. Apoptosis is averted by mutations in the p53 pathway or by other mechanisms. The expression of telomerase is a common feature in cancers. Overexpression of growth factors and their receptors is frequently detected. Activation of the *Ras* proto-oncogene or other changes leading to a constitutively activeMAPkinase cascade are common. Changes in cytoskeleton and responsiveness to contact-mediated growth inhibition are frequent in cancer cells. Usually when a mutation occurs in one component of a signaling pathway, other mutations are seen in other pathways rather than in another component of the same pathway. The high level of mutability of cancer cells facilitates adaptation to the environment, including the development of resistance to anticancer drugs. As tumors progress, they acquire the ability to secrete proteases that aid in the escape from local barriers so that they may metastasize (Chap. 83). Discrete steps in tumor progression lead to the production of factors by the tumor cells that permit neovascularization to supply nutrients to the growing tumor. Other mutations allow the tumor to escape immune surveillance mechanisms; for example, some tumors downregulate expression of class I major histocompatibility complex antigens so that they become invisible to T cells. The wide range of changes that must occur in a single cell to permit the behavior associated with a malignant neoplasm makes it clear why carcinogenesis is a multistep process and why human cancers may have 10 or more genetic lesions that account for the biology.

The characteristic of cancer cells that has dominated clinical thinking is their uncontrolled proliferation. However, the growth fraction of most human cancers is usually not higher than the growth fraction of normal gut epithelia or normal bone marrow, and most human tumor explants are difficult to propagate for long periods of time in culture. Cancer cell lines immortal in vitro may have additional genetic lesions that permit their growth in vitro. Naturally occurring tumors growing in vivo show a Gompertzian or exponential decline in their growth fraction because the daughter cells of a division are not uniformly capable of further division. The accumulation of genetic damage, poor oxygen or nutrient supply, and other unknown factors contribute to the senescence of some tumor cells, so that by the time a tumor becomes clinically apparent at a tumor burden of 10^8 to 10^9cells, most of the proliferative capability of the tumor is finished. Often by this time, more malignant and highly selected clonal derivatives of the tumor have metastasized to other sites where new tumor deposits with more aggressive characteristics are formed. Thus, cancer cells can be viewed as having lost the altruism that usually characterizes cell behavior in multicellular organisms. Cancer cells operate under natural selection imposed by a hostile environment. Ironically, the more successful they are at achieving independence from environmental influences, the more assured is the destruction of their host and ultimately themselves.

Many potential therapeutic agents are in clinical development based on our concepts of tumor cell biology. They include the development of growth factor and growth factor receptor antagonists; inhibitors of phosphoryl transfer to block key kinases; selective inhibitors ofPKC, P13K, phospholipase C, and other targets; farnesyl transferase inhibitors that block the insertion of *Ras* into the membrane; mutant versions of proteins such as *Ras* and p53 that may make the cell vulnerable to immunologic attack if employed as a vaccine; and inhibitors of angiogenesis or the steps in metastasis that may limit tumor growth and prevent its spread. However, it seems unlikely that a single target will be the highly sought-after point of vulnerability. More likely, combinations of inhibitors will be required to improve antitumor effects. For example, the combination of chemotherapy and antibody toEGFreceptors appears to produce greater antitumor effects than the sum of the effects produced individually.

(Bibliography omitted in Palm version)

Back to Table of Contents

83. ANGIOGENESIS - *Judah Folkman*

Virtually every cell in the body lives adjacent to a capillary blood vessel, or at least no further than the mean oxygen diffusion distance of 100 to 200 um. Some cell types, such as beta cells in the pancreatic islets, fat cells, and skeletal muscle cells, are surrounded by at least two capillaries (Fig. 83-1). Capillaries of 8 to 20 um diameter are lined by a single layer of endothelial cells. These cells cover ~1000 m_2, an area the size of a tennis court. The length of capillary tubing in 1 mm_3 of human heart muscle is ~2500 mm, and 1 kg of fat contains ~3500 m of capillaries. During normal conditions vascular endothelial cell proliferation is barely detectable -- <0.01% of endothelial cells are in cycle. Endothelial cell turnover is >1000 days and in retinal vasculature may be >5000 days. In contrast, in the normal adult bone marrow, ~6 billion cell divisions occur per hour and the turnover time is ~5 days, (i.e., the time during which bone marrow is completely replaced). Endothelial cells can emerge from their resting state and proliferate as rapidly as bone marrow cells during formation of new capillaries. This process is called *angiogenesis* and leads to *neovascularization*.

Physiologic angiogenesis is tightly regulated and of limited duration. It is essential to reproduction and embryonic development. During postnatal and adult life, angiogenesis in wound repair and in exercised muscle is restricted to days or weeks.

Pathologic angiogenesis, in contrast, is usually persistent and unabated. Angiogenesis that continues for months or years supports the growth and progression of solid tumors and leukemias, provides a conduit for the entry of inflammatory cells into sites of chronic inflammation (e.g., Crohn's disease and chronic cystitis), is the most common cause of blindness, destroys cartilage in rheumatoid arthritis, contributes to growth and hemorrhage of atherosclerotic plaques, leads to intraperitoneal bleeding in endometriosis, is the basis of life-threatening hemangiomas of infancy, and permits prostate growth in benign prostatic hypertrophy. These are just a few of the "angiogenic disease processes," which are found in almost all specialties of medicine. Angiogenesis inhibitors are a new *class* of drugs that suppress or reverse the pathologic neovascularization upon which these diseases are dependent.

NEOPLASTIC DISEASE

HISTORIC BACKGROUND

Tumor hyperemia, observed during surgery since the 1870s, was for the next 100 years attributed to simple dilation of existing host vessels. Two reports, in 1939 and 1945, suggested that tumor vascularity was due to the induction of new blood vessels. This idea was dismissed by most investigators. The few who accepted it believed that new vessels were an inflammatory side effect of tumor growth.

In 1971, based on experiments carried out in the 1960s, a hypothesis was proposed that tumor growth could be angiogenesis-dependent, i.e., tumors could recruit their own private blood supply by releasing a diffusible chemical signal that stimulated angiogenesis. Tumor angiogenesis could then be a novel second target for anticancer therapy. These concepts were not accepted at the time. The conventional wisdom was that tumor neovascularization was (1) an inflammatory host response to necrotic tumor cells, (2) a host response detrimental to the tumor, or (3) "established" vasculature that could not regress. From these assumptions most scientists concluded that it was fruitless to attempt to discover an angiogenesis stimulator, to say nothing of discovering angiogenesis inhibitors. Eventual acceptance of the 1971 hypothesis was slow because it would be 2 more years before the first vascular endothelial cells were successfully cultured in vitro, 8 more years before *capillary* endothelial cells could be cultured in vitro, 11 years before the discovery of the first angiogenesis inhibitor, and 13 years before the purification of the first angiogenic protein. By the mid-1980s, after a series of reports from several laboratories demonstrating that tumor growth was angiogenesis-dependent, this hypothesis had been confirmed by genetic methods.

THE ANGIOGENIC SWITCH

The Prevascular Phase Most human tumors arise without angiogenic activity and exist in situ as microscopic-sized lesions of 0.2 to 2 mm diameter for months to years, after which a small percentage may switch to the angiogenic phenotype. Autopsy studies of people who died of trauma but who never had cancer during their lifetime reveal that in women from 40 to 50 years of age, 39% had in situ carcinomas in their breast, but breast cancer is diagnosed in only 1% of women in this age range. In men from age 50 to 70, 46% had in situ prostate cancers at the time of death, but only 1% are diagnosed in this age range during life. In people from age 50 to 70, >98% had small carcinomas of the thyroid (Fig. 83-2), but thyroid cancer is diagnosed in only 0.1% in this age range. In the majority of human tumors, the angiogenic phenotype appears after the malignant phenotype is recognized histologically. However, for certain human tumors (e.g., carcinoma of the cervix), the preneoplastic stage of dysplasia becomes angiogenic before the malignant phenotype is recognized histologically. When a nonangiogenic in situ carcinoma emerges in avascular epidermis or mucosa (e.g., melanoma or breast cancer), it is separated from host vessels by a basement membrane (Fig. 83-3). If a nonangiogenic tumor emerges in the midst of a vascularized tissue (e.g., an islet cell carcinoma), it may form an in situ microcylinder of tumor cells around capillary vessels (called *cooption*).

At the clinical level the angiogenic switch is recognized by expansion of tumor mass to a detectable size, local bleeding, and metastasis. For example, a positive mammogram usually represents a neovascularized tumor -- a non-neovascularized in situ carcinoma is below the detectable limits of mammography. Hematuria in bladder cancer, melena in colorectal cancer, and hemoptysis in lung cancer all result from neovascularized tumors. Tumor cells are not usually shed into the circulation until after neovascularization has occurred. Furthermore, distant metastases themselves cannot be detected until they have "turned on" the angiogenic switch.

At the cellular level at least four mechanisms of the angiogenic switch have been identified in human and mouse tumors: (1) avascular in situ carcinomas can recruit their own blood supply by stimulating neovascularization in an adjacent host vascular bed -- the most common process in human tumors; (2) circulating precursor endothelial cells from bone marrow may incorporate into an angiogenic focus; (3) tumors may induce host fibroblasts and/or macrophages in the tumor bed to overexpress an angiogenic factor [e.g., vascular endothelial growth factor (VEGF)]; and (4) preexisting vessels can be coopted by tumor cells. The angiogenic switch may also include combinations of these mechanisms. Once tumors have switched on angiogenesis, they rarely revert to the nonangiogenic phenotype. Neuroblastoma and retinoblastoma may be exceptions, but spontaneous loss of angiogenic activity (and tumor regression) is rare even in these two tumors. After the angiogenic switch, new microvessels converge on the tiny in situ tumor. Tumor cells grow as microcylinders, or "perivascular cuffs," around each new vessel. One endothelial cell can support from 5 to 100 tumor cells.

At the molecular level, the angiogenic switch operates as a shift in the balance of production by tumor cells of molecules that positively or negatively regulate angiogenesis. The overexpression of positive regulators of angiogenesis and the downregulation of inhibitors of angiogenesis during early tumor development are generally triggered by genetic mutations that control angiogenesis. For example, overexpression of the *ras* oncogene increases production of the angiogenic proteinVEGF, while a mutation in the p53 tumor-suppressor gene or its deletion decreases production of the angiogenesis inhibitor protein, thrombospondin-1. In the normal cell wild-type p53 upregulates thrombospondin-1 and downregulates VEGF.

The angiogenic switch can be further modified by environmental conditions such as hypoxia, endogenous angiogenesis inhibitors, and genetic background of the host.

1. *Hypoxia*: After a tumor has become neovascularized, its continued expansion may lead to increased tissue pressure. This increased interstitial pressure is caused mainly by plasma that leaks from new vessels but is slow to efflux from the tumor because of a dearth of intratumoral lymphatics. Microvessels in the center of the tumor are the first to be compressed, which leads to central necrosis. Hypoxia activates an hypoxia-inducible factor (HIF-1) binding sequence in theVEGFpromoter. This leads to transcription of VEGF mRNA, increased stability of VEGF message, and increased production of VEGF protein beyond what may have been triggered genetically. Tumors, therefore, do not "outgrow their blood supply" but compress it. A counterintuitive lesson is that in situ tumors arising in an avascular compartment (e.g., epidermis) are *not* hypoxic, but larger neovascularized tumors become hypoxic after compressing their blood supply. Low pH and low glucose in a tumor may also upregulate production of angiogenic factors, especially VEGF.

2. *Endogenous angiogenesis inhibitors*: At the molecular level, the angiogenic switch can also be modified by endothelial inhibitors that either circulate [e.g., interferon (IFN) b, platelet factor 4, angiostatin] or are releasable from extracellular matrix [e.g., endostatin, thrombospondin-1, and tissue inhibitors of metalloproteinases (TIMPs)]. Therapeutic administration of an endogenous angiogenesis inhibitor, such as angiostatin or endostatin, can tip the balance of the angiogenic switch so that angiogenic output of a tumor is opposed or abrogated.

3. *Genetic control of host response*: Just as the angiogenic output of a given tumor is governed by oncogenes and tumor-suppressor genes, the angiogenic response of the host is genetically regulated. It is known that hemangiomas predominate in white infants and that ocular neovascularization in macular degeneration is almost never found in black patients. The genes that regulate these effects are not yet known.

Endogenous Angiogenesis Promoters The known endogenous angiogenic promoters are listed inTable 83-1. Virtually all of these proteins are produced by different types of tumors. However, acidic FGF (aFGF), basic FGF (bFGF),VEGF, and angiopoietin-1 and -2 are the most well studied and have been found in a wide variety of human tumors.

Fibroblast Growth Factors aFGF andbFGFstimulate endothelial cell mitosis and migration in vitro and are among the most potent angiogenic proteins in vivo. They have high affinity for heparin and heparan sulfate. They lack a signal sequence for secretion but are stored in extracellular matrix. An unsolved problem is how bFGF is exported from tumor cells in the absence of a signal sequence. Many different cells synthesize bFGF, including tumor cells of the central nervous system, sarcomas, genitourinary tumors, and even endothelial cells in the tumor vasculature. Proteinases and heparanases are thought to mobilize bFGF from the extracellular matrix. Furthermore, some tumors recruit macrophages and activate them to secrete bFGF, while others attract mast cells, which, because of their high heparin content, sequester bFGF. bFGF is not a specific endothelial mitogen but has several cell targets including fibroblasts, smooth-muscle cells, and neurons. However, experimental tumors transfected with bFGF containing an engineered signal sequence stimulate endothelial proliferation almost to the exclusion of smooth-muscle and fibroblast proliferation. This is similar to the process in human tumors. This selective attraction of vascular endothelial cells by bFGF released from a tumor may be explained by the smooth-muscle repellant activity of angiopoietin-2 elaborated from proliferating endothelial cells in a tumor bed (see below). bFGF interferes with adhesion of leukocytes to endothelium; thus, tumors that elaborate bFGF may produce a form of local immunologic tolerance.

Abnormally elevated levels ofbFGF are found in the serum and urine of cancer patients and in the cerebrospinal fluid of patients with different types of brain tumors. High bFGF levels in renal carcinoma correlate with poor outcome. Also, bFGF levels in the urine of children with Wilms' tumor correlate with stage of disease and tumor grade.

Vascular Endothelial Growth Factor/Vascular Permeability Factor The first proposal that tumor angiogenesis is associated with increased microvascular permeability led to the identification of vascular permeability factor (VPF). VPF was subsequently sequenced and shown to be a specific inducer of angiogenesis; it was called *vascular endothelial growth factor*.VEGF is an endothelial cell mitogen and motogen that is angiogenic in vivo. Its permeability effect on capillaries is more potent than histamine and contributes to ascites in ovarian cancer and to edema in brain tumors. Its expression correlates with blood vessel growth during embryogenesis and with angiogenesis in the female reproductive tract and in tumors. VEGF is a 40- to 45-kDa homodimeric protein with a signal sequence secreted by a wide variety of cells and by the majority of human tumor cells. For example,>60% of breast cancers overexpress VEGF. VEGF exists as five different isoforms of 121, 145, 165, 189, and 206 amino acids, of which VEGF$_{165}$is the predominant molecular species produced by a variety of normal and neoplastic cells. Two receptors for VEGF are found mainly on vascular endothelial cells, the 180-kDa fms-like tyrosine kinase (Flt-1) and the 200-kDa human kinase insert domain-containing receptor (KDR) and its mouse homologue, Flk-1. VEGF binds to both receptors, but KDR/Flk-1 transduces the signals for endothelial proliferation and chemotaxis. Other structural homologues of the VEGF family have recently been identified, including VEGF-B, -C, -D, and -E. VEGF-C stimulates lymphatic growth and binds to Flt4, which is preferentially expressed on lymphatic endothelium. Neuropilin-1, a neuronal guidance molecule, is a recently discovered receptor for $VEGF_{165}$. Neuropilin is not a tyrosine kinase receptor and is expressed on nonendothelial cells including tumor cells. This allows VEGF that is synthesized by tumor cells to bind to their surface. Surface-bound VEGF could make endothelial cells chemotactic to tumor cells or it could act in a paracrine manner to mediate cooption of tumor cells around microvessels (Fig. 83-3).

VEGFexpression is upregulated by the *ras* oncogene. The farnesyl transferase inhibitors inhibit *ras* expression. One mechanism of their antitumor effect is to inhibit angiogenesis by inhibiting VEGF expression.

VEGFexpression is inhibited by the von Hippel-Lindau (VHL) protein. The VHL-tumor suppressor gene is inactivated in patients with VHL disease and in most sporadic clear-cell renal carcinomas, which leads to VEGF-mediated angiogenesis. The VHL gene normally suppresses hypoxia-inducible genes including erythropoietin. When VHL is mutated or deleted, these genes are overexpressed even under normoxic conditions. This explains why renal cell carcinomas driven by mutant VHL are associated with a high hematocrit.

Experimental evidence indicates thatbFGFfunction may be in part dependent uponVEGF. bFGF induces the expression of VEGF. The two endothelial mitogens act synergistically to stimulate capillary tube formation in vitro. Systemic administration of a soluble receptor for VEGF (flk-1) completely blocks cornea angiogenesis induced by implanted bFGF. An important implication of these studies is that angiogenesis inhibitors that block VEGF (currently in clinical trial) may also inhibit bFGF.

Angiopoietins Tie2 is a receptor found only on vascular endothelial cells. It is a specific tyrosine kinase whose ligand is angiopoietin-1. Angiopoietin-1 induces endothelial cells to recruit pericytes and smooth-muscle cells [mainly by producing platelet-derived growth factor (PDGF) BB] to become incorporated in the vessel wall. Vessels stimulated by angiopoietin-1 are not leaky and are analogous to new vessels in a healing wound. Angiopoietin-2 blocks the Tie-2 receptor and acts to repel pericytes and smooth muscle. It is produced by vascular endothelium in a tumor bed, but it is unclear how tumor cells mediate this. Nevertheless, tumor vessels remain thin "endothelial-lined tubes" even though some of these microvessels reach the diameter of venules (Fig. 83-3). A key point is that angiopoietin-2 andVEGFtogether increase angiogenesis. However, if VEGF is neutralized or withdrawn, endothelial cells in the absence of perivascular smooth muscle and pericytes undergo apoptosis and new microvessels regress. These differences indicate that angiogenesis in tumors may be more vulnerable to certain angiogenesis inhibitors than angiogenesis in healing wounds. Endostatin inhibits tumor growth in mice without delaying wound healing.

Endogenous Angiogenesis Inhibitors Certain endogenous inhibitors of angiogenesis are known to play a role in the angiogenic switch, including:IFN-aand platelet factor 4, and the class of angiostatic steroids typified by tetrahydrocortisol (Table 83-2).

Thrombospondin-1 The production of thrombospondin-1 has been shown to be inversely related to the ability of a cell line to produce a tumor and vessels in vivo; loss of thrombospondin-1 production allowed non-tumorigenic cells to become tumorigenic. Thrombospondin-1 is regulated by wild-type p53. Loss of p53 function in tumor cells dramatically decreased the level of angiogenesis inhibitor. Restoration of p53 increased the inhibitor and suppressed the angiogenic activity of the tumor cells. The angiogenic switch was controlled by a negative regulator of angiogenesis generated by the tumor. The switch itself was viewed as a result of a shift in the "net balance" of angiogenesis stimulators and inhibitors. This led to the discovery of angiostatin, a second inhibitor found to be involved in the angiogenic switch.

Angiostatin, Endostatin, and Antiangiogenic Antithrombin III Several clinical and experimental observations suggested that certain tumors may produce angiogenesis inhibitors. The removal of certain tumors (e.g., breast carcinomas, colon carcinomas, and osteogenic sarcomas) can be followed by rapid growth of distant metastases. A primary tumor can suppress metastases from a different type of tumor, e.g., a breast cancer can inhibit melanoma metastases. In melanoma, partial spontaneous regression of the primary tumor may be followed by rapid growth of metastases. Regression of small cell lung cancer by ionizing radiation may be followed by rapid growth of distant metastases. If one portion of a primary tumor is removed (e.g., cytoreductive surgery for testicular cancer), the residual tumor increases its rate of expansion. A similar phenomenon is observed in animal tumors, i.e., certain primary tumors inhibit the growth, but not the number, of their own metastases. Surgical removal of a primary tumor increases growth rate of the residual tumors. Many primary tumors can suppress the growth of a second tumor inoculation. This "resistance" to a second tumor challenge is inversely proportional to the size of the tumor inoculum and directly proportional to the size of the first tumor. A threshold size is necessary for the inhibitory effect to occur.

At least three hypotheses have been advanced to explain these diverse observations and experiments: (1) "concomitant immunity" -- a primary tumor induces an immunologic response against a secondary tumor or a metastasis in the same host; (2) depletion of nutrients by the primary tumor; or (3) production of antimitotic factors from the primary tumor that directly inhibit the proliferation of the secondary tumor. However, none of these ideas offers a molecular mechanism to explain all of the experiments cited above, and overall they have not been confirmed. Concomitant immunity has been ruled out as a mechanism because tumors can suppress metastasis in mice with severe combined immunodeficiency (SCID).

Once it was realized that a tumor could generate both positive and negative regulators of angiogenesis, then it also became clear that a primary tumor, while stimulating angiogenesis in its own vascular bed, could possibly inhibit angiogenesis in the vascular bed of a distant metastasis. However, at least two conditions would be necessary: (1) the primary tumor (i.e., the first tumor to grow) would need to generate an angiogenic promoter in excess of an inhibitor in its own local vascular bed, and (2) the putative inhibitor would need to have a longer half-life in the circulation than the angiogenic promoter. Research done over the past decade has identified angiostatin, endostatin, and antiangiogenic antithrombin as negative regulators of angiogenesis.

Lewis lung carcinoma generated angiostatin, a 38-kDa cleavage product of plasminogen. Systemic administration of purified angiostatin completely inhibited growth of metastases, producing dormant tumors of microscopic size (<200 um in diameter) in the lung, and inhibited the growth of primary tumors. Angiostatin is not secreted by tumor cells but is generated through proteolytic cleavage of circulating plasminogen by