Back to Table of Contents

290. ACUTE INTESTINAL OBSTRUCTION - *William Silen*

ETIOLOGY AND CLASSIFICATION

Intestinal obstruction may be *mechanical* or *nonmechanical* (resulting from neuromuscular disturbances that produce either adynamic or dynamic ileus). The causes of mechanical obstruction of the lumen are conveniently divided into (1) lesions *extrinsic* to the intestine, e.g., adhesive bands, internal and external hernias; (2) lesions *intrinsic* to the wall of the intestine, e.g., diverticulitis, carcinoma, regional enteritis; and (3) obturation of the lumen, e.g., gallstone obstruction, intussusception. Clinically, however, it is most useful to consider whether the obstructive mechanism involves the small or large intestine, because the causes, symptoms, and treatments are different (see below). Adhesions and external hernias are the most common causes of obstruction of the small intestine, constituting 70 to 75% of cases of this type. Adhesions, however, almost never produce obstruction of the colon, where carcinoma, sigmoid diverticulitis, and volvulus, in that order, are the most common causes and together account for about 90% of the cases. Primary intestinal pseudoobstruction (Chap. 289) is a chronic motility disorder that frequently mimics mechanical obstruction. Unnecessary operations in such patients should be avoided.

Adynamic ileus is probably the most common overall cause of obstruction. The development of this condition is mediated via the hormonal component of the sympathoadrenal system. Adynamic ileus may occur after any peritoneal insult, and its severity and duration will be dependent to some degree on the type of peritoneal injury. Hydrochloric acid, colonic contents, and pancreatic enzymes are among the most irritating substances, whereas blood and urine are less so. Adynamic ileus occurs to some degree after any abdominal operation. Retroperitoneal hematomas, particularly associated with vertebral fracture, commonly cause severe adynamic ileus, and the latter may occur with other retroperitoneal conditions, such as ureteral calculus or severe pyelonephritis. Thoracic diseases, including lower-lobe pneumonia, fractured ribs, and myocardial infarction, frequently produce adynamic ileus, as do electrolyte disturbances, particularly potassium depletion. Finally, intestinal ischemia, whether the result of vascular occlusion or intestinal distention itself, may perpetuate an adynamic ileus.

Spastic ileus or *dynamic ileus* is very uncommon and results from extreme and prolonged contraction of the intestine. It has been observed in heavy metal poisoning, uremia, porphyria, and extensive intestinal ulcerations.

PATHOPHYSIOLOGY

Distention of the intestine is caused by the accumulation of gas and fluid proximal to and within the obstructed segment. Between 70 and 80% of intestinal gas consists of swallowed air, and because this is composed mainly of nitrogen, which is poorly absorbed from the intestinal lumen, removal of air by continuous gastric suction is a useful adjunct in the treatment of intestinal distention. The accumulation of fluid proximal to the obstructing mechanism results not only from ingested fluid, swallowed saliva, gastric juice, and biliary and pancreatic secretions but also from interference with normal sodium and water transport. During the first 12 to 24 h of obstruction, there is a marked depression of flux from lumen to blood of sodium and consequently water in the distended proximal intestine. After 24 h, there is movement of sodium and water into the lumen, contributing further to the distention and fluid losses. Intraluminal pressure rises from a normal of 2 to 4 cmH_2O to 8 to 10 cmH_2O. During peristalsis, when simple obstruction or a "closed loop" is present, pressures reach 30 to 60 cmH_2O. Closed-loop obstruction of the small intestine results when the lumen is occluded at two points by a single mechanism such as a hernial ring or adhesive band, thus producing a closed loop whose blood supply is often obstructed at the same time. Strangulation of the loop itself is thus common in association with marked distention proximal to the involved loop. A form of closed-loop obstruction is encountered when complete obstruction of the colon exists in the presence of a competent ileocecal valve (85% of individuals). Although the blood supply of the colon is not entrapped within the obstructing mechanism, distention of the cecum is extreme because of its greater diameter (Laplace's law), and impairment of the intramural blood supply is considerable with consequent gangrene of the cecal wall, usually anteriorly. Necrosis of the small intestine may occur by the same mechanism of interference with intramural blood flow when distention is extreme, but this sequence is uncommon in the small intestine. Once impairment of blood supply occurs, bacterial invasion supervenes, and peritonitis develops. The systemic effects of extreme distention include elevation of the diaphragm with restricted ventilation and subsequent atelectasis. Venous return via the inferior vena cava may also be impaired.

The loss of fluids and electrolytes may be extreme and, unless replacement is prompt, leads to hemoconcentration, hypovolemia, renal insufficiency, shock, and death. Vomiting, accumulation of fluids within the lumen by the mechanisms described above, and the sequestration of fluid into the edematous intestinal wall and peritoneal cavity as a result of impairment of venous return from the intestine all contribute to massive loss of fluid and electrolytes, especially potassium. As soon as significant impedance to venous return is present, the intestine becomes severely congested, and blood begins to seep into the intestinal lumen. Blood loss may reach significant levels when long segments of intestine are involved.

SYMPTOMS

Mechanical small-intestinal obstruction is characterized by cramping midabdominal pain, which tends to be more severe the higher the obstruction. The pain occurs in paroxysms, and the patient is relatively comfortable in the intervals between the pains. Audible borborygmi are often noted by the patient simultaneously with the paroxysms of pain. The pain may become less severe as distention progresses, probably because motility is impaired in the edematous intestine. When strangulation is present, the pain is usually more localized and may be steady and severe without a colicky component, a fact that often causes delay in diagnosis of obstruction. Vomiting is almost invariable, and it is earlier and more profuse the higher the obstruction. The vomitus initially contains bile and mucus and remains as such if the obstruction is high in the intestine. With low ileal obstruction, the vomitus becomes feculent, i.e., orange-brown in color with a foul odor, which results from the overgrowth of bacteria proximal to the obstruction. Hiccups (singultus) are common. Obstipation and failure to pass gas by rectum are invariably present when the obstruction is complete, although some stool and gas may be passed spontaneously or after an enema shortly after onset of the complete obstruction. Diarrhea is occasionally observed in partial obstruction. Blood in the stool is rare but does occur in cases of intussusception. Other than some minor but inconsistent differences in pain patterns noted above, the symptoms of strangulating obstructions cannot be distinguished from those of nonstrangulating obstructions.

Mechanical colonic obstruction produces colicky abdominal pain similar in quality to that of small-intestinal obstruction but of much lower intensity. Complaints of pain are occasionally absent in stoic elderly patients. Vomiting occurs late, if at all, particularly if the ileocecal valve is competent. Paradoxically, feculent vomitus is very rare. A history of recent alterations in bowel habits and blood in the stool is common because carcinoma and diverticulitis are the most frequent causes. Constipation becomes progressive, and obstipation with failure to pass gas ensues. Acute symptoms may develop over a period of a week. Cecal volvulus more closely resembles obstruction of the small intestine clinically, whereas patients with sigmoid volvulus more typically have the picture of colonic obstruction in which marked distention predominates, with relatively less pain.

In *adynamic ileus*, colicky pain is absent, and only discomfort from distention is evident. Vomiting may be frequent but is rarely profuse. It usually consists of gastric contents and bile and is almost never feculent. Complete obstipation may or may not occur. Singultus (hiccups) is common.

PHYSICAL FINDINGS

Abdominal distention is the hallmark of all forms of intestinal obstruction. It is least marked in cases of obstruction high in the small intestine and most marked in colonic obstruction. Early, especially in closed-loop strangulating small-bowel obstruction, distention may be barely perceptible or absent. Tenderness and rigidity are usually minimal; the temperature is rarely above 37.8°C (100°F) in nonstrangulating obstruction of the small and large intestine. Contrary to popular belief, the same is true of strangulating obstruction until very late, a fact that has often resulted in unfortunate delay in treatment. Signs and symptoms of shock also occur *very late* in strangulating obstruction. The appearance of shock, tenderness, rigidity, and fever often means that contamination of the peritoneum with infected intestinal content has occurred. Hernial orifices should always be carefully examined for the presence of a mass. The presence of a palpable abdominal mass usually signifies a closed-loop strangulating small-bowel obstruction because the tense fluid-filled loop is the palpable lesion. Auscultation may reveal loud, high-pitched borborygmi coincident with the colicky pain, but this finding is often absent late in strangulating or nonstrangulating obstruction. A quiet abdomen does not eliminate the possibility of obstruction, nor does it necessarily establish the diagnosis of adynamic ileus.

LABORATORY AND X-RAY FINDINGS

Leukocytosis, with shift to the left, usually occurs when strangulation is present, but a normal white blood cell count does not exclude strangulation. Elevation of the serum amylase level is encountered occasionally in all forms of intestinal obstruction, especially the strangulating variety.

The x-ray is extremely valuable but under certain circumstances may also be misleading. In nonstrangulating complete small-bowel obstruction, x-rays are almost completely reliable. Distention of fluid- and gas-filled loops of small intestine usually arranged in a "stepladder" pattern with air-fluid levels and an absence or paucity of colonic gas are pathognomonic (Fig. 290-1). These findings, however, are absent in slightly over half the cases of strangulating small-bowel obstruction, especially early in the disease. A general haze due to peritoneal fluid and sometimes a "coffee bean"-shaped mass are seen in strangulating obstruction. Occasionally, the films are normal, but when symptoms are consistent with obstruction of the small intestine, a normal film should suggest strangulation. In these circumstances, computed tomography may be very useful. Roentgenographic differentiation of partial mechanical small-bowel obstruction from adynamic ileus may be impossible because gas is present in both the small and large intestines; however, colonic distention is usually more prominent in adynamic ileus. A radiopaque dye given by mouth is useful in making this distinction.

Colonic obstruction with a competent ileocecal valve is easily recognized because distention with gas is mainly confined to the colon. Barium enema, sigmoidoscopy, or colonoscopy, depending on the suspected site of obstruction, is usually advisable to determine the nature of the lesion, except when concomitant perforation is suspected, a rare occurrence. Sigmoidoscopy may be therapeutic in cases of sigmoid volvulus. When the ileocecal valve is incompetent, the films resemble those of partial small-bowel obstruction or adynamic ileus, and barium enema or colonoscopy is necessary to establish the correct diagnosis. Barium given by mouth is perfectly safe when obstruction is in the small intestine, since the barium sulfate does not become inspissated in this location. *Barium should never be given by mouth to a patient with possible colonic obstruction* until that possibility has been excluded by barium enema.

TREATMENT

Small-Intestinal Obstruction The overall mortality rate for obstruction of the small intestine is about 10%, even under the most optimal conditions. While the mortality rate for nonstrangulating obstruction is as low as 5 to 8%, that for strangulating obstruction has been reported to be between 20 and 75%. Well over half the deaths from small-bowel obstruction occur in those with strangulation; however, the latter constitute only one-fourth to one-third of the cases. Careful studies indicate that the clinical, laboratory, and x-ray findings are not reliable in distinguishing strangulating from nonstrangulating obstruction when obstruction is complete. Complete obstruction is suggested when passage of gas or stool per rectum has ceased and when gas is absent in the distal intestine by x-ray. Since strangulating small-bowel obstruction is always complete, operation should always be undertaken in such patients after suitable preparation. Before operation, fluid and electrolyte balance should be restored and decompression instituted by means of a nasogastric tube. Replacement of potassium is especially important because intake is nil and losses in vomitus are large. >From 6 to 8 h of preparation may be necessary. During this period, broad-spectrum antibiotics are indicated if strangulation is felt to be likely, but operation should not be delayed unless there is unequivocal clinical and roentgenographic evidence of resolution of the obstruction during the period of preparation. Attempts to pass a long tube into the small intestine usually fail while putting the patient through uncomfortable, unproductive manipulations that delay appropriate fluid replacement and decompression. *There are* *few, if any, indications for the use of a long intestinal tube*. Procrastination of operation because of improvement in well-being of the patient during resuscitation and gastric decompression usually leads to unnecessary and hazardous delay in proper treatment. Purely nonoperative therapy is safe only in the presence of incomplete obstruction and is best utilized in patients with (1) repeated episodes of partial obstruction, (2) recent postoperative partial obstruction, and (3) partial obstruction following a recent episode of diffuse peritonitis.

Colonic Obstruction The mortality rate for colonic obstruction is about 20%. As in small-bowel obstruction, nonoperative treatment is contraindicated unless the obstruction is incomplete. Occasionally, but not always, when the obstruction is incomplete, nonoperative therapy may result in sufficient decompression that a definitive operative procedure can be undertaken at a later date. This can usually be accomplished by discontinuation of all oral intake and perhaps by nasogastric suction, although attempts to decompress a *completely* obstructed colon by intubation are almost invariably futile. A long intestinal tube will not decompress an obstructed colon with a competent ileocecal valve. When obstruction is complete, early operation is mandatory, especially when the ileocecal valve is competent; cecal gangrene is likely if the cecal diameter exceeds 10 cm on plain abdominal film. For obstruction on the left side of the colon, the most common site, preliminary operative decompression by cecostomy or transverse colostomy followed by definitive resection of the primary lesion has been the treatment of choice. Recently, primary resection of obstructing left-sided lesions with on-table washout of the colon has been accomplished safely. For a lesion of the right or transverse colon, primary resection and anastomosis can be performed safely because distention of the ileum with consequent discrepancy in size and hazard in suture are not present.

Adynamic Ileus This type of ileus usually responds to nonoperative continuous decompression and adequate treatment of the primary disease. The prognosis is usually good. Successful decompression of severe colonic ileus has been accomplished by colonoscopy, but this should be avoided if tenderness in the right lower quadrant suggests possible cecal gangrene. Neostigmine is effective in cases of colonic ileus that have not responded to other conservative treatment. Rarely, adynamic colonic distention may become so great that cecostomy is required if cecal gangrene is feared. Spastic ileus usually responds to treatment of the primary disease.

(Bibliography omitted in Palm version)

Back to Table of Contents

291. ACUTE APPENDICITIS - *William Silen*

INCIDENCE AND EPIDEMIOLOGY

The peak incidence of acute appendicitis is in the second and third decades of life; it is relatively rare at the extremes of age. Males and females are equally affected, except between puberty and age 25, when males predominate in a 3:2 ratio. Perforation is more common in infancy and in the aged, during which periods mortality rates are highest. The mortality rate has decreased steadily in Europe and the United States from 8.1 per 100,000 of the population in 1941 to less than 1 per 100,000 in 1970 and subsequently. The absolute incidence of the disease also decreased by about 40% between 1940 and 1960 but since then has remained unchanged. Although various factors such as changing dietary habits, altered intestinal flora, and better nutrition and intake of vitamins have been suggested to explain the reduced incidence, the exact reasons have not been elucidated. The overall incidence of appendicitis is much lower in underdeveloped countries, especially parts of Africa, and in lower socioeconomic groups.

PATHOGENESIS

Luminal obstruction has long been considered the pathogenetic hallmark. However, obstruction can be identified in only 30 to 40% of cases; ulceration of the mucosa is the initial event in the majority. The cause of the ulceration is unknown, although a viral etiology has been postulated. Infection with *Yersinia* organisms may cause the disease, since high complement fixation antibody titers have been found in up to 30% of cases of proven appendicitis. Whether the inflammatory reaction seen with ulceration is sufficient to obstruct the tiny appendiceal lumen even transiently is not clear. Obstruction, when present, is most commonly caused by a fecalith, which results from accumulation and inspissation of fecal matter around vegetable fibers. Enlarged lymphoid follicles associated with viral infections (e.g., measles), inspissated barium, worms (e.g., pinworms, *Ascaris*, and *Taenia*), and tumors (e.g., carcinoid or carcinoma) may also obstruct the lumen. Secretion of mucus distends the organ, which has a capacity of only 0.1 to 0.2 mL, and luminal pressures rise as high as 60 cmH_2O. Luminal bacteria multiply and invade the appendiceal wall as venous engorgement and subsequent arterial compromise result from the high intraluminal pressures. Finally, gangrene and perforation occur. If the process evolves slowly, adjacent organs such as the terminal ileum, cecum, and omentum may wall off the appendiceal area so that a localized abscess will develop, whereas rapid progression of vascular impairment may cause perforation with free access to the peritoneal cavity. Subsequent rupture of primary appendiceal abscesses may produce fistulas between the appendix and bladder, small intestine, sigmoid, or cecum. Occasionally, acute appendicitis may be the first manifestation of Crohn's disease.

While chronic infection of the appendix with tuberculosis, amebiasis, and actinomycosis may occur, a useful clinical aphorism states that *chronic appendiceal inflammation is not usually the cause of prolonged abdominal pain of weeks' or months' duration*. In contrast, recurrent acute appendicitis does occur, often with complete resolution of inflammation and symptoms between attacks. Recurrent acute appendicitis may become more frequent as antibiotics are dispensed more freely and if a long appendiceal stump is left after laparoscopic appendectomy.

CLINICAL MANIFESTATIONS

The history and sequence of symptoms are important diagnostic features of appendicitis. The initial symptom is almost invariably *abdominal pain* of the visceral type, resulting from appendiceal contractions or distention of the lumen. It is usually poorly localized in the periumbilical or epigastric region with an accompanying urge to defecate or pass flatus, neither of which relieves the distress. This visceral pain is mild, often cramping, and rarely catastrophic in nature, usually lasting 4 to 6 h, but it may not be noted by stoic individuals or by some patients during sleep. As inflammation spreads to the parietal peritoneal surfaces, the pain becomes somatic, steady, and more severe, aggravated by motion or cough, and usually located in the *right lower quadrant*. *Anorexia* is nearly universal; a hungry patient does not have acute appendicitis. *Nausea* and *vomiting* occur in 50 to 60% of cases, but vomiting is usually self-limited. The development of nausea and vomiting before the onset of pain is extremely rare. Change in bowel habit is of little diagnostic value, since any or no alteration may be observed, although the presence of diarrhea caused by an inflamed appendix in juxtaposition to the sigmoid may cause serious diagnostic difficulties. Urinary frequency and dysuria occur if the appendix lies adjacent to the bladder. The typical sequence of symptoms (poorly localized periumbilical pain followed by nausea and vomiting with subsequent shift of pain to the right lower quadrant) occurs in only 50 to 60% of patients.

Physical findings vary with time after onset of the illness and according to the location of the appendix, which may be situated deep in the pelvic cul-de-sac; in the right lower quadrant in any relation to the peritoneum, cecum, and small intestine; in the right upper quadrant (especially during pregnancy); or even in the left lower quadrant. *The diagnosis cannot be established unless tenderness can be elicited*. While tenderness is sometimes absent in the early visceral stage of the disease, it ultimately always develops and is found in any location corresponding to the position of the appendix. Abdominal tenderness may be completely absent if a retrocecal or pelvic appendix is present, in which case the sole physical finding may be tenderness in the flank or on rectal or pelvic examination. Percussion, rebound tenderness, and referred rebound tenderness are often, but not invariably, present; they are most likely to be absent early in the illness. Flexion of the right hip and guarded movement by the patient are due to parietal peritoneal involvement. Hyperesthesia of the skin of the right lower quadrant and a positive psoas or obturator sign are often late findings and are rarely of diagnostic value. When the inflamed appendix is in close proximity to the anterior parietal peritoneum, muscular rigidity is present yet is often minimal early.

The temperature is usually normal or slightly elevated [37.2 to 38°C (99 to 100.5°F)], but a temperature> 38.3°C (101°F) should suggest perforation. Tachycardia is commensurate with the elevation of the temperature. Rigidity and tenderness become more marked as the disease progresses to perforation and localized or diffuse peritonitis. Distention is rare unless severe diffuse peritonitis has developed. The disappearance of pain and tenderness just before perforation is extremely unusual. A mass may develop if localized perforation has occurred but usually will not be detectable before 3 days after onset. Earlier presence of a mass suggests carcinoma of the cecum or Crohn's disease. Perforation is rare before 24 h after onset of symptoms, but the rate may be as high as 80% after 48 h.

Diagnosis is based primarily on clinical grounds. Although moderate leukocytosis of 10,000 to 18,000 cells/uL is frequent (with a concomitant left shift), the absence of leukocytosis does not rule out acute appendicitis. Leukocytosis of >20,000 cells/uL suggests probable perforation. Anemia and blood in the stool suggest a primary diagnosis of carcinoma of the cecum, especially in elderly individuals. The urine may contain a few white or red blood cells without bacteria if the appendix lies close to the right ureter or bladder. Urinalysis is most useful in excluding genitourinary conditions that may mimic acute appendicitis.

Radiographs are rarely of value except when an opaque fecalith (5% of patients) is observed in the right lower quadrant (especially in children). Consequently, abdominal films are not routinely obtained unless other conditions such as intestinal obstruction or ureteral calculus may be present. In some patients with recurrent or prolonged symptoms, a careful barium enema or computed tomography (CT) scan may reveal an extrinsic defect on the medial wall of the cecum or a calcified fecalith. The value of CT scan in acute appendicitis is being evaluated. The diagnosis may also be established by the ultrasonic demonstration of an enlarged and thick-walled appendix. Ultrasound is most useful to exclude ovarian cysts, ectopic pregnancy, or tuboovarian abscess.

While the typical historic sequence and physical findings are present in 50 to 60% of cases, a wide variety of atypical patterns of disease are encountered, especially at the age extremes and during pregnancy. Infants under 2 years of age have a 70 to 80% incidence of perforation and generalized peritonitis. Any infant or child with diarrhea, vomiting, and abdominal pain is highly suspect. Fever is much more common in this age group, and abdominal distention is often the only physical finding. In the elderly, pain and tenderness are often blunted, and thus the diagnosis is frequently delayed and leads to a 30% incidence of perforation in patients over 70. Elderly patients often present initially with a slightly painful mass (a primary appendiceal abscess) or with adhesive intestinal obstruction 5 or 6 days after a previously undetected perforated appendix.

Appendicitis occurs about once in every 1000 pregnancies and is the most common extrauterine condition requiring abdominal operation. The diagnosis may be missed or delayed because of the frequent occurrence of mild abdominal discomfort and nausea and vomiting during pregnancy. During the last trimester, when the mortality rate from appendicitis is highest, uterine displacement of the appendix to the right upper quadrant and laterally leads to confusion in diagnosis because pain and tenderness are similarly displaced.

DIFFERENTIAL DIAGNOSIS

Appendicitis can be confused with any condition that causes abdominal pain. Diagnostic accuracy is about 75 to 80% for experienced clinicians and must be based solely on the clinical criteria outlined. It is probably better to err slightly in the direction of overdiagnosis, since delay is associated with perforation and increased morbidity and mortality. In unperforated appendicitis, the mortality rate is 0.1%, little more than that associated with general anesthesia; for perforated appendicitis, overall mortality is 3%, (15% in the elderly). In doubtful cases, 4 to 6 h of observation is always more beneficial than harmful. The most common conditions discovered at operation when acute appendicitis is erroneously diagnosed are, in order of frequency, mesenteric lymphadenitis, no organic disease, acute pelvic inflammatory disease, ruptured graafian follicle or corpus luteum cyst, and acute gastroenteritis. In addition, acute cholecystitis, perforated ulcer, acute pancreatitis, acute diverticulitis, strangulating intestinal obstruction, ureteral calculus, and pyelonephritis may present diagnostic difficulties.

Differentiation of *pelvic inflammatory disease* from acute appendicitis on clinical grounds may be virtually impossible. Gram-negative intracellular diplococci on cervical smear are not pathognomonic unless *Neisseria gonorrhoeae* can be cultured. Pain on movement of the cervix is not specific and may occur in appendicitis if perforation has occurred or if the appendix lies adjacent to the uterus or adnexa. *Rupture of a graafian follicle* (mittelschmerz) occurs at midcycle and will spill off blood and fluid to produce pain and tenderness more diffuse and usually of a less severe degree than in appendicitis. Fever and leukocytosis are usually absent. *Rupture of a corpus luteum cyst* is identical clinically to rupture of a graafian follicle but develops about the time of menstruation. The presence of an adnexal mass, evidence of blood loss, and a positive pregnancy test help differentiate *ruptured tubal pregnancy*, but a negative pregnancy test is present when tubal abortion has occurred. *Twisted ovarian cyst* and *endometriosis* are occasionally difficult to distinguish from appendicitis. In all these female conditions, ultrasonography, laparoscopy, and occasionallyCT may be of great value.

Acute mesenteric lymphadenitis is the diagnosis usually given when enlarged, slightly reddened lymph nodes at the root of the mesentery and a normal appendix are encountered at operation in a patient who usually has right lower quadrant tenderness. Whether this is a single, discrete entity is unclear, since the causative factor is not known. Some of these patients have infection with *Y. pseudotuberculosis* or *Y. enterocolytica*, in which case the diagnosis can be established by culture of the mesenteric nodes or by serologic titers (Chap. 162). The diagnosis is essentially impossible clinically, although retrospectively these patients may have a higher temperature and more diffuse pain and tenderness. Children seem to be affected more frequently than adults. *Acute gastroenteritis* usually causes profuse watery diarrhea, often with nausea and vomiting, but without localized findings. Between cramps, the abdomen is completely relaxed. In *Salmonella* gastroenteritis, the abdominal findings are similar, although the pain may be more severe and more localized, and fever and chills are common. The occurrence of similar symptoms among other members of the family may be helpful. When the diagnosis of acute pelvic appendicitis with perforation has been missed, gastroenteritis is the most common previous working diagnosis. Persistent abdominal or rectal tenderness should eliminate the diagnosis of gastroenteritis. *Regional enteritis* (Crohn's disease) is usually associated with a more prolonged history, often with previous exacerbations regarded as episodes of gastroenteritis unless the diagnosis has been established previously. *Meckel's diverticulitis* usually cannot be distinguished from acute appendicitis but is very rare.

TREATMENT

Cathartics and enemas should be avoided if appendicitis is under consideration, and antibiotics should not be administered when the diagnosis is in question, since they will only mask the perforation. The treatment is early operation and appendectomy as soon as the patient can be prepared. Appendectomy is increasingly accomplished laparoscopically and may have some benefits over the open technique. Preparation for operation rarely takes more than 1 to 2 h in early appendicitis but may require 6 to 8 h in cases of severe sepsis and dehydration associated with late perforation. The *only* circumstance in which operation is *not* indicated is the presence of a palpable mass 3 to 5 days after the onset of symptoms. Should operation be undertaken at that time, a phlegmon rather than a definitive abscess will be found, and complications from its dissection are frequent. Such patients treated with broad-spectrum antibiotics, parenteral fluids, and rest usually show resolution of the mass and symptoms within 1 week. *Interval appendectomy* should be done safely 3 months later. Should the mass enlarge or the patient become more toxic, drainage of the abscess is necessary. The complications of subphrenic, pelvic, or other intraabdominal abscesses usually follow perforation with generalized peritonitis and can be avoided by early diagnosis of the disease.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 2 -LIVER AND BILIARY TRACT DISEASE

292. APPROACH TO THE PATIENT WITH LIVER DISEASE - *Marc Ghany, Jay H. Hoofnagle*

In most instances, a diagnosis of liver disease can be made accurately by a careful history, physical examination, and application of a few laboratory tests. In some instances, radiologic examinations are helpful or, indeed, diagnostic. Liver biopsy is considered the "gold standard" in evaluation of liver disease but is now needed less for diagnosis than for grading and staging disease. This chapter provides an introduction to diagnosis and management of liver disease, briefly reviewing the structure and function of the liver; the major clinical manifestations of liver disease; and the use of clinical history, physical examination, laboratory tests, imaging studies, and liver biopsy.

LIVER STRUCTURE AND FUNCTION

The liver is the largest organ of the body, weighing 1 to 1.5 kg and representing 1.5 to 2.5% of the lean body mass. The size and shape of the liver vary and generally match the general body shape -- long and lean or squat and square. The liver is located in the right upper quadrant of the abdomen under the right lower rib cage against the diaphragm and projects for a variable extent into the left upper quadrant. The liver is held in place by ligamentous attachments to the diaphragm, peritoneum, great vessels, and upper gastrointestinal organs. It receives a dual blood supply; approximately 20% of the blood flow is oxygen-rich blood from the hepatic artery, and 80% is nutrient-rich blood from the portal vein arising from the stomach, intestines, and spleen.

The majority of cells in the liver are hepatocytes, which constitute two-thirds of the mass of the liver. The remaining cell types are Kupffer cells (members of the reticuloendothelial system), stellate (Ito or fat-storing) cells, endothelial cells and blood vessels, bile ductular cells, and supporting structures. Viewed by light microscopy, the liver appears to be organized in lobules, with portal areas at the periphery and central veins in the center of each lobule. However, from a functional point of view, the liver is organized into acini, with both hepatic arterial and portal venous blood entering the acinus from the portal areas and then flowing through the sinusoids to the terminal hepatic veins. The advantage of viewing the acinus as the physiologic unit of the liver is that it helps to explain the morphologic patterns of many vascular and biliary diseases not explained by the lobular arrangement.

Portal areas of the liver consist of small veins, arteries, bile ducts, and lymphatics organized in a loose stroma of supporting matrix and small amounts of collagen. Blood flowing into the portal areas is distributed through the sinusoids, passing from zone 1 to zone 3 of the acinus and draining into the terminal hepatic veins ("central veins"). The sinusoids are lined by unique endothelial cells that have prominent fenestrae of variable size, allowing the free flow of plasma but not cellular elements. The plasma is thus in direct contact with hepatocytes in the subendothelial space of Disse.

Hepatocytes have distinct polarity. The basolateral side of the hepatocyte lines the space of Disse and is richly lined with microvilli; it demonstrates endocytotic and pinocytotic activity, with passive and active uptake of nutrients, proteins, and other molecules. The apical pole of the hepatocyte forms the cannicular membranes through which bile components are secreted. The canniculi of hepatocytes form a fine network, which fuses into the bile ductular elements near the portal areas. Kupffer cells usually lie within the sinusoidal vascular space and represent the largest group of fixed macrophages in the body. The stellate cells are located in the space of Disse but are not usually prominent unless activated, when they produce collagen and matrix. Red blood cells stay in the sinusoidal space as blood flows through the lobules, but white blood cells can migrate through or around endothelial cells into the space of Disse and from there to portal areas, where they can return to the circulation through lymphatics.

Hepatocytes perform numerous and vital roles in maintaining homeostasis and health. These functions include the synthesis of most essential serum proteins (albumin, carrier proteins, coagulation factors, many hormonal and growth factors), the production of bile and its carriers (bile acids, cholesterol, lecithin, phospholipids), the regulation of nutrients (glucose, glycogen, lipids, cholesterol, amino acids), and metabolism and conjugation of lipophilic compounds (bilirubin, cations, drugs) for excretion in the bile or urine. Measurement of these activities to assess liver function is complicated by the multiplicity and variability of these functions. The most commonly used liver "function" tests are measurements of serum bilirubin, albumin, and prothrombin time. The serum bilirubin level is a measure of hepatic conjugation and excretion, and the serum albumin level and prothrombin time are measures of protein synthesis. Abnormalities of bilirubin, albumin, and prothrombin time are typical of hepatic dysfunction. Frank liver failure is incompatible with life, and the functions of the liver are too complex and diverse to be subserved by a mechanical pump; dialysis membrane; or concoction of infused hormones, proteins, and growth factors.

LIVER DISEASES

While there are many causes of liver disease (Table 292-1), they generally present clinically in a few distinct patterns, usually classified as either hepatocellular or cholestatic (obstructive). In *hepatocellular diseases* (such as viral hepatitis or alcoholic liver disease), features of liver injury, inflammation, and necrosis predominate. In *cholestatic diseases* (such as gall stone or malignant obstruction, primary biliary cirrhosis, many drug-induced liver diseases), features of inhibition of bile flow predominate. The pattern of onset and prominence of symptoms can rapidly suggest a diagnosis, particularly if major risk factors are considered, such as the age and sex of the patient and a history of exposure or risk behaviors.

Typical presenting symptoms of liver disease include jaundice, fatigue, itching, right upper quadrant pain, abdominal distention, and intestinal bleeding. At present, however, many patients are diagnosed with liver disease who have no symptoms and who have been found to have abnormalities in biochemical liver tests as a part of a routine physical examination or screening for blood donation or for insurance or employment. The wide availability of batteries of liver tests makes it relatively simple to demonstrate the presence of liver injury as well as to rule it out in someone suspected of liver disease.

Evaluation of patients with liver disease should be directed at (1) establishing the etiologic diagnosis, (2) estimating the disease severity (grading), and (3) establishing the disease stage (staging). *Diagnosis* should focus on the category of disease, such as hepatocellular versus cholestatic injury, as well as on the specific etiologic diagnosis. *Grading* refers to assessing the severity or activity of disease -- active or inactive, and mild, moderate, or severe. *Staging* refers to estimating the place in the course of the natural history of the disease, whether acute or chronic; early or late; precirrhotic, cirrhotic, or end-stage.

The goal of this chapter is to introduce general, salient concepts in the evaluation of patients with liver disease that help lead to the diagnoses discussed in subsequent chapters.

CLINICAL HISTORY

The clinical history should focus on the symptoms of liver disease -- their nature, pattern of onset, and progression -- and on potential risk factors for liver disease. The symptoms of liver disease include constitutional symptoms such as fatigue, weakness, nausea, poor appetite, and malaise and the more liver-specific symptoms of jaundice, dark urine, light stools, itching, abdominal pain, and bloating. Symptoms can also suggest the presence of cirrhosis, end-stage liver disease, or complications of cirrhosis such as portal hypertension. Generally, the constellation of symptoms and their pattern of onset rather than a specific symptom points to an etiology.

Fatigue is the most common and most characteristic symptom of liver disease. It is variously described as lethargy, weakness, listlessness, malaise, increased need for sleep, lack of stamina, and poor energy. The fatigue of liver disease typically arises after activity or exercise and is rarely present or severe in the morning after adequate rest (afternoon versus morning fatigue). Fatigue in liver disease is often intermittent and variable in severity from hour to hour and day to day. In some patients, it may not be clear whether fatigue is due to the liver disease or to other problems such as stress, anxiety, sleep disturbance, or a concurrent illness.

Nausea occurs with more severe liver disease and may accompany fatigue or be provoked by odors of food or eating fatty foods. Vomiting can occur but is rarely persistent or prominent. Poor appetite with weight loss occurs commonly in acute liver diseases but is rare in chronic disease, except when cirrhosis is present and advanced. Diarrhea is uncommon in liver disease, except with severe jaundice, in which case lack of bile acids reaching the intestine can lead to steatorrhea.

Right upper quadrant discomfort or ache ("liver pain") occurs in many liver diseases and is usually marked by tenderness over the liver area. The pain arises from stretching or irritation of Glisson's capsule, which surrounds the liver and is rich in nerve endings. Severe pain is most typical of gall bladder disease, liver abscess, and severe venoocclusive disease but is an occasional accompaniment of acute hepatitis.

Itching occurs with acute liver disease, appearing early in obstructive jaundice (from biliary obstruction or drug-induced cholestasis) and somewhat later in hepatocellular disease (acute hepatitis). Itching also occurs in chronic liver diseases, typically the cholestatic forms such as primary biliary cirrhosis and sclerosing cholangitis where it is often the presenting symptom, occurring before the onset of jaundice. However, itching can occur in any liver disease, particularly once cirrhosis is present.

Jaundice is the hallmark symptom of liver disease and perhaps the most reliable marker of severity. Patients usually report darkening of the urine before they notice scleral icterus. Jaundice is rarely detectable with a bilirubin level less than 43 umol/L (2.5 mg/dL). With severe cholestasis there will also be lightening of the color of the stools and steatorrhea. Jaundice without dark urine usually indicates indirect (unconjugated) hyperbilirubinemia and is typical of hemolytic anemia and the genetic disorders of bilirubin conjugation, the common and benign form being Gilbert's syndrome and the rare and severe form being Crigler-Najjar syndrome. Gilbert's syndrome affects up to 5% of the population; the jaundice is more noticeable after fasting and with stress.

Major risk factors for liver disease that should be sought in the clinical history include details of alcohol use, medications (including herbal compounds, birth control pills, and over-the-counter medications), personal habits, sexual activity, travel, exposure to jaundiced or other high-risk persons, injection drug use, recent surgery, remote or recent transfusion with blood and blood products, occupation, accidental exposure to blood or needlestick, and familial history of liver disease.

For assessing the risk of viral hepatitis, a careful history of sexual activity is of particular importance and should include life-time number of sexual partners and, for men, a history of having sex with men. Sexual exposure is a common mode of spread of hepatitis B but is rare for hepatitis C. Maternal-infant transmission occurs with both hepatitis B and C. Vertical spread of hepatitis B can now be prevented by passive and active immunization of the infant at birth. Vertical spread of hepatitis C is uncommon, but there are no known means of prevention. A history of injection drug use, even in the remote past, is of great importance in assessing the risk for hepatitis B and C. Injection drug use is now the single most common risk factor for hepatitis C. Transfusion with blood or blood products is no longer an important risk factor for acute viral hepatitis. However, blood transfusions received before the introduction of sensitive enzyme immunoassays for antibody to hepatitis C virus (anti-HCV) in 1992 is an important risk factor for chronic hepatitis C. Blood transfusion before 1986, when screening for antibody to hepatitis B core antigen (anti-HBc) was introduced, is also a risk factor for hepatitis B. Travel to an underdeveloped area of the world, exposure to persons with jaundice, and exposure to young children in day-care centers are risk factors for hepatitis A. Tattooing and body piercing (for hepatitis B and C) and eating shellfish (for hepatitis A) are frequently mentioned but actually quite rate types of exposure for acquiring hepatitis.

A history of alcohol intake is important in assessing the cause of liver disease and also in planning management and recommendations. In the United States, for example, at least 70% of adults drink alcohol to some degree, but significant alcohol intake is less common; in population-based surveys, only 5% have more than two drinks per day, the average drink representing 11 to 15 g alcohol. Alcohol consumption associated with an increased rate of alcoholic liver disease is probably more than two drinks (22 to 30 g) per day in women and three drinks (33 to 45 g) in men. Most patients with alcoholic cirrhosis have a much higher daily intake and have drunk excessively for 10 years or more before onset of liver disease. In assessing alcohol intake, the history should also focus upon whether alcohol abuse or dependence is present. Alcoholism is usually defined on the behavioral patterns and consequences of alcohol intake, not on the basis of the amount of alcohol intake. *Abuse* is defined by a repetitive pattern of drinking alcohol that has adverse effects on social, family, occupational, or health status. *Dependence* is defined by alcohol-seeking behavior, despite its adverse effects. Many alcoholics demonstrate both dependence and abuse, and dependence is considered the more serious and advanced form of alcoholism. A clinically helpful approach to diagnosis of alcohol dependence and abuse is the use of the CAGE questionnaire (Table 292-2), which is recommended in all medical history taking.

Family history can be helpful in assessing liver disease. Familial causes of liver disease include Wilson's disease; hemochromatosis anda_1-antitrypsin (a_1AT) deficiency; and the more uncommon inherited pediatric liver diseases of familial intrahepatic cholestasis (FIC), benign recurrent intrahepatic cholestasis (BRIC), and Alagille's syndrome. Onset of severe liver disease in childhood or adolescence with a family history of liver disease or neuropsychiatric disturbance should lead to investigation for Wilson's disease. A family history of cirrhosis, diabetes, or endocrine failure and the appearance of liver disease in adulthood should suggest hemochromatosis and lead to investigation of iron status. Patients with abnormal iron studies warrant genotyping of the HFE gene for the C282Y and H63D mutations typical of genetic hemochromatosis. A family history of emphysema should provoke investigation ofa_1AT levels and, if low, for Pi genotype.

PHYSICAL EXAMINATION

The physical examination rarely demonstrates evidence of liver dysfunction in a patient without symptoms or laboratory findings, nor are most signs of liver disease specific to one diagnosis. Thus, the physical examination usually complements rather than replaces the need for other diagnostic approaches. In many patients, the physical examination is normal unless the disease is acute or severe and advanced. Nevertheless, the physical examination is important in that it can be the first evidence for the presence of hepatic failure, portal hypertension, and liver decompensation. In addition, the physical examination can reveal signs that point to a specific diagnosis, either in risk factors or in associated diseases or findings.

Typical physical findings in liver disease are icterus, hepatomegaly, hepatic tenderness, splenomegaly, spider angiomata, palmar erythema, and excoriations. Signs of advanced disease include muscle-wasting, ascites, edema, dilated abdominal veins, hepatic fetor, asterixis, mental confusion, stupor, and coma.

Icterus is best appreciated by inspecting the sclera under natural light. In fair-skinned individuals, a yellow color of the skin may be obvious. In dark-skinned individuals, the mucous membranes below the tongue can demonstrate jaundice. Jaundice is rarely detectable if the serum bilirubin level is <43 umol/L (2.5 ug/dL) but may remain detectable below this level during recovery from jaundice (because of protein and tissue binding of conjugated bilirubin).

Spider angiomata and palmar erythema occur in both acute and chronic liver disease and may be especially prominent in persons with cirrhosis, but they can occur in normal individuals and are frequently present during pregnancy. Spider angiomata are superficial, tortuous arterioles and, unlike simple telangiectases, typically fill from the center outwards. Spider angiomata occur only on the arms, face, and upper torso; they can be pulsatile and may be difficult to detect in dark-skinned individuals.

Hepatomegaly is not a very reliable sign of liver disease, because of the variability of the size and shape of the liver and the physical impediments to assessing liver size by percussion and palpation. Marked hepatomegaly is typical of cirrhosis, venoocclusive disease, metastatic or primary cancers of the liver, and alcoholic hepatitis. Careful assessment of the liver edge may also demonstrate unusual firmness, irregularity of the surface, or frank nodules. Perhaps the most reliable physical finding in examining the liver is hepatic tenderness. Discomfort on touching or pressing on the liver should be carefully sought with percussive comparison of the right and left upper quadrants.

Splenomegaly occurs in many medical conditions but can be a subtle but significant physical finding in liver disease. The availability of ultrasound (US) assessment of the spleen allows for confirmation of the physical finding.

Signs of advanced liver disease include muscle-wasting and weight loss as well as hepatomegaly, bruising, ascites, and edema. Ascites is best appreciated by attempts to detect shifting dullness by careful percussion.USexamination will confirm the finding of ascites in equivocal cases. Peripheral edema can occur with or without ascites. In patients with advanced liver disease, other factors frequently contribute to edema formation, including hypoalbuminemia, venous insufficiency, heart failure, and medications.

Hepatic failure is defined as the occurrence of signs or symptoms of hepatic encephalopathy in a person with severe acute or chronic liver disease. The first signs of hepatic encephalopathy can be subtle and nonspecific -- change in sleep patterns, change in personality, irritability, and mental dullness. Thereafter, confusion, disorientation, stupor, and eventually coma supervene. Physical findings include asterixis and flapping tremors of the body and tongue. *Fetor hepaticus* refers to the slightly sweet, ammoniacal odor that is common in patients with liver failure, particularly if there is portal-venous shunting of blood around the liver. Other causes of coma and confusion should be excluded, mainly electrolyte imbalances, sedative use, and renal or respiratory failure. A helpful measure of hepatic encephalopathy is a careful mental status examination and use of the trail-making test, which consists of a series of 20 numbered circles that the patient is asked to connect as rapidly as possible using a pencil. The normal range for the connect-the-dot test is 15 to 30 s; it is considerably delayed in patients with early hepatic encephalopathy. Other tests include drawing abstract objects or comparison of a signature to previous examples.

Other signs of advanced liver disease include umbilical hernia from ascites, prominent veins over the abdomen, and *caput medusa*, which consists of collateral veins seen radiating from the umbilicus and resulting from the recanulation of the umbilical vein. Widened pulse pressure and signs of a hyperdynamic circulation can occur in patients with cirrhosis as a result of fluid and sodium retention, increased cardiac output, and reduced peripheral resistance. Patients with long-standing cirrhosis are prone to develop the hepatopulmonary syndrome with hypoxemia due to pulmonary arteriovenous shunting, characterized by hypoxia that worsens when lying flat.

Several skin disorders and changes occur commonly in liver disease. Hyperpigmentation is typical of advanced chronic cholestatic diseases such as primary biliary cirrhosis and sclerosing cholangitis. In these same conditions, xanthelasma and tendon xanthomata occur as a result of retention and high serum levels of lipids and cholesterol. A slate-gray pigmentation to the skin also occurs with hemochromatosis if iron levels are high for a prolonged period. Mucocutaneous vasculitis with palpable purpura, especially on the lower extremities, is typical of cryoglobulinemia of chronic hepatitis C but can also occur in chronic hepatitis B.

Some physical signs point to specific liver diseases. Kayser-Fleischer rings occur in Wilson's disease and consist of a golden-brown copper pigment deposited at the periphery of the cornea; they are best seen by slit-lamp examination. In metastatic liver disease or primary hepatocellular carcinoma, signs of cachexia and wasting may be prominent, as well as firm hepatomegaly and a hepatic bruit.

LABORATORY TESTING

Diagnosis in liver disease is greatly aided by the availability of reliable and sensitive tests of liver injury and function. Use and interpretation of liver function tests is summarized in Chap. 293. A typical battery of blood tests used for initial assessment of liver disease includes measuring levels of serum alanine and aspartate aminotransferases (ALT and AST), alkaline phosphatase, direct and total serum bilirubin, and albumin and assessing prothrombin time. The pattern of abnormalities generally points to hepatocellular versus cholestatic liver disease and will help to decide whether the disease is acute or chronic and whether cirrhosis and hepatic failure are present. Based on these results, further testing over time may be necessary. Other laboratory tests may be helpful, such as g-glutamyl transpeptidase (GGT) to define whether alkaline phosphatase elevations are due to liver disease; hepatitis serology to define the type of viral hepatitis; and autoimmune markers to diagnose primary biliary cirrhosis (antimitochondrial antibody; AMA), sclerosing cholangitis (peripheral antineutrophil cytoplasmic antibody; pANCA), autoimmune hepatitis (antinuclear, smooth-muscle, and liver-kidney microsomal antibody). A simple delineation of laboratory abnormalities and common liver diseases is given in Table 292-3.

DIAGNOSTIC IMAGING

There have been great advances made in hepatic imaging, although no method is suitably accurate in demonstrating underlying cirrhosis. There are many modalities available for imaging the liver.US, computed tomography (CT), and magnetic resonance imaging (MRI) are the most commonly employed and are complementary to each other. In general, US and CT have a high sensitivity for detecting biliary duct dilatation and are the first-line options for investigating the patient with suspected obstructive jaundice. Both US and CT can detect a fatty liver, which appears bright on both studies. Endoscopic retrograde cholangiopancreatography (ERCP) is the procedure of choice for visualization of the biliary tree. ERCP also provides several therapeutic options in patients with obstructive jaundice, such as sphincterotomy, stone extraction, and placement of nasobiliary catheters and biliary stents. Doppler US and MRI are used to assess hepatic vasculature and hemodynamics and to monitor surgically or radiologically placed vascular shunts such as transjugular intrahepatic portosystemic shunts (TIPS). CT and MRI are indicated for the identification and evaluation of hepatic masses, staging of liver tumors, and preoperative assessment. With regard to mass lesions, sensitivity of hepatic imaging continues to increase; unfortunately, specificity remains a problem, and often two and sometimes three studies are needed before a diagnosis can be reached. Finally, interventional radiologic techniques allow the biopsy of solitary lesions, insertion of drains into hepatic abscesses, and creation of vascular shunts in patients with portal hypertension. Which modality to use depends on factors such as availability, cost, and experience of the radiologist with each technique.

LIVER BIOPSY

Liver biopsy remains the gold standard in the evaluation of patients with liver disease, particularly in patients with chronic liver diseases. In selected instances, liver biopsy is necessary for diagnosis but is more often useful in assessing the severity (grade) and stage of liver damage, in predicting prognosis, and in monitoring response to treatment.

Diagnosis of Liver Disease The major causes of liver disease and key diagnostic features are outlined in Table 292-3 (specifics of diagnosis are discussed in later chapters). The most common causes of acute liver disease are viral hepatitis (particularly hepatitis A, B, and C), drug-induced liver injury, cholangitis, and alcoholic liver disease. Liver biopsy is usually not needed in the diagnosis and management of acute liver disease, exceptions being situations where the diagnosis remains unclear despite thorough clinical and laboratory investigation. Liver biopsy can be helpful in the diagnosis of drug-induced liver disease and in establishing the diagnosis of acute alcoholic hepatitis.

The most common causes of chronic liver disease in general order of frequency are chronic hepatitis C, alcoholic liver disease, nonalcoholic steatohepatitis, chronic hepatitis B, autoimmune hepatitis, sclerosing cholangitis, primary biliary cirrhosis, hemochromatosis, and Wilson's disease. Strict diagnostic criteria have not been developed for most liver diseases, but liver biopsy plays an important role in the diagnosis of autoimmune hepatitis, primary biliary cirrhosis, nonalcoholic and alcoholic steatohepatitis, and Wilson's disease (with a quantitative hepatic copper level).

Grading and Staging of Liver Disease Grading refers to an assessment of the severity or activity of liver disease, whether acute or chronic; active or inactive; and mild, moderate, or severe. Liver biopsy is the most accurate means of assessing severity, particularly in chronic liver disease. Serum aminotransferase levels are used as a convenient and noninvasive means to follow disease activity, but aminotransferases are not always reliable in reflecting disease severity. Thus, normal serum aminotransferases in patients with hepatitis B surface antigen (HBsAg) in serum may indicate the inactive HBsAg carrier state or may reflect mild chronic hepatitis B or hepatitis B with fluctuating disease activity. Serum testing for hepatitis B e antigen and hepatitis B virus DNA can help resolve these different patterns, but these markers can also fluctuate and change over time. Similarly, in chronic hepatitis C, serum aminotransferases can be normal despite moderate activity of disease. Finally, in both alcoholic and nonalcoholic steatohepatitis, aminotransferases are quite unreliable in reflecting severity. In these conditions, liver biopsy is helpful in guiding management and recommending therapy, particularly if therapy is difficult, prolonged, and expensive as is often the case in chronic viral hepatitis. There are several well-verified numerical scales for grading activity in chronic liver disease, the most common being the histology activity index and the Ishak histology scale.

Liver biopsy is also the most accurate means of assessing stage of disease as early or advanced, precirrhotic, and cirrhotic. Staging of disease pertains largely to chronic liver diseases in which progression to cirrhosis and end-stage liver disease can occur, but which may require years or decades to develop. Clinical features, biochemical tests, and hepatic imaging studies are helpful in assessing stage but generally become abnormal only in the middle to late stages of cirrhosis. Early stages of cirrhosis are generally detectable only by liver biopsy. In assessing stage, the degree of fibrosis is usually used as its quantitative measure. The amount of fibrosis is generally staged on a 0 to 4+ (histology activity index) or 0 to 6+ scale (Ishak scale).

Cirrhosis can also be staged clinically. A reliable staging system is the modified Child-Pugh classification with a scoring system of 5 to 15: scores of 5 and 6 being Child-Pugh class A (consistent with "compensated cirrhosis"), scores of 7 to 9 indicating class B, and 10 to 15 class C (Table 292-4). This scoring system was initially devised to stratify patients into risk groups prior to undergoing portal decompressive surgery. It is now used to assess prognosis in cirrhosis and provides the standard criteria for listing for liver transplantation (Child-Pugh class B). The Child-Pugh score is a reasonably reliable predictor of survival in many liver diseases and predicts the likelihood of major complications of cirrhosis such as bleeding from varices and spontaneous bacterial peritonitis. Other means of assessing stage and survival have been developed for primary biliary cirrhosis and sclerosing cholangitis (Mayo Risk scores), which are somewhat more accurate but which actually rely mostly on the same measurements as the Child-Pugh score.

Thus, liver biopsy is helpful not only in diagnosis but also in management of chronic liver disease and assessment of prognosis. Because liver biopsy is an invasive procedure and not without complications, it should be used only when it will contribute materially to management and therapeutic decisions.

(Bibliography omitted in Palm version)

Back to Table of Contents

293. EVALUATION OF LIVER FUNCTION - *Daniel S. Pratt*, *Marshall M. Kaplan*

Several biochemical tests are useful in the evaluation and management of patients with hepatic dysfunction. These tests can be used to (1) detect the presence of liver disease; (2) distinguish among different types of liver disorders; (3) gauge the extent of known liver damage; and (4) follow the response to treatment.

Liver tests have shortcomings. They can be normal in patients with serious liver disease and abnormal in patients with diseases that do not affect the liver. Liver tests rarely suggest a specific diagnosis; rather, they suggest a general category of liver disease, such as hepatocellular or cholestatic, which then further directs the evaluation.

The liver carries out thousands of biochemical functions, most of which cannot be easily measured by blood tests. Laboratory tests measure only a limited number of these functions. In fact, many tests, such as the aminotransferases or alkaline phosphatase, do not measure liver function at all. Rather, they detect liver cell damage or interference with bile flow. Thus, no one test enables the clinician to accurately assess the liver's total functional capacity.

To increase both the sensitivity and the specificity of laboratory tests in the detection of liver disease, it is best to use them as a battery. Those tests usually employed in clinical practice include the bilirubin, aminotransferases, alkaline phosphatase, albumin, and prothrombin time tests. When more than one of these tests provide abnormal findings, or the findings are persistently abnormal on serial determinations, the probability of liver disease is high. When all test results are normal, the probability of missing occult liver disease is low.

When evaluating patients with liver disorders, it is helpful to group these tests into general categories. The classification we have found most useful is given below.

TESTS BASED ON DETOXIFICATION AND EXCRETORY FUNCTIONS

Serum Bilirubin Bilirubin, a breakdown product of the porphyrin ring of heme-containing proteins, is found in the blood in two fractions -- conjugated and unconjugated. The *van den Bergh assay*, or a variation of it, is still used in most clinical chemistry laboratories to determine the total serum bilirubin level and what amount is conjugated or unconjugated bilirubin. In this assay, 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. The unconjugated fraction, also termed the indirect fraction, is insoluble in water and is bound to albumin in the blood. The conjugated (direct) bilirubin fraction is water soluble and can therefore be excreted by the kidney. When measured by the original van den Bergh method, the normal total serum bilirubin concentration is less than 1 mg/dL. Up to 30%, or 0.3 mg/dL, of the total is direct-reacting (or conjugated) bilirubin.

Elevation of the unconjugated fraction of bilirubin is rarely due to liver disease. An isolated elevation of unconjugated bilirubin is seen primarily in hemolytic disorders and in a number of genetic conditions such as Crigler-Najjar and Gilbert's syndromes. *Gilbert's syndrome* is a common, benign condition with a reported incidence in 3 to 7% of the population. It is marked by the impaired conjugation of bilirubin due to reduced bilirubin uridine diphosphate (UDP) glucuronosyltransferase activity. This results in mild unconjugated hyperbilirubinemia, which is marked by considerable fluctuations and is sometimes only identified during periods of fasting. One molecular defect that has been identified in patients with Gilbert's syndrome is in the TATAA element in the 5¢ promoter region of the bilirubin UDP-glucuronosyltransferase gene upstream of exon 1. This defect alone is not necessarily sufficient for producing the clinical syndrome of Gilbert's, as there are patients who are homozygous for this defect yet do not have the levels of hyperbilirubinemia typically seen in Gilbert's syndrome. Isolated unconjugated hyperbilirubinemia (bilirubin elevated, but less than 15% direct) should prompt a workup for hemolysis (Fig. 293-1). In the absence of hemolysis, an isolated unconjugated hyperbilirubinemia can be attributed to Gilbert's syndrome and no further evaluation is required.

In contrast, conjugated hyperbilirubinemia almost always implies liver or biliary tract disease. The rate-limiting step in bilirubin metabolism is not conjugation of bilirubin, but rather the transport of conjugated bilirubin into the bile canaliculi. Thus, elevation of the conjugated fraction may be seen in any type of liver disease. In most liver diseases, both conjugated and unconjugated fractions of the bilirubin tend to be elevated. Except in the presence of a purely unconjugated hyperbilirubinemia, fractionation of the bilirubin is rarely helpful in determining the cause of jaundice.

Concern may be generated by a slower than expected decline of the serum bilirubin during convalescence from certain liver diseases. This can be attributed to the covalent binding of conjugated bilirubin to serum albumin that occurs when there is a prolonged episode of conjugated hyperbilirubinemia. The serum half-life of the albumin-bilirubin complex (15 days) is much longer than that of conjugated bilirubin and closer to that of albumin.

Urine Bilirubin Unconjugated bilirubin always binds to albumin in the serum and is not filtered by the kidney. Therefore, any bilirubin found in the urine is conjugated bilirubin; the presence of bilirubinuria implies the presence of liver disease. A urine dipstick test can theoretically give the same information as fractionation of the serum bilirubin. This test is almost 100% accurate. Phenothiazines may give a false positive reading with the Ictotest tablet.

Blood Ammonia Ammonia is produced in the body during normal protein metabolism and by intestinal bacteria, primarily those in the colon. The liver plays a role in the detoxification of ammonia by converting it to urea, which is excreted by the kidneys. Striated muscle also plays a role in detoxification of ammonia, which is combined with glutamic acid to form glutamine. Patients with advanced liver disease typically have significant muscle wasting, which likely contributes to hyperammonemia in these patients. Some physicians use the blood ammonia for detecting encephalopathy or for monitoring hepatic synthetic function, although its use for either of these indications has problems. There is very poor correlation between either the presence or the degree of acute encephalopathy and elevation of blood ammonia; it can be occasionally useful for identifying the occult liver disease in patients with mental status changes. There is also a poor correlation of the blood serum ammonia and hepatic function. The ammonia can be elevated in patients with severe portal hypertension and portal blood shunting around the liver even in the presence of normal or near normal hepatic function.

Serum Enzymes The liver contains thousands of enzymes, some of which are also present in the serum in very low concentrations. These enzymes have no known function in the serum and behave like other serum proteins. They are distributed in the plasma and in interstitial fluid and have characteristic half-lives, usually measured in days. Very little is known about the catabolism of serum enzymes, although they are probably cleared by cells in the reticuloendothelial system. The elevation of a given enzyme activity in the serum is thought to primarily reflect its increased rate of entrance into serum from damaged liver cells.

Serum enzyme tests can be grouped into three categories: (1) enzymes whose elevation in serum reflects damage to hepatocytes; (2) enzymes whose elevation in serum reflects cholestasis; and (3) enzyme tests that do not fit precisely into either pattern.

Enzymes that Reflect Damage to Hepatocytes The aminotransferases (transaminases) are sensitive indicators of liver cell injury and are most helpful in recognizing acute hepatocellular diseases such as hepatitis. They include the aspartate aminotransferase (AST) and the alanine aminotransferase (ALT). AST is found in the liver, cardiac muscle, skeletal muscle, kidneys, brain, pancreas, lungs, leukocytes, and erythrocytes in decreasing order of concentration. ALT is found primarily in the liver. The aminotransferases are normally present in the serum in low concentrations. These enzymes are released into the blood in greater amounts when there is damage to the liver cell membrane resulting in increased permeability. Liver cell necrosis is not required for the release of the aminotransferases and there is a poor correlation between the degree of liver cell damage and the level of the aminotransferases. Thus, the absolute elevation of the aminotransferases is of no prognostic significance in acute hepatocellular disorders.

Any type of liver cell injury can cause modest elevations in the serum aminotransferases. Levels of up to 300 U/L are nonspecific and may be found in any type of liver disorder. Striking elevations -- i.e., aminotransferases>1000 U/L -- occur almost exclusively in disorders associated with extensive hepatocellular injury such as (1) viral hepatitis, (2) ischemic liver injury (prolonged hypotension or acute heart failure), or (3) toxin or drug-induced liver injury.

The pattern of the aminotransferase elevation can be helpful diagnostically. In most acute hepatocellular disorders, the ALTis higher than or equal to theAST. An AST:ALT ratio >2:1 is suggestive while a ratio>3:1 is highly suggestive of alcoholic liver disease. The AST in alcoholic liver disease is rarely>300 U/L and the ALT is often normal. A low level of ALT in the serum is due to an alcohol-induced deficiency of pyridoxal phosphate.

The aminotransferases are usually not greatly elevated in obstructive jaundice. One notable exception occurs during the acute phase of biliary obstruction caused by the passage of a gallstone into the common bile duct. In this setting, the aminotransferases can briefly be in the 1,000 to 2,000 U/L range. However, aminotransferase levels decrease quickly and the liver function tests rapidly evolve into one typical of cholestasis.

Enzymes that Reflect Cholestasis The activities of three enzymes -- alkaline phosphatase, 5¢-nucleotidase, and gamma glutamyl transpeptidase (GGT) -- are usually elevated in cholestasis. Alkaline phosphatase and 5¢-nucleotidase are found in or near the bile canalicular membrane of hepatocytes, while GGT is located in the endoplasmic reticulum and in bile duct epithelial cells. Reflecting its more diffuse localization in the liver, GGT elevation in serum is less specific for cholestasis than are elevations of alkaline phosphatase or 5¢-nucleotidase. Some have advocated the use of GGT to identify patients with occult alcohol use. Its lack of specificity makes its use in this setting questionable.

The normal serum alkaline phosphatase consists of many distinct isoenzymes found in the liver, bone, placenta, and, less commonly, small intestine. Patients over age 60 can have a mildly elevated alkaline phosphatase (1 to 1$\frac{1}{2}$times normal), while individuals with blood types O and B can have an elevation of the serum alkaline phosphatase after eating a fatty meal due to the influx of intestinal alkaline phosphatase into the blood. It is also nonpathologically elevated in children and adolescents undergoing rapid bone growth because of bone alkaline phosphatase, and late in normal pregnancies due to the influx of placental alkaline phosphatase.

Elevation of liver-derived alkaline phosphatase is not totally specific for cholestasis and a less than threefold elevation can be seen in almost any type of liver disease. Alkaline phosphatase elevations greater than four times normal occur primarily in patients with cholestatic liver disorders, infiltrative liver diseases such as cancer, and bone conditions characterized by rapid bone turnover (e.g., Paget's disease). In bone diseases, the elevation is due to increased amounts of the bone isoenzymes. In liver diseases, the elevation is almost always due to increased amounts of the liver isoenzyme.

If an elevated serum alkaline phosphatase is the only abnormal finding in an apparently healthy person, or if the degree of elevation is higher than expected in the clinical setting, identification of the source of elevated isoenzymes is helpful (Fig. 293-1). This problem can be approached in several ways. First, and most precise, is the fractionation of the alkaline phosphatase by electrophoresis. The second approach is based on the observation that alkaline phosphatases from individual tissues differ in susceptibility to inactivation by heat. The finding of an elevated serum alkaline phosphatase level in a patient with a heat-stable fraction strongly suggests that the placenta or a tumor is the source of the elevated enzyme in serum. Susceptibility to inactivation by heat increases, respectively, for the intestinal, liver, and bone alkaline phosphatases, bone being by far the most sensitive. The third, best substantiated, and most available approach involves the measurement of serum 5¢-nucleotidase orGGT. These enzymes are rarely elevated in conditions other than liver disease.

In the absence of jaundice or elevated aminotransferases, an elevated alkaline phosphatase of liver origin often, but not always, suggests early cholestasis and, less often, hepatic infiltration by tumor or granulomata. Other conditions that cause isolated elevations of the alkaline phosphatase include Hodgkin's disease, diabetes, hyperthyroidism, congestive heart failure, and inflammatory bowel disease.

The level of serum alkaline phosphatase elevation is not helpful in distinguishing between intrahepatic and extrahepatic cholestasis. There is essentially no difference among the values found in obstructive jaundice due to cancer, common duct stone, sclerosing cholangitis, or bile duct stricture. Values are similarly increased in patients with intrahepatic cholestasis due to drug-induced hepatitis, primary biliary cirrhosis, rejection of transplanted livers, and, rarely, alcohol-induced steatonecrosis. Values are also greatly elevated in hepatobiliary disorders seen in patients with AIDS (e.g., AIDS cholangiopathy due to cytomegalovirus or cryptosporidial infection and tuberculosis with hepatic involvement).

TESTS THAT MEASURE BIOSYNTHETIC FUNCTION OF THE LIVER

Serum Albumin Serum albumin is synthesized exclusively by hepatocytes. Serum albumin has a long half-life: 15 to 20 days, with approximately 4% degraded per day. Because of this slow turnover, the serum albumin is not a good indicator of acute or mild hepatic dysfunction; only minimal changes in the serum albumin are seen in acute liver conditions such as viral hepatitis, drug-related hepatoxicity, and obstructive jaundice. In hepatitis, albumin levels below 3 g/dL should raise the possibility of chronic liver disease. Hypoalbuminemia is more common in chronic liver disorders such as cirrhosis and usually reflects severe liver damage and decreased albumin synthesis. One exception is the patient with ascites in whom synthesis may be normal or even increased, but levels are low because of the increased volume of distribution. However, hypoalbuminemia is not specific for liver disease and may occur in protein malnutrition of any cause, as well as protein-losing enteropathies, nephrotic syndrome, and chronic infections that are associated with prolonged increases in serum interleukin-1 and/or tumor necrosis factor levels that inhibit albumin synthesis. Serum albumin should not be measured for screening in patients in whom there is no suspicion of liver disease. A general medical clinic study of consecutive patients in whom no indications were present for albumin measurement showed that while 12% of patients had abnormal test results, the finding was of clinical importance in only 0.4%.

Serum Globulins Serum globulins are a group of proteins made up of gamma globulins (immunoglobulins) produced by B lymphocytes and alpha and beta globulins produced primarily in hepatocytes. Gamma globulins are increased in chronic liver disease, such as chronic hepatitis and cirrhosis. In cirrhosis, the increased serum gamma globulin concentration is due to the increased synthesis of antibodies, some of which are directed against intestinal bacteria. This occurs because the cirrhotic liver fails to clear bacterial antigens that normally reach the liver through the hepatic circulation.

Increases in the concentration of specific isotypes of gamma globulins are often helpful in the recognition of certain chronic liver diseases. Diffuse polyclonal increases in IgG levels are common in autoimmune hepatitis; increases greater than 100% should alert the clinician to this possibility. Increases in the IgM levels are common in primary biliary cirrhosis, while increases in the IgA levels occur in alcoholic liver disease.

Coagulation Factors With the exception of factor VIII, the blood clotting factors are made exclusively in hepatocytes. Their serum half-lives are much shorter than albumin, ranging from 6 hours for factor VII to 5 days for fibrinogen. Because of their rapid turnover, measurement of the clotting factors is the single best acute measure of hepatic synthetic function and helpful in both the diagnosis and assessing the prognosis of acute parenchymal liver disease. Useful for this purpose is the *serum prothrombin time*, which collectively measures factors II, V, VII, and X. Biosynthesis of factors II, VII, IX, and X depends on vitamin K. The prothrombin time may be elevated in hepatitis and cirrhosis as well as in disorders that lead to vitamin K deficiency such as obstructive jaundice or fat malabsorption of any kind. Marked prolongation of the prothrombin time,>5 s above control and not corrected by parenteral vitamin K administration, is a poor prognostic sign in acute viral hepatitis and other acute and chronic liver diseases.

OTHER DIAGNOSTIC TESTS

While tests may direct the physician to a category of liver disease, additional radiologic testing and procedures are often necessary to make the proper diagnosis, as shown inFig. 293-1. The two most commonly-used ancillary tests are reviewed here.

Percutaneous Liver Biopsy Percutaneous biopsy of the liver is a safe procedure that can be easily performed at the bedside with local anesthesia. Liver biopsy is of proven value in the following situations: (1) hepatocellular disease of uncertain cause; (2) prolonged hepatitis with the possibility of chronic active hepatitis; (3) unexplained hepatomegaly; (4) unexplained splenomegaly; (5) hepatic filling defects by radiologic imaging; (6) fever of unknown origin; (7) staging of malignant lymphoma. Liver biopsy is most accurate in disorders causing diffuse changes throughout the liver and is subject to sampling error in focal infiltrative disorders such as hepatic metastases. Liver biopsy should not be the initial procedure in the diagnosis of cholestasis. The biliary tree should first be assessed for signs of obstruction.

Ultrasonography Ultrasonography is the first diagnostic test to use in patients whose liver tests suggest cholestasis, to look for the presence of a dilated intrahepatic or extrahepatic biliary tree, or to identify gallstones. In addition, it shows space-occupying lesions within the liver, enables the clinician to distinguish between cystic and solid masses, and helps direct percutaneous biopsies. Ultrasound with Doppler imaging can detect the patency of the portal vein, hepatic artery, and hepatic veins and determine the direction of blood flow. This is the first test ordered in patients suspected of having Budd-Chiari syndrome.

USE OF LIVER TESTS

As previously noted, the best way to increase the sensitivity and specificity of laboratory tests in the detection of liver disease is to employ a battery of tests that include the aminotransferases, alkaline phosphatase, bilirubin, albumin, and prothrombin time along with the judicious use of the other tests described in this chapter.Table 293-1 shows how patterns of liver tests can lead the clinician to a category of disease which will direct further evaluation. However, it is important to remember that no single set of liver tests will necessarily provide a diagnosis. It is often necessary to repeat these tests on several occasions over days to weeks for a diagnostic pattern to emerge.Figure 293-1 is an algorithm for the evaluation of chronically abnormal liver tests.

(Bibliography omitted in Palm version)

Back to Table of Contents

294. BILIRUBIN METABOLISM AND THE HYPERBILIRUBINEMIAS - *Paul D. Berk*, *Allan W. Wolkoff*

BILIRUBIN METABOLISM

SOURCES OF BILIRUBIN

Bilirubin is the end-product of the metabolic degradation of heme, the prosthetic group of hemoglobin, myoglobin, the cytochrome P450s, and various other hemoproteins. The first step in the conversion of heme to bilirubin is the stereospecific oxidative opening of the heme molecule at its a-bridge carbon by the microsomal enzyme *heme oxygenase*, resulting in the formation of equimolar quantities of carbon monoxide and of the green tetrapyrrole biliverdin. Biliverdin is then reduced by a second enzyme, biliverdin reductase, to bilirubin. Between 70 and 90% of bilirubin is derived from degradation of the hemoglobin of senescent or injured circulating red blood cells. The remainder has several sources, including hemoglobin produced during the process of ineffective erythropoiesis within the bone marrow and the turnover of nonhemoglobin hemoproteins in cells throughout the body. Degradation of red-cell hemoglobin occurs principally in the spleen but also throughout the rest of the peripheral reticuloendothelial system, including the Kupffer cells within the liver. Bilirubin produced in the periphery is transported to the liver within the plasma, where, due to its insolubility in aqueous solutions, it is tightly bound to albumin.

The anatomy of the hepatic acinus is highly specialized to facilitate the extraction of such tightly protein-bound compounds (Fig. 294-1). Cuboidal hepatocytes within the hepatic cell plates are immediately adjacent to sinusoids on two surfaces. The endothelial cells of the sinusoids are fenestrated, allowing ready exchange of plasma between the sinusoidal blood and the extracellular space of Disse and affording direct access of the bilirubin-albumin complex to the surface of the hepatocyte, which is greatly expanded by the elaboration of microvilli.

HEPATIC DISPOSITION OF BILIRUBIN

Since bilirubin is a potentially toxic waste product, hepatic handling is designed to eliminate it from the body via the biliary tract. Transfer of bilirubin from blood to bile involves four distinct but interrelated steps, described below (Fig. 294-1).

Hepatocellular Uptake Bilirubin most likely enters the hepatocyte both by a facilitated transport mechanism and by passive diffusion. While kinetic data suggest that facilitated transport is the predominant process and several putative bilirubin transporters have been identified, none has been cloned successfully. Cloned transporters such as *organic anion transport protein* (OATP) and *sodium taurocholate co-transporting polypeptide* (NTCP), which are responsible for the hepatocellular uptake of other organic substrates, including sulfobromophthalein and bile acids, specifically do not transport bilirubin. Therefore, the precise mechanism of bilirubin uptake remains to be determined.

Intracellular Binding Having crossed the plasma membrane to enter the cell, bilirubin partitions between the lipid environment of intracellular membranes and the aqueous cytosol, in which it is kept in solution by binding as a nonsubstrate ligand to several of the glutathione-S-transferases, formerly called ligandins.

Conjugation The aqueous insolubility of bilirubin reflects a rigid, highly ordered molecular structure in which internal hydrogen bonding involving the propionic acid carboxyl groups of one dipyrrolic half of the molecule and the imino and lactam groups of the opposite half blocks solvent access to these polar residues. When the carboxyl groups are esterified by conjugation with glucuronic acid residues, the internal hydrogen bonding is disrupted, rendering the resulting mono- and diglucuronide conjugates highly soluble in aqueous solution.

Bilirubin glucuronidation is catalyzed by a specific UDP-glucuronosyltransferase. The UDP-glucuronosyltransferases have been classified into gene families based on the degree of homology between the various protein isoforms. Those that conjugate bilirubin and certain other substrates have been designated the *UGT1* family and have been shown to be expressed from a single gene complex by alternative splicing. This gene complex contains multiple substrate-specific first exons, designated A1, A2, . . . (Fig. 294-2), each with its own promoter and each encoding the amino-terminal end of a specific isoform, as well as four common exons (exons 2 to 5) that encode the shared carboxyl-terminal end of all of the *UGT1* isoforms. The various first exons encode the specific substrate-binding sites for each isoform, while the shared exons encode common glycosylation, UDP-glucuronic acid-binding, transmembrane, and stop transfer domains. Exon A1 and the four common exons, collectively designated the *UGT1A1* gene (Fig. 294-2), encode the physiologically critical enzyme bilirubin-UDP-glucuronosyltransferase (UGT1A1). A critical corollary of the organization of the *UGT1* gene is that a mutation in one of the first exons will affect only a single enzyme isoform. By contrast, a mutation in exons 2 to 5 will alter all isoforms encoded by the *UGT1* gene complex.

Biliary Excretion Normal bile typically contains less than 5% unconjugated bilirubin, an average of 7% bilirubin monoconjugates, and 90% bilirubin diconjugates. The proportion of monoconjugates increases in the presence of an increased bilirubin load (hemolysis) or a reduced bilirubin-conjugating capacity. Bilirubin mono- and diglucuronides are excreted across the canalicular plasma membrane into the canaliculus by an ATP-dependent transport process mediated by a canalicular membrane protein called *multidrug resistance-associated protein 2* (MRP2). MRP2 is a member of the MRP gene family, other members of which pump certain types of drug conjugates, as well as unmodified anticancer drugs, out of cells. It is also a member of the ATP-binding cassette (ABC) superfamily. Mutations in the rat homologue of MRP2 result in conjugated hyperbilirubinemia in several jaundiced strains that serve as models of the Dubin-Johnson syndrome. It has recently been established that the Dubin-Johnson syndrome in humans also results from mutations in MRP2 (see below).

BILIRUBIN IN PLASMA

Although physicians equate the direct-reacting fraction of bilirubin in plasma with conjugated bilirubin and the indirect fraction with unconjugated bilirubin, modern analytical methods document that normal plasma contains virtually no bilirubin conjugates. The 10 to 20% of bilirubin in normal plasma that gives a prompt (direct) diazo reaction is an artifact of the kinetics of the van den Bergh reaction, which, along with various modifications, is the method most commonly used to quantitate bilirubin in clinical laboratories. Indeed, when the direct-reacting fraction is less than 15% of total bilirubin at virtually any total bilirubin concentration, the bilirubin in the sample can be considered as essentially all unconjugated. The canalicular transport mechanism for excretion of bilirubin conjugates is very sensitive to injury. Accordingly, in hepatocellular disease, as well as with either cholestasis or mechanical obstruction to the bile ducts, bilirubin conjugates within the hepatocyte, prevented from taking their normal path into the canaliculi and down the bile ducts, may reflux into the bloodstream, resulting in a mixed or, less often, a truly conjugated hyperbilirubinemia.

EXTRAHEPATIC ASPECTS OF BILIRUBIN DISPOSITION

Bilirubin in the Gut Following secretion into bile, conjugated bilirubin reaches the duodenum and passes down the gastrointestinal tract without reabsorption by the intestinal mucosa. Although some reaches the feces unaltered, an appreciable fraction is converted to urobilinogen and related compounds by bacterial metabolism within the ileum and colon. Urobilinogen is reabsorbed from these sites, reaches the liver via the portal circulation, and is reexcreted into bile, undergoing an enterohepatic circulation. Urobilinogen not taken up by the liver reaches the systemic circulation, from which some is cleared by the kidneys. Urinary urobilinogen excretion normally does not exceed 4 mg/d. In the presence of hemolysis, which increases the amount of bilirubin entering the gut (and hence the amount of urobilinogen formed and reabsorbed), or in the presence of hepatic disease, which decreases hepatic extraction of urobilinogen, plasma urobilinogen levels rise, as does the amount excreted in the urine. Severe cholestasis, bile duct obstruction, or administration of broad-spectrum antibiotics that eliminate the enteric flora required for the conversion of bilirubin to urobilinogens, markedly decrease formation of urobilinogen and its urinary excretion.

Unconjugated bilirubin ordinarily does not reach the gut except in neonates or, by ill-defined alternative pathways, in the presence of severe unconjugated hyperbilirubinemia (e.g., Crigler-Najjar syndrome type I). In these circumstances, however, unconjugated bilirubin is readily reabsorbed from the gut lumen, amplifying the underlying hyperbilirubinemia.

Renal Excretion of Bilirubin Conjugates Unconjugated bilirubin is not excreted in urine no matter how high its plasma concentration, since it is too tightly bound to albumin for effective glomerular filtration and there is no tubular mechanism for its renal secretion. By contrast, the polar bilirubin conjugates are far less tightly bound to albumin and are readily filtered at the glomerulus. Bilirubin conjugates are not secreted by the renal tubules but may be minimally reabsorbed. Since normal plasma contains virtually exclusively unconjugated bilirubin, no bilirubin normally appears in the urine. Indeed, bilirubinuria indicates the presence of conjugated bilirubin in plasma and, therefore, hepatobiliary dysfunction.

CLINICAL PHYSIOLOGY

The plasma concentration of unconjugated bilirubin ([Br]) is determined by the rate at which newly synthesized bilirubin enters the plasma (plasma bilirubin turnover, BrT) and hepatic bilirubin clearance (C_{Br}), according to the following relationship:

where k is a constant related to the different units of time employed in the conventional expression of BrTand C_{Br}. BrT closely reflects total bilirubin production; C_{Br}, analogous to the creatinine clearance test widely used to assess kidney function, is a measure of the rate at which bilirubin is extracted from plasma and is a true quantitative test of liver function. While not easily quantified in routine clinical settings, investigative measurements of BrT and C_{Br} have yielded useful pathophysiologic insights into the unconjugated hyperbilirubinemias.

Equation (1) indicates that the unconjugated bilirubin concentration will increase in the presence of either an increase inBrT or a reduction in hepatic C_{Br}. This equation therefore provides a basis for classifying unconjugated hyperbilirubinemias according to pathogenesis. Furthermore, for an individual with a given value for C_{Br}, or for a population in which C_{Br}varies within a narrow range, [Br] will increase as a linear function of BrT, with a slope relating increases in the plasma [Br] to increased BrT equal to k/C_{Br}. For individuals or populations with reduced bilirubin clearance (e.g., in Gilbert's syndrome, see below), this slope will be steeper than in normal individuals. Conversely, for a given BrT, the relationship between [Br] and C_{Br} is hyperbolic, like the relation between serum creatinine concentration and creatinine clearance. In any patient, if C_{Br} is reduced from its baseline value, [Br] will be increased in consequence, in direct proportion to the extent of the decrease in C_{Br}.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO UNCONJUGATED HYPERBILIRUBINEMIA

INCREASED BILIRUBIN PRODUCTION

Hemolysis Increased destruction of erythrocytes leads to increased bilirubin turnover and unconjugated hyperbilirubinemia. With normal liver function, the hyperbilirubinemia is usually modest. In particular, since the bone marrow is only capable of a sustained eightfold increase in erythrocyte production in response to a hemolytic stress, hemolysis alone cannot result in a sustained hyperbilirubinemia of more than approximately 68 umol/L (4 mg/dL). Higher values imply concomitant hepatic dysfunction.

The causes of hemolysis are numerous. Besides specific hemolytic disorders, mild hemolytic processes accompany many acquired systemic diseases. When hemolysis is the only abnormality in an otherwise healthy individual, the result is a purely unconjugated hyperbilirubinemia, with the direct-reacting fraction as measured in a typical clinical laboratory being £15% of the total serum bilirubin. In the presence of systemic disease, which may include a degree of hepatic dysfunction, hemolysis may produce a component of conjugated hyperbilirubinemia in addition to an elevated unconjugated bilirubin concentration.

Prolonged hemolysis may lead to the precipitation of bilirubin salts within the gall bladder or biliary tree, resulting in the formation of gallstones in which bilirubin, rather than cholesterol, is the major component. Such pigment stones may lead to acute or chronic cholecystitis, biliary obstruction, or any other biliary tract consequence of calculous disease.

Ineffective Erythropoiesis During erythroid maturation, small amounts of hemoglobin may be lost during nuclear extrusion, and a fraction of developing erythroid cells is destroyed within the marrow. These processes normally account for 10 to 15% of bilirubin produced. In various disorders, including thalassemia major, frankly megaloblastic anemias due to folate or vitamin B_{12}deficiency, congenital erythropoietic porphyria, lead poisoning, and various congenital and acquired dyserythropoietic anemias, the fraction of total bilirubin production derived from ineffective erythropoiesis is increased, reaching as much as 70% of the total, and may be sufficient to produce modest degrees of unconjugated hyperbilirubinemia.

Miscellaneous Degradation of the hemoglobin of extravascular collections of erythrocytes, such as those seen in massive tissue infarctions or large hematomas, may lead transiently to unconjugated hyperbilirubinemia.

DECREASED HEPATIC BILIRUBIN CLEARANCE

Decreased Hepatic Uptake As noted above, the mechanisms by which bilirubin enters hepatocytes are not fully defined but probably include both diffusion and facilitated transport. Decreased hepatic bilirubin uptake is believed to contribute to the unconjugated hyperbilirubinemia of Gilbert's syndrome (GS), although the molecular basis for this finding remains unclear (see below). Several drugs, including flavispidic acid, novobiocin, and various cholecystographic contrast agents, have been reported to inhibit bilirubin uptake. The resulting unconjugated hyperbilirubinemia resolves with cessation of the medication.

Impaired Conjugation

Physiologic Neonatal Jaundice Bilirubin produced by the fetus is cleared by the placenta and eliminated by the maternal liver. Consequently, bilirubin concentrations in normal neonates at birth are low. The presence of jaundice at birth is pathologic and requires investigation. Immediately after birth, the neonatal liver must assume responsibility for bilirubin clearance and excretion. However, many aspects of hepatic physiology are incompletely developed at birth. Levels of UGT1A1 are low, and alternative pathways allow passage of unconjugated bilirubin into the gut. Since the intestinal flora that converts bilirubin to urobilinogen is also undeveloped, an enterohepatic circulation of unconjugated bilirubin ensues. In consequence, most neonates develop mild unconjugated hyperbilirubinemia between days 2 and 5 after birth. Peak levels are typically less than 85 to 170 umol/L (5 to 10 mg/dL) and decline to normal adult concentrations within 2 weeks, as mechanisms required for bilirubin disposition mature.

Prematurity, with more profound immaturity of hepatic function, or hemolysis, such as occurs with erythroblastosis fetalis, results in higher levels of unconjugated hyperbilirubinemia. A rapidly rising unconjugated bilirubin concentration, or absolute levels in excess of 340 umol/L (20 mg/dL), puts the infant at risk for bilirubin encephalopathy, or *kernicterus*, in which bilirubin crosses an immature blood-brain barrier and precipitates in the basal ganglia and other areas of the brain. The consequences range from appreciable neurologic deficits to death. Principal treatment options include phototherapy, which converts bilirubin into photoisomers that are soluble in aqueous media and readily excretable in bile without conjugation, and exchange transfusion.

The canalicular mechanisms responsible for bilirubin excretion are also immature at birth, and their maturation may, on occasion, lag behind that of UGT1A1. This may lead to transient conjugated neonatal hyperbilirubinemia, especially in infants with hemolysis.

Acquired Conjugation Defects A modest reduction in bilirubin-conjugating capacity may be observed in advanced hepatitis or cirrhosis. However, in this setting, conjugation is better preserved than other aspects of bilirubin disposition, such as canalicular excretion. Various drugs, including pregnanediol, novobiocin, chloramphenicol, and gentamicin, may produce unconjugated hyperbilirubinemia by inhibiting UGT1A1 activity. Finally, certain fatty acids and the progestational steroid 3a,20b-pregnanediol, identified in the breast milk but not the serum of mothers whose infants have excessive neonatal hyperbilirubinemia (*breast milk jaundice*), inhibit bilirubin conjugation. The pathogenesis of breast milk jaundice appears to differ from that of transient familial neonatal hyperbilirubinemia (Lucey-Driscoll syndrome), in which a UGT1A1 inhibitor is found in maternal serum.

HEREDITARY DEFECTS IN BILIRUBIN CONJUGATION

Three familial disorders characterized by differing degrees of unconjugated hyperbilirubinemia have long been recognized. The defining clinical features of each are described below (Table 294-1). While these disorders have been recognized for decades to reflect differing degrees of deficiency in the ability to conjugate bilirubin, recent advances in the molecular biology of the *UGT1* gene complex have elucidated their interrelationships and clarified previously puzzling features.

Crigler-Najjar Syndrome, Type I (CN-I) This disorder is characterized by striking unconjugated hyperbilirubinemia of about 340 to 765 umol/L (20 to 45 mg/dL) that appears in the neonatal period and persists for life. Other conventional hepatic biochemical tests such as serum aminotransferases and alkaline phosphatase are normal, and there is no evidence of hemolysis. Hepatic histology is also essentially normal except for the occasional presence of bile plugs within canaliculi.

Bilirubin glucuronides are markedly reduced or absent from the nearly colorless bile, and there is no detectable constitutive expression of UGT1A1 activity in hepatic tissue. Neither UGT1A1 activity nor the serum bilirubin concentration responds to administration of phenobarbital or other enzyme inducers. In the absence of conjugation, unconjugated bilirubin accumulates in plasma, from which it is eliminated very slowly by alternative pathways that include direct passage into the bile and small intestine. These account for the small amounts of urobilinogen found in feces. No bilirubin is found in the urine.

First described in 1952, the disorder is rare (estimated prevalence of 0.6 to 1.0 per million). Many patients are from geographically or socially isolated communities in which consanguinity is common, and pedigree analyses suggest an autosomal recessive pattern of inheritance. The majority of patients (type IA) exhibit defects in the glucuronide conjugation of a spectrum of substrates in addition to bilirubin, including various drugs and other xenobiotics. These individuals have mutations in one of the common exons (2 to 5) of the *UGT1* gene (Fig. 294-2). In a smaller subset (type IB), the defect is limited largely to bilirubin conjugation, and the causative mutation is in the bilirubin-specific exon A1. More than 30 different *UGT1A1* mutations responsible forCN-Ihave been identified, including deletions, frameshifts, alterations in intronic splice donor and acceptor sites, and point mutations that introduce premature stop codons or alter critical aminoacids. Their common feature is that they all encode proteins with absent or, at most, traces of bilirubin-UDP-glucuronosyltransferase enzymatic activity.

Prior to the availability of phototherapy, most patients withCN-Idied of bilirubin encephalopathy (kernicterus) in infancy or early childhood. A few lived as long as early adult life without overt neurologic damage, although more subtle testing usually indicated mild but progressive brain damage. In all such cases, in the absence of liver transplantation, death eventually supervened from late-onset bilirubin encephalopathy, which often followed a nonspecific febrile illness. Recent data suggest that the best hope for survival of a neurologically intact patient involves the following regimen: (1) about 12 h/d of phototherapy from birth throughout childhood, perhaps supplemented by exchange transfusion in the immediate neonatal period; (2) use of tin-protoporphyrin to blunt transient episodes of increased hyperbilirubinemia; and (3) early liver transplantation, prior to the onset of brain damage. In a single patient, transplantation with isolated allogeneic hepatocytes produced a clinically significant reduction in serum bilirubin concentration.

Crigler-Najjar Syndrome, Type II (CN-II) Characterized by marked unconjugated hyperbilirubinemia in the absence of abnormalities of other conventional hepatic biochemical tests, hepatic histology, or hemolysis, this condition was recognized as a distinct entity in 1962. It differs fromCN-I in several specific ways (Table 294-1). (1) Although there is considerable overlap, average bilirubin concentrations are lower in CN-II; (2) accordingly, CN-II is only infrequently associated with kernicterus; (3) bile is deeply colored and bilirubin glucuronides are present, with a striking, characteristic increase in monoglucuronides; (4) UGT1A1 in liver is usually present at reduced levels (typically £10% of normal) but may be undetectable by less sensitive older assays; (5) while typically detected in infancy, hyperbilirubinemia was not recognized in some cases until later in life, and in one instance, until age 34. As with CN-I, most CN-II cases exhibit abnormalities in the conjugation of other compounds, such as salicylamide and menthol, but in some instances the defect appears limited to bilirubin.

Reduction of serum bilirubin concentrations by more than 25% in response to enzyme inducers such as phenobarbital distinguishesCN-IIfromCN-I, although this response may not be elicited in early infancy and often is not accompanied by measurable UGT1A1 induction. Bilirubin concentrations during phenobarbital administration do not return to normal but are typically in the range of 51 to 86 umol/L (3 to 5 mg/dL). Although the incidence of kernicterus in CN-II is low, instances have occurred, not only in infants but in adolescents and adults, often in the setting of an intercurrent illness, fasting, or any other factor that temporarily raises the serum bilirubin concentration above baseline. For this reason, phenobarbital therapy is widely recommended, a single bedtime dose often sufficing to maintain clinically safe plasma bilirubin concentrations.

At least 10 different mutations of *UGT1* associated withCN-IIhave been identified. Their common feature is that they encode for a bilirubin-UDP-glucuronosyltransferase with markedly reduced but detectable enzymatic activity. The spectrum of residual enzyme activity explains the spectrum of phenotypic severity of the resulting hyperbilirubinemia. Molecular analysis has established that a large majority of CN-II patients are either homozygotes or compound heterozygotes for CN-II mutations and that individuals carrying one mutated and one entirely normal allele have normal bilirubin concentrations. Possible inheritance in one case as a dominant negative mutation remains to be confirmed.

Gilbert's Syndrome This syndrome is characterized by mild unconjugated hyperbilirubinemia, normal values for standard hepatic biochemical tests, and normal hepatic histology other than a modest increase of lipofuscin pigment in some patients. Serum bilirubin concentrations are most often<51 umol/L (<3 mg/dL), although both higher and lower values are frequent. The spectrum of hyperbilirubinemia fades into that ofCN-II at serum bilirubin concentrations of 86 to 136 umol/L (5 to 8 mg/dL). At the other end of the scale, the distinction between mild cases ofGS and a normal state is often blurred. Bilirubin concentrations may fluctuate substantially in any given individual, and at least 25% of patients will exhibit temporarily normal values during prolonged follow-up. More elevated values are associated with stress, fatigue, alcohol use, reduced caloric intake, and intercurrent illness, while increased caloric intake or administration of enzyme-inducing agents produce lower bilirubin levels. GS is most often diagnosed at or shortly after puberty or in adult life during routine examinations that include multichannel biochemical analyses.

UGT1A1 activity is typically reduced to 10 to 35% of normal, and bile pigments in bile exhibit a characteristic increase in bilirubin monoglucuronides. Studies of radiobilirubin kinetics indicate that hepatic bilirubin clearance is reduced to an average of one-third of normal. Administration of phenobarbital normalizes both the serum bilirubin concentration and hepatic bilirubin clearance. However, failure of UGT1A1 activity to improve in many such instances suggests the possible coexistence of an additional defect. Compartmental analysis of bilirubin kinetic data suggests that GSpatients have a defect in bilirubin uptake as well as in conjugation. Defect(s) in the hepatic uptake of other organic anions that at least partially share an uptake mechanism with bilirubin, such as sulfobromophthalein and indocyanine green, are observed in some, but not all, patients. The disposition of bile acids, which do not utilize the bilirubin uptake mechanism, is normal.

The magnitude of changes in the plasma bilirubin concentration induced by provocation tests such as 48 h of fasting or the intravenous administration of nicotinic acid have been reported to be of help in separatingGSpatients from normal individuals. Other studies dispute this assertion. Moreover, on theoretical grounds, the results of such studies should provide no more information than simple measurements of the baseline plasma bilirubin concentration.

Family studies indicate thatGS and hereditary hemolytic anemias such as hereditary spherocytosis, glucose-6-phosphate dehydrogenase deficiency, andb-thalassemia trait sort independently. Reports of hemolysis in up to 50% of GS patients are believed to reflect better case finding, since patients with both GS and hemolysis have higher bilirubin concentrations, and are more likely to be jaundiced, than patients with either defect alone.

GSis common, with many series placing its prevalence at 8% or more. Males predominate over females by reported ratios ranging from 1.5:1 to more than 7:1. However, these ratios may have a large artifactual component since normal males have higher mean bilirubin levels than normal females, but the diagnosis of GS is often based on comparison to normal ranges established in men. The high prevalence of GS in the general population may explain the reported frequency of mild unconjugated hyperbilirubinemia in liver transplant recipients.

The disposition of most xenobiotics metabolized by glucuronidation appears to be normal inGS, as is oxidative drug metabolism in the majority of reported studies. The principal exception is the metabolism of the anti-tumor agent irinotecan (CPT-11). Its active metabolite (SN-38) is glucuronidated specifically by bilirubin-UDP-glucuronosyltransferase. Administration of CPT-11 to patients with GS has resulted in several toxicities, including intractable diarrhea and myelosuppression. Some reports also suggest abnormal disposition of menthol, estradiol benzoate, acetaminophen, tolbutamide, and rifamycin SV. Although some of these studies have been disputed, and there have been no reports of clinical complications from use of these agents in GS, prudence should be exercised in prescribing them, or any agents metabolized primarily by glucuronidation, in this condition.

Most older pedigree studies ofGS were consistent with autosomal dominant inheritance with variable expressivity. However, studies of the *UGT1* gene in GS have indicated a variety of molecular genetic bases for the phenotypic picture and several different patterns of inheritance. Studies in European and U.S. patients found that the majority of GS patients had normal coding regions for UGT1A1 but were homozygous for an abnormality consisting of an extra TA (i.e., A[TA]$_7$TAA rather than A[TA]$_6$TAA) in the promoter region of the first exon. This appeared to be a necessary but not a sufficient genetic basis for clinically expressed GS, since 15% of normal controls were also homozygous for this variant. While normal by standard criteria, these individuals had somewhat higher bilirubin concentrations than the rest of the controls studied. Heterozygotes for this abnormality had bilirubin concentrations identical to those homozygous for the A[TA]$_6$TAA allele. The prevalence of the A[TA]$_7$TAA allele in a general western population is 30%, in which case 9% would be homozygotes. This is slightly higher than the prevalence of GS based on purely phenotypic parameters. It was suggested that additional variables, such as mild hemolysis or a defect in bilirubin uptake, might be among the factors enhancing phenotypic expression of the defect. Phenotypic expression of GS due solely to the A[TA]$_7$TAA promoter abnormality is inherited as an autosomal recessive trait.

A number ofCN-IIkindreds have been identified in which there is also an allele containing a normal coding region but the A[TA]$_7$TAA promoter abnormality. CN-II heterozygotes who have the A[TA]$_6$TAA promoter are phenotypically normal, whereas those with the A[TA]$_7$TAA promoter express the phenotypic picture ofGS. GS in such kindreds may also result from homozygosity for the A[TA]$_7$TAA promoter abnormality.

Seven different missense mutations in the *UGT1* gene that reportedly causeGS with dominant inheritance have been found in Japanese individuals. Another Japanese patient with mild unconjugated hyperbilirubinemia was homozygous for a missense mutation in exon 5. GS in her family appeared to be recessive. Missense mutations causing GS have not been reported outside of Japan.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO MIXED OR PREDOMINANTLY CONJUGATED HYPERBILIRUBINEMIA

In hyperbilirubinemia due to acquired liver disease (e.g., acute hepatitis, common bile duct stone), there are usually elevations in the serum concentrations of both conjugated and unconjugated bilirubin. Although biliary tract obstruction or hepatocellular cholestatic injury may present on occasion with a predominantly conjugated hyperbilirubinemia, it is generally not possible to differentiate intrahepatic from extrahepatic causes of jaundice based upon the serum levels or relative proportions of unconjugated and conjugated bilirubin. The major reason for determining the amounts of conjugated and unconjugated bilirubin in the serum is for the initial differentiation of hepatic parenchymal and obstructive disorders (mixed conjugated and unconjugated hyperbilirubinemia) from the inheritable and hemolytic disorders discussed above that are associated with unconjugated hyperbilirubinemia.

FAMILIAL DEFECTS IN HEPATIC EXCRETORY FUNCTION

Dubin-Johnson Syndrome This benign, relatively rare disorder is characterized by low-grade, predominantly conjugated hyperbilirubinemia. Total bilirubin concentrations are typically between 34 and 85 umol/L (2 and 5 mg/dL) but on occasion can be in the normal range or as high as 340 to 430 umol/L (20 to 25 mg/dL) and can fluctuate widely in any given patient. The degree of hyperbilirubinemia may be increased by intercurrent illness, oral contraceptive use, and pregnancy. As the hyperbilirubinemia is due to a predominant rise in conjugated bilirubin, bilirubinuria is characteristically present. Aside from elevated serum bilirubin levels, other routine laboratory tests are normal. Physical examination is usually normal except for jaundice, although an occasional patient may have hepatosplenomegaly.

Patients with Dubin-Johnson syndrome are usually asymptomatic, although some may have vague constitutional symptoms. These latter patients have usually undergone extensive and often unnecessary diagnostic examinations for unexplained jaundice and have high levels of anxiety. In women, the condition may be subclinical until the patient becomes pregnant or receives oral contraceptives, at which time chemical hyperbilirubinemia becomes frank jaundice. Even in these situations, other routine liver function tests, including serum alkaline phosphatase and transaminase activities, are normal.

A cardinal feature of Dubin-Johnson syndrome is the accumulation in the lysosomes of centrilobular hepatocytes of dark, coarsely granular pigment. As a result, the liver may be grossly black in appearance. This pigment is thought to be derived from epinephrine metabolites that are not excreted normally. The pigment may disappear during bouts of viral hepatitis, only to reaccumulate slowly after recovery.

Biliary excretion of a number of anionic compounds is compromised in Dubin-Johnson syndrome. These include various cholecystographic agents, as well as sulfobromophthalein (Bromsulphalein, BSP), a synthetic dye formerly used in a test of liver function. In this test, the rate of disappearance of BSP from plasma was determined following bolus intravenous administration. BSP is conjugated with glutathione in the hepatocyte; the resulting conjugate is normally excreted rapidly into the canaliculus. Patients with Dubin-Johnson syndrome exhibit a characteristic rise in its plasma concentration at 90 min after injection, due to reflux of conjugated BSP into the circulation from the hepatocyte. Dyes such as indocyanine green (ICG) that are taken up by hepatocytes but are not further metabolized prior to biliary excretion do not show this reflux phenomenon. Continuous BSP infusion studies suggest a reduction in the $t_{\max}$for biliary excretion. Bile acid disposition, including hepatocellular uptake and biliary excretion, are normal in Dubin-Johnson syndrome. These patients have normal serum and biliary bile acid concentrations and do not have pruritus.

By analogy with findings in several mutant rat strains, the selective defect in biliary excretion of bilirubin conjugates and certain other classes of organic compounds, but not of bile acids, that characterizes the Dubin-Johnson syndrome was found to reflect defective expression of MRP2, an ATP-dependent canalicular membrane transporter. Several different mutations in the *MRP2* gene produce the Dubin-Johnson phenotype, which has an autosomal recessive pattern of inheritance. Although MRP2 is undoubtedly important in the biliary excretion of conjugated bilirubin, the fact that this pigment is still excreted in the absence of MRP2 suggests that other, as yet uncharacterized, transport proteins may serve in a secondary role in this process.

Patients with Dubin-Johnson syndrome also have a diagnostic abnormality in urinary coproporphyrin excretion. There are two naturally occurring coproporphyrin isomers, I and III. Normally, approximately 75% of the coproporphyrin in urine is isomer III. In urine from Dubin-Johnson syndrome patients, total coproporphyrin content is normal, but more than 80% is isomer I. Heterozygotes for the syndrome show an intermediate pattern. The molecular basis for this phenomenon remains unclear.

Rotor Syndrome This benign, autosomal recessive disorder is clinically similar to the Dubin-Johnson syndrome, although it is seen even less frequently. A major phenotypic difference is that the liver in patients with Rotor syndrome has no increased pigmentation and appears totally normal. The only abnormality in routine laboratory tests is an elevation of total serum bilirubin, due to a predominant rise in conjugated bilirubin. This is accompanied by bilirubinuria. Several additional features differentiate Rotor and Dubin-Johnson syndromes. In Rotor syndrome, the gallbladder is usually visualized on oral cholecystography, in contrast to the nonvisualization that is typical of Dubin-Johnson syndrome. The pattern of urinary coproporphyrin excretion also differs. The pattern in Rotor syndrome resembles that of many acquired disorders of hepatobiliary function, in which coproporphyrin I, the major coproporphyrin isomer in bile, refluxes from the hepatocyte back into the circulation and is excreted in urine. Thus, total urinary coproporphyrin excretion is substantially increased in Rotor syndrome, in contrast to the normal levels seen in Dubin-Johnson syndrome. Although the fraction of coproporphyrin I in urine is elevated, it is usually less than 70% of the total, as compared to 80% or more in Dubin-Johnson syndrome. The disorders also can be distinguished by their patterns ofBSPexcretion. Although clearance of BSP from plasma is delayed in Rotor syndrome, there is no reflux of conjugated BSP back into the circulation as seen in Dubin-Johnson syndrome. Kinetic analysis of plasma BSP infusion studies suggests the presence of a defect in intrahepatocellular storage of this compound. This has never been demonstrated directly, and the molecular basis of Rotor syndrome remains unknown.

Benign Recurrent Intrahepatic Cholestasis (BRIC) This rare disorder is characterized by recurrent attacks of pruritus and jaundice. The typical episode begins with mild malaise and elevations in serum aminotransferase levels, followed rapidly by rises in alkaline phosphatase and bilirubin and onset of jaundice and itching. The first one or two episodes may be misdiagnosed as acute viral hepatitis. The cholestatic episodes, which may begin in childhood or adulthood, can vary in duration from several weeks to months, following which there is complete clinical and biochemical resolution. Intervals between attacks may vary from several months to years. Between episodes, physical examination is normal, as are serum levels of bile acids, bilirubin, transaminases, and alkaline phosphatase. The disorder is familial and has an autosomal recessive pattern of inheritance.BRIC is considered a benign disorder in that it does not lead to cirrhosis or end-stage liver disease. However, the episodes of jaundice and pruritus can be prolonged and debilitating, and some patients have undergone liver transplantation to relieve the intractable and disabling symptoms. Treatment during the cholestatic episodes is symptomatic; there is no specific treatment to prevent or shorten the occurrence of episodes.

A gene termed *FIC1* was recently identified and found to be mutated in patients withBRIC. Curiously, this gene is expressed strongly in the small intestine but only weakly in the liver. The protein encoded by *FIC1* shows little similarity to genes that have been shown to play a role in bile canalicular excretion of various compounds. Rather, it appears to be a member of a P-type ATPase family that transports aminophospholipids from the outer to the inner leaflet of a variety of cell membranes.

Progressive Familial Intrahepatic Cholestasis (FIC) This name is applied to three phenotypically related syndromes. Progressive FIC type 1 (Byler disease) presents in early infancy as cholestasis that may be initially episodic. However, in contrast toBRIC, Byler disease progresses to malnutrition, growth retardation, and end-stage liver disease during childhood. This disorder is also a consequence of an FIC1 mutation. The functional relationship of the *FIC1* protein to the pathogenesis of cholestasis in these disorders is unknown. Two other types of progressive FIC (types 2 and 3) have been described. Type 2 is associated with a mutation in the protein named *sister of p-glycoprotein*, which is the major bile canalicular exporter of bile acids. Type 3 has been associated with a mutation of MDR3, a protein that is essential for normal bile canalicular excretion of phospholipids. Although all three types of progressive FIC have similar clinical phenotypes, only type 3 is associated with high serum levels ofg-glutamyltransferase activity. In contrast, activity of this enzyme is normal or only mildly elevated in symptomatic BRIC and progressive FIC types 1 and 2.

(Bibliography omitted in Palm version)

Back to Table of Contents

295. ACUTE VIRAL HEPATITIS - *Jules L. Dienstag*, *Kurt J. Isselbacher*

Acute viral hepatitis is a systemic infection affecting the liver predominantly. Almost all cases of acute viral hepatitis are caused by one of five viral agents: hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), the HBV-associated delta agent or hepatitis D virus (HDV), and hepatitis E virus (HEV). Other transfusion-transmitted agents, e.g., "hepatitis G" virus and "TT" virus, have been identified but do not cause hepatitis. All these human hepatitis viruses are RNA viruses, except for hepatitis B, which is a DNA virus. Although these agents can be distinguished by their molecular and antigenic properties, all types of viral hepatitis produce clinically similar illnesses. These range from asymptomatic and inapparent to fulminant and fatal acute infections common to all types, on the one hand, and from subclinical persistent infections to rapidly progressive chronic liver disease with cirrhosis and even hepatocellular carcinoma, common to the bloodborne types (HBV, HCV, and HDV), on the other.

VIROLOGY AND ETIOLOGY

Hepatitis A Hepatitis A virus is a nonenveloped 27-nm, heat-, acid-, and ether-resistant RNA virus in the hepatovirus genus of the picornavirus family (Fig. 295-1). Its virion contains four capsid polypeptides, designated VP1 to VP4, which are cleaved posttranslationally from the polyprotein product of a 7500-nucleotide genome. Inactivation of viral activity can be achieved by boiling for 1 min, by contact with formaldehyde and chlorine, or by ultraviolet irradiation. Despite nucleotide sequence variation of up to 20% among isolates ofHAV, all strains of this virus are immunologically indistinguishable and belong to one serotype. Hepatitis A has an incubation period of approximately 4 weeks. Its replication is limited to the liver, but the virus is present in the liver, bile, stools, and blood during the late incubation period and acute preicteric phase of illness. Despite persistence of virus in the liver, viral shedding in feces, viremia, and infectivity diminish rapidly once jaundice becomes apparent. HAV is the only one of the human hepatitis viruses that can be cultivated reproducibly in vitro.

Antibodies toHAV(anti-HAV) can be detected during acute illness when serum aminotransferase activity is elevated and fecal HAV shedding is still occurring. This early antibody response is predominantly of the IgM class and persists for several months, rarely for 6 to 12 months. During convalescence, however, anti-HAV of the IgG class becomes the predominant antibody (Fig. 295-2). Therefore, the diagnosis of hepatitis A is made during acute illness by demonstrating anti-HAV of the IgM class. After acute illness, anti-HAV of the IgG class remains detectable indefinitely, and patients with serum anti-HAV are immune to reinfection. Neutralizing antibody activity parallels the appearance of anti-HAV, and the IgG anti-HAV present in immune globulin accounts for the protection it affords against HAV infection.

Hepatitis B Hepatitis B virus is a DNA virus with a remarkably compact genomic structure; despite its small, circular, 3200-basepair size,HBV DNA codes for four sets of viral products and has a complex, multiparticle structure. HBV achieves its genomic economy by relying on an efficient strategy of encoding proteins from four overlapping genes: S, C, P, and X (Fig. 295-3), as detailed below. Once thought to be unique among viruses, HBV is now recognized as one of a family of animal viruses, hepadnaviruses (hepatotropic DNA viruses), and is classified as hepadnavirus type 1. Similar viruses infect certain species of woodchucks, ground and tree squirrels, and Pekin ducks, to mention the most carefully characterized. Like HBV, all have the same distinctive three morphologic forms, have counterparts to the envelope and nucleocapsid virus antigens of HBV, replicate in the liver but exist in extrahepatic sites, contain their own endogenous DNA polymerase, have partially double-stranded and partially single-stranded genomes, are associated with acute and chronic hepatitis and hepatocellular carcinoma, and rely on a replicative strategy unique among DNA viruses but typical of retroviruses. Instead of DNA replication directly from a DNA template, hepadnaviruses rely on reverse transcription (effected by the DNA polymerase) of minus-strand DNA from a "pregenomic" RNA intermediate. Then plus-strand DNA is transcribed from the minus-strand DNA template by the DNA-dependent DNA polymerase. Viral proteins are translated by the pregenomic RNA, and the proteins and genome are packaged into virions and secreted from the hepatocyte. Although HBV is difficult to cultivate in vitro in the conventional sense from clinical material, several cell lines have been transfected with HBV DNA. Such transfected cells support in vitro replication of the intact virus and its component proteins.

Viral proteins and particles Three particulate forms ofHBV(Table 295-1) can be demonstrated by electron microscopy (Fig. 295-1). The most numerous are the 22-nm particles, which appear as spherical or long filamentous forms; these are antigenically indistinguishable from the outer surface or envelope protein of HBV and are thought to represent excess viral envelope protein. Outnumbered in serum by a factor of 100 or 1000 to 1 compared with the spheres and tubules are large, 42-nm, double-shelled spherical particles, which represent the intact hepatitis B virion. The envelope protein expressed on the outer surface of the virion and on the smaller spherical and tubular structures is referred to as *hepatitis B surface antigen* (HBsAg). The concentration of HBsAg and virus particles in the blood may reach 500 ug/mL and 10 trillion particles per milliliter, respectively. The envelope protein, HBsAg, is the product of the S gene of HBV.

A number of differentHBsAgsubdeterminants have been identified. There is a common group-reactive antigen, *a*, shared by all HBsAg isolates. In addition, HBsAg may contain one of several subtype-specific antigens, namely, *d* or *y*, *w* or *r*, as well as other more recently characterized specificities. Hepatitis B isolates fall into one of at least eight subtypes and six genotypes (A-F); however, clinical course and outcome are independent of subtype and genotype [except for an increase in "precore" mutations (see below) in certain genotypes].

Upstream of the S gene are the pre-S genes (Fig. 295-3), which code for pre-S gene products, including receptors on theHBVsurface for polymerized human serum albumin and for hepatocyte membrane proteins. The pre-S region actually consists of both pre-S1 and pre-S2. Depending on where translation is initiated, three potentialHBsAggene products are synthesized. The protein product of the S gene is HBsAg (*major protein*), the product of the S region plus the adjacent pre-S2 region is the *middle protein*, and the product of the pre-S1 plus pre-S2 plus S regions is the *large protein*. Compared with the smaller spherical and tubular particles of HBV, complete 42-nm virions are enriched in the large protein. Both pre-S proteins and their respective antibodies can be detected during HBV infection, and the period of pre-S antigenemia appears to coincide with other markers of virus replication, as detailed below.

The intact 42-nm virion can be disrupted by mild detergents, and the 27-nm nucleocapsid core particle isolated. Nucleocapsid proteins are coded for by the C gene. The antigen expressed on the surface of the nucleocapsid core is referred to as *hepatitis B core antigen* (HBcAg), and its corresponding antibody is anti-HBc. A thirdHBVantigen is *hepatitis B e antigen* (HBeAg), a soluble, nonparticulate, nucleocapsid protein that is immunologically distinct from intact HBcAg but is a product of the same C gene. The C gene has two initiation codons, a precore and a core region (Fig. 295-3). If translation is initiated at the precore region, the protein product is HBeAg, which has a signal peptide that binds it to the smooth endoplasmic reticulum and leads to its secretion into the circulation. If translation begins with the core region, HBcAg is the protein product; it has no signal peptide, it is not secreted, but it assembles into nucleocapsid particles, which bind to and incorporate RNA and which, ultimately, contain HBV DNA. Also packaged within the nucleocapsid core is a DNA polymerase, which directs replication and repair of HBV DNA. When packaging within viral proteins is complete, synthesis of the incomplete plus strand stops; this accounts for the single-stranded gap and for differences in the size of the gap. HBcAg particles remain in the hepatocyte, where they are readily detectable by immunohistochemical staining, and are exported after encapsidation by an envelope ofHBsAg. Therefore, naked core particles do not circulate in the serum. The secreted nucleocapsid protein, HBeAg, provides a convenient, readily detectable, qualitative marker of HBV replication and relative infectivity.

HBsAg-positive serum containingHBeAg is more likely to be highly infectious and to be associated with the presence of hepatitis B virions (and detectableHBV DNA, see below) than HBeAg-negative or anti-HBe-positive serum. For example, HBsAg carrier mothers who are HBeAg-positive almost invariably (>90%) transmit hepatitis B infection to their offspring, whereas HBsAg carrier mothers with anti-HBe rarely (10 to 15%) infect their offspring.

Early during the course of acute hepatitis B,HBeAgappears transiently; its disappearance may be a harbinger of clinical improvement and resolution of infection. Persistence of HBeAg in serum beyond the first 3 months of acute infection may be predictive of the development of chronic infection, and the presence of HBeAg during chronic hepatitis B is associated with ongoing viral replication, infectivity, and inflammatory liver injury.

The third of theHBVgenes is the largest, the P gene (Fig. 295-3), which codes for the DNA polymerase; as noted above, this enzyme has both DNA-dependent DNA polymerase and RNA-dependent reverse transcriptase activities. The fourth gene, X, codes for a small, nonparticulate protein that is capable of transactivating the transcription of both viral and cellular genes (Fig. 295-3). Such transactivation may enhance the replication of HBV, leading to the clinical association observed between the expression of the product of the X gene, hepatitis B x antigen (HBxAg), and antibodies to it in patients with severe chronic hepatitis and hepatocellular carcinoma. The transactivating activity can enhance the transcription and replication of other viruses besides HBV, such as HIV. Cellular processes transactivated by X include the human interferon ggene and class I major histocompatibility genes; potentially, these effects could contribute to enhanced susceptibility of HBV-infected hepatocytes to cytolytic T cells. The expression of X can also induce programmed cell death (apoptosis). The X gene and its protein product, however, are absent in nonmammalian hepadnaviruses; therefore, X is not essential for hepadnavirus replication.

Serologic and virologic markers After infection withHBV, the first virologic marker detectable in serum isHBsAg(Fig. 295-4). Circulating HBsAg precedes elevations of serum aminotransferase activity and clinical symptoms and remains detectable during the entire icteric or symptomatic phase of acute hepatitis B and beyond. In typical cases, HBsAg becomes undetectable 1 to 2 months after the onset of jaundice and rarely persists beyond 6 months. After HBsAg disappears, antibody to HBsAg (anti-HBs) becomes detectable in serum and remains detectable indefinitely thereafter. BecauseHBcAg is sequestered within an HBsAg coat, HBcAg is not detectable routinely in the serum of patients with HBV infection. By contrast, anti-HBc is readily demonstrable in serum, beginning within the first 1 to 2 weeks after the appearance of HBsAg and preceding detectable levels of anti-HBs by weeks to months. Because variability exists in the time of appearance of anti-HBs after HBV infection, occasionally a gap of several weeks or longer may separate the disappearance of HBsAg and the appearance of anti-HBs. During this "gap" or "window" period, anti-HBc may represent serologic evidence of current or recent HBV infection, and blood containing anti-HBc in the absence of HBsAg and anti-HBs has been implicated in the development of transfusion-associated hepatitis B. In part because the sensitivity of immunoassays for HBsAg and anti-HBs has increased, however, this window period is rarely encountered. In some persons, years after HBV infection, anti-HBc may persist in the circulation longer than anti-HBs. Therefore, isolated anti-HBc does not necessarily indicate active virus replication; most instances of isolated anti-HBc represent hepatitis B infection in the remote past. Rarely, however, isolated anti-HBc represents low-level hepatitis B viremia, with HBsAg below the detection threshold; occasionally, isolated anti-HBc represents a cross-reacting or false-positive immunologic specificity. Recent and remote HBV infections can be distinguished by determination of the immunoglobulin class of anti-HBc. Anti-HBc of the IgM class (IgM anti-HBc) predominates during the first 6 months after acute infection, whereas IgG anti-HBc is the predominant class of anti-HBc beyond 6 months. Therefore, patients with current or recent acute hepatitis B, including those in the anti-HBc window, have IgM anti-HBc in their serum. In patients who have recovered from hepatitis B in the remote past as well as those with chronic HBV infection, anti-HBc is predominantly of the IgG class. Infrequently, in no more than 1 to 5% of patients with acute HBV infection, levels of HBsAg are too low to be detected; in such cases, the presence of IgM anti-HBc establishes the diagnosis of acute hepatitis B. When isolated anti-HBc occurs in the rare patient with chronic hepatitis B whose HBsAg level is below the sensitivity threshold of contemporary immunoassays (a low-level carrier), the anti-HBc is of the IgG class. Generally, in persons who have recovered from hepatitis B, anti-HBs and anti-HBc persist indefinitely.

The temporal association between the appearance of anti-HBs and resolution ofHBVinfection as well as the observation that persons with anti-HBs in serum are protected against reinfection with HBV suggest that *anti-HBs is the protective antibody*. Therefore, strategies for prevention of HBV infection are based on providing susceptible persons with circulating anti-HBs (see below). Occasionally, in 10 to 20% of patients with chronic hepatitis B, low-level, low-affinity anti-HBs can be detected. This antibody is directed against a subtype determinant different from that represented by the patient'sHBsAg; its presence is thought to reflect the stimulation of a related clone of antibody-forming cells, but it has no clinical relevance and does not signal imminent clearance of hepatitis B.

The other readily detectable serologic marker ofHBVinfection,HBeAg, appears concurrently with or shortly after HBsAg. Its appearance coincides temporally with high levels of virus replication and reflects the presence of circulating intact virions and detectable HBV DNA. Pre-S1 and pre-S2 proteins are also expressed during periods of peak replication, but assays for these gene products are not routinely available. In self-limited HBV infections, HBeAg becomes undetectable shortly after peak elevations in aminotransferase activity, before the disappearance of HBsAg, and anti-HBe then becomes detectable, coinciding with a period of relatively lower infectivity (Fig. 295-4). Because markers of HBV replication appear transiently during acute infection, testing for such markers is of little clinical utility in typical cases of acute HBV infection. In contrast, markers of HBV replication provide valuable information in patients with protracted infections.

Departing from the pattern typical of acute HBV infections, in chronic HBV infection, HBsAg remains detectable beyond 6 months, anti-HBc is primarily of the IgG class, and anti-HBs is either undetectable or detectable at low levels (see "Laboratory Features," below) (Fig. 295-5). During early chronic HBV infection, HBV DNA can be detected both in serum and in hepatocyte nuclei, where it is present in free or episomal form. This *replicative stage* of HBV infection is the time of maximal infectivity and liver injury; HBeAg is a qualitative marker and HBV DNA a quantitative marker of this replicative phase, during which all three forms of HBV circulate, including intact virions. Over time, the replicative phase of chronic HBV infection gives way to a relatively *nonreplicative phase*. This occurs at a rate of approximately 10% per year and is accompanied by seroconversion from HBeAg-positive to anti-HBe-positive. In most cases, this seroconversion coincides with a transient, acute hepatitis-like elevation in aminotransferase activity, believed to reflect cell-mediated clearance of virus-infected hepatocytes. In the nonreplicative phase of chronic infection, when HBV DNA is demonstrable in hepatocyte nuclei, it tends to be integrated into the host genome. In this phase, only spherical and tubular forms of HBV, *not intact virions*, circulate, and liver injury tends to subside. Most such patients would be characterized as asymptomatic HBV *carriers*. In reality, the designations *replicative* and *nonreplicative* are only relative; even in the so-called nonreplicative phase, HBV replication can be detected with highly sensitive amplification probes such as the polymerase chain reaction. Still, the distinctions are pathophysiologically and clinically meaningful. Occasionally, nonreplicative HBV infection converts back to replicative infection. Such spontaneous reactivations are accompanied by reexpression of HBeAg and HBV DNA, and sometimes of IgM anti-HBc, as well as by exacerbations of liver injury.

Molecular variants Variation occurs throughout theHBVgenome, and clinical isolates of HBV that do not express typical viral proteins have been attributed to mutations in individual or even multiple gene locations. For example, variants have been described that lack nucleocapsid proteins, envelope proteins, or both. Two categories of HBV have attracted the most attention. One of these was identified initially in Mediterranean countries among patients with an unusual serologic-clinical profile. They have severe chronic HBV infection and detectable HBV DNA but with anti-HBe instead ofHBeAg.

These patients were found to be infected with an HBV mutant that contained an alteration in the precore region rendering the virus incapable of encoding HBeAg. Although several potential mutation sites exist in the pre-C region, the region of the C gene necessary for the expression of HBeAg (see "Virology and Etiology," above), the most commonly encountered in such patients is a single base substitution, from G to A, which occurs in the second to last codon of the pre-C gene at nucleotide 1896. This substitution results in the replacement of the TGG tryptophan codon by a stop codon (T*A*G), which prevents the translation of HBeAg. Another mutation in the core promoter region prevents transcription of the coding region for HBeAg and yields an HBeAg-negative phenotype. Patients with such precore mutants that are unable to secrete HBeAg tend to have severe liver disease that progresses rapidly to cirrhosis and that does not respond readily to antiviral therapy. Both "wild-type" HBV and precore mutant HBV can coexist in the same patient, or mutant HBV may arise during wild-type HBV infection. In addition, clusters of fulminant hepatitis B in Israel and Japan have been attributed to common-source infection with a precore mutant. Fulminant hepatitis B in North America and western Europe, however, occurs in patients infected with wild-type HBV, in the absence of precore mutants, and both precore mutants and other mutations throughout the HBV genome occur commonly even in patients with typical, self-limited, milder forms of HBV infection. In areas where chronic HBV infection is common, precore mutations are more frequent and may reflect viral evolution driven by immune selection. Additional investigation is necessary to define the effect of precore mutants on the pathogenicity and natural history of HBV infection.

The second important category ofHBVmutants consists of *escape mutants*, in which a single amino acid substitution, from glycine to arginine, occurs at position 145 of the immunodominant *a* determinant common to all subtypes ofHBsAg. This change in HBsAg leads to a critical conformational change that results in a loss of neutralizing activity by anti-HBs. This specific HBV/*a* mutant has been observed in two situations, active and passive immunization, in which humoral immunologic pressure may favor evolutionary change ("escape") in the virus -- in a small number of hepatitis B vaccine recipients who acquired HBV infection despite the prior appearance of neutralizing anti-HBs and in liver transplant recipients who underwent the procedure for hepatitis B and who were treated with a high-potency human monoclonal anti-HBs preparation. Although such mutants have not been recognized frequently, their existence raises a concern that may complicate vaccination strategies and serologic diagnosis.

Extrahepatic sites Hepatitis B antigens andHBV DNA have been identified in extrahepatic sites, including lymph nodes, bone marrow, circulating lymphocytes, spleen, and pancreas. Although the virus does not appear to be associated with tissue injury in any of these extrahepatic sites, its presence in these "remote" reservoirs has been invoked to explain the recurrence of HBV infection after orthotopic liver transplantation. A more complete understanding of the clinical relevance of extrahepatic HBV remains to be defined.

Hepatitis D The delta hepatitis agent, orHDV, is a defective RNA virus that coinfects with and requires the helper function ofHBV (or other hepadnaviruses) for its replication and expression. Slightly smaller than HBV, delta is a formalin-sensitive, 35- to 37-nm virus with a hybrid structure. Its nucleocapsid expresses delta antigen, which bears no antigenic homology with any of the HBV antigens, and contains the virus genome. The delta core is "encapsidated" by an outer envelope ofHBsAg, indistinguishable from that of HBV except in its relative compositions of major, middle, and large HBsAg component proteins. The genome is a small, 1700-nucleotide, circular, single-stranded RNA (minus strand) that is nonhomologous with HBV DNA (except for a small area of the polymerase gene) but that has features and the rolling circle model of replication common to genomes of plant satellite viruses or viroids. HDV RNA contains many areas of internal complementarity; therefore, it can fold on itself by internal base pairing to form an unusual, very stable, rodlike structure. HDV RNA replicates via RNA-directed RNA synthesis by transcription of genomic RNA to a complementary antigenomic (plus strand) RNA; the antigenomic RNA, in turn, serves as a template for subsequent genomic RNA synthesis. Between the genomic and antigenomic RNAs of HDV, there are coding regions for nine proteins. Delta antigen, which is a product of the antigenomic strand, exists in two forms, a small, 195-amino-acid species, which plays a role in facilitating HDV RNA replication, and a large, 214-amino-acid species, which appears to suppress replication but is required for assembly of the antigen into virions. Although complete hepatitis D virions and liver injury require the cooperative helper function of HBV, intracellular replication of HDV RNA can occur without HBV. Genomic heterogeneity among HDV isolates has been described; however, pathophysiologic and clinical consequences of this genetic diversity have not been recognized.

HDVcan either infect a person simultaneously withHBV(*coinfection*) or superinfect a person already infected with HBV (*superinfection*); when HDV infection is transmitted from a donor with oneHBsAgsubtype to an HBsAg-positive recipient with a different subtype, the HDV agent assumes the HBsAg subtype of the recipient, rather than the donor. Because HDV relies absolutely on HBV, the duration of HDV infection is determined by the duration of (and cannot outlast) HBV infection. HDV antigen is expressed primarily in hepatocyte nuclei and is occasionally detectable in serum. During acute HDV infection, anti-HDV of the IgM class predominates, and 30 to 40 days may elapse after symptoms appear before anti-HDV can be detected. In self-limited infection, anti-HDV is low titer and transient, rarely remaining detectable beyond the clearance of HBsAg and HDV antigen. In chronic HDV infection, anti-HDV circulates in high titer, and both IgM and IgG anti-HDV can be detected. HDV antigen in the liver and HDV RNA in serum and liver can be detected during HDV replication.

Hepatitis C Hepatitis C virus, which, before its identification was labeled "non-A, non-B hepatitis," is a linear, single-stranded, positive-sense, 9400-nucleotide RNA virus, the genome of which is similar in organization to that of flaviviruses and pestiviruses;HCVconstitutes its own genus in the family Flaviviridae. The HCV genome contains a single large open reading frame (gene) that codes for a virus polyprotein of approximately 3000 amino acids. The 5¢ end of the genome consists of an untranslated region adjacent to the genes for structural proteins, the nucleocapsid core protein and two envelope glycoproteins, E1 and E2/NS1. The 5¢ untranslated region and core gene are highly conserved among genotypes, but the envelope proteins are coded for by the hypervariable region, which varies from isolate to isolate and may allow the virus to evade host immunologic containment directed at accessible virus-envelope proteins. The 3¢ end of the genome contains the genes for nonstructural (NS) proteins. The first reported HCV clone, 5-1-1, and the nucleotide sequence coding for C100-3, the recombinant virus protein used in the first immunoassay for antibodies to HCV, reside within the NS4 gene, and the RNA-dependent RNA polymerase, through which HCV replicates, is encoded by the NS5 region (Fig. 295-6). Because HCV does not replicate via a DNA intermediate, it does not integrate into the host genome. Because HCV tends to circulate in very low titer, visualization of virus particles, estimated to be 40 to 60 nm in diameter, has been difficult. Although in vitro HCV replication remains difficult to accomplish convincingly, the chimpanzee has proven to be an invaluable experimental animal model.

At least six distinct genotypes, as well as subtypes within genotypes, ofHCV have been identified by nucleotide sequencing. Genotypes differ one from another in sequence homology by ³30%. Because divergence of HCV isolates within a genotype or subtype, and within the same host, may vary insufficiently to define a distinct genotype, these intragenotypic differences are referred to as *quasispecies* and differ in sequence homology by only a few percent. The genotypic and quasispecies diversity of HCV, resulting from its high mutation rate, interferes with effective humoral immunity. Neutralizing antibodies to HCV have been demonstrated, but they tend to be short-lived; and HCV infection does not induce lasting immunity against reinfection with different virus isolates or even the same virus isolate. Thus, neither *heterologous* nor *homologous* immunity appears to develop after acute HCV infection. Some HCV genotypes are distributed worldwide, while others are more geographically confined. In addition, differences in pathogenicity and responsiveness to antiviral therapy have been reported among genotypes; however, the biologic impact of genotype and quasispecies differences remains incompletely defined.

As noted above, the first assay detected antibodies to C100-3, a recombinant polypeptide derived from theNS4 region of the genome. In most patients with acute hepatitis C, antibody detected with this assay appears between 1 to 3 months after the onset of acute hepatitis but sometimes not for a year or longer. Second-generation assays incorporate recombinant proteins from the nucleocapsid core region, C22-3, and the NS3 region, C33c (expressed in combination with C100-3 as C200); these assays are more sensitive (by approximately 20%) and detect anti-HCV30 to 90 days earlier, during the period of acute hepatitis. A third-generation immunoassay, which incorporates proteins from the NS5 region and replaces some recombinant proteins with synthetic peptides, may detect anti-HCV even earlier. Because nonspecificity has been encountered in clinical samples tested for anti-HCV, a supplementary recombinant immunoblot assay was introduced. Reactivity in an immunoassay is "confirmed" by incubation with a nitrocellulose strip that contains individual bands of recombinant or synthetic HCV proteins. This approach allows the demonstration of individual antibodies to nonstructural and structural viral proteins and identifies false-positive reactivity associated with nonviral specificities. It is useful to support the validity of anti-HCV-reactive samples, especially in patients with a low prior probability of true infection (e.g., blood donors) or in patients with confounding activity in serum (such as a rheumatoid factor) that may yield false-positive antibody reactivity. Still, detection of anti-HCV is insufficient to identify all persons infected with HCV. The most sensitive indicator is the presence of HCV RNA, which requires molecular amplification by polymerase chain reaction (PCR) (Fig. 295-7). An alternative method for detection of HCV RNA, more easily automated but one or two orders of magnitude less sensitive, is branched-chain complementary DNA hybridization. HCV RNA can be detected within a few days of exposure to HCV, well before the appearance of anti-HCV, and tends to persist for the duration of HCV infection; however, in patients with chronic HCV infection, occasionally, HCV RNA may be detectable only intermittently. Application of sensitive molecular probes for HCV RNA has revealed the presence of replicative HCV in peripheral blood lymphocytes of infected persons; however, as is the case forHBV in lymphocytes, the clinical relevance of HCV lymphocyte infection is not known.

Hepatitis E Previously labeled *epidemic* or *enterically transmitted non-A, non-B hepatitis*,HEV is an enterically transmitted virus that occurs primarily in India, Asia, Africa, and Central America. This agent, with epidemiologic features resembling those of hepatitis A, is a 32- to 34-nm, nonenveloped,HAV-like virus with a 7600-nucleotide, single-stranded, positive-sense RNA genome. HEV has three open reading frames (genes), the largest of which encodes nonstructural proteins involved in virus replication. A middle-sized gene encodes the nucleocapsid protein, and the smallest, whose function is not known, encodes protein specificities to which antibodies appear in human serum. All HEV isolates appear to belong to a single serotype, despite genomic heterogeneity of up to 25%. There is no genomic or antigenic homology, however, between HEV and HAV or other picornaviruses; and HEV, although resembling calciviruses, appears to be sufficiently distinct from any known agent to merit a new classification of its own within the alphavirus group. The virus has been detected in stool, bile, and liver and is excreted in the stool during the late incubation period; immune responses to viral antigens occur very early during the course of acute infection. Both IgM anti-HEV and IgG anti-HEV can be detected, but both fall rapidly after acute infection, reaching low levels within 9 to 12 months. Currently, serologic testing for HEV infection is not available routinely.

PATHOGENESIS

Under ordinary circumstances, none of the hepatitis viruses is known to be directly cytopathic to hepatocytes. Evidence suggests that the clinical manifestations and outcomes after acute liver injury associated with viral hepatitis are determined by the immunologic responses of the host.

Hepatitis B Among the viral hepatitides, the immunopathogenesis of hepatitis B has been studied most extensively. Certainly for this agent, the existence of asymptomatic hepatitis B carriers with normal liver histology and function suggests that the virus is not directly cytopathic. The fact that patients with defects in cellular immune competence are more likely to remain chronically infected rather than to clear the virus is cited to support the role of cellular immune responses in the pathogenesis of hepatitis B-related liver injury. The model that has the most experimental support involves cytolytic T cells sensitized specifically to recognize host and hepatitis B viral antigens on the liver cell surface. Recent laboratory observations suggest that nucleocapsid proteins (HBcAgand possiblyHBeAg), present on the cell membrane in minute quantities, are the viral target antigens that, with host antigens, invite cytolytic T cells to destroyHBV-infected hepatocytes. Differences in the robustness of cytolytic T cell responsiveness and in the elaboration of antiviral cytokines by T cells have been invoked to explain differences in outcomes between those who recover after acute hepatitis and those who progress to chronic hepatitis or between those with mild and those with severe (fulminant) acute HBV infection.

A recent observation provides further insight into the mechanism of viral clearance in acute hepatitis B. Although a robust cytolytic T cell response occurs and eliminates virus-infected liver cells during acute hepatitis B, more than 90% ofHBV DNA has been found in experimentally infected chimpanzees to disappear from the liver and blood before maximal T cell infiltration of the liver and before most of the biochemical and histologic evidence of liver injury. This observation suggests that inflammatory cytokines, independent of cytopathic antiviral mechanisms, participate in early viral clearance; this effect has been shown to represent elimination of HBV replicative intermediates from the cytoplasm and covalently closed circular viral DNA from the nucleus of infected hepatocytes.

Debate continues over the relative importance of viral and host factors in the pathogenesis ofHBV-associated liver injury and its outcome. As noted above, precore genetic mutants of HBV have been associated with the more severe outcomes of HBV infection (severe chronic and fulminant hepatitis), suggesting that, under certain circumstances, relative pathogenicity is a property of the virus, not the host. The fact that concomitantHDV and HBV infections are associated with more severe liver injury than HBV infection alone and the fact that cells transfected in vitro with the gene for HDV (delta) antigen express HDV antigen and then become necrotic in the absence of any immunologic influences are also consistent with a viral effect on pathogenicity. Similarly, in patients who undergo liver transplantation for end-stage chronic hepatitis B, occasionally, rapidly progressive liver injury appears in the new liver. This clinical pattern is associated with an unusual histologic pattern in the new liver, *fibrosing cholestatic hepatitis*, which, ultrastructurally, appears to represent a choking of the cell with overwhelming quantities ofHBsAg. This observation suggests that under the influence of the potent immunosuppressive agents required to prevent allograft rejection, HBV may have a direct cytopathic effect on liver cells, independent of the immune system.

Although the precise mechanism of liver injury inHBVinfection remains elusive, studies of nucleocapsid proteins have shed light on the profound immunologic tolerance to HBV of babies born to mothers with highly replicative (HBeAg-positive), chronic HBV infection. In HBeAg-expressing transgenic mice, in utero exposure to HBeAg, which is sufficiently small to traverse the placenta, induces T cell tolerance to both nucleocapsid proteins. This, in turn, may explain why, when infection occurs so early in life, immunologic clearance does not occur, and protracted, lifelong infection ensues.

Hepatitis C Undoubtedly, cell-mediated immune responses and elaboration by T cells of antiviral cytokines contribute to the containment of infection and pathogenesis of liver injury associated with hepatitis C. PerhapsHCVinfection of lymphoid cells plays a role in moderating immune responsiveness to the virus, as well. Intrahepatic HLA class-I-restricted cytolytic T cells directed at nucleocapsid, envelope, andNS viral protein antigens have been demonstrated in patients with chronic hepatitis C. Such virus-specific cytolytic T cell responses, however, do not correlate adequately with the degree of liver injury or with recovery. Several HLA alleles have been linked with self-limited hepatitis C, but such associations do not apply universally. Finally, cross-reactivity between viral and host autoantigens has been invoked to explain the association between hepatitis C and a subset of patients with autoimmune hepatitis and antibodies to liver kidney microsomal antigen (anti-LKM) (Chap. 297).

Extrahepatic Manifestations Immune complex-mediated tissue damage appears to play a pathogenetic role in the extrahepatic manifestations of acute hepatitis B. The occasional prodromal serum sickness-like syndrome observed in acute hepatitis B appears to be related to the deposition in tissue blood vessel walls of circulating immune complexes leading to activation of the complement system. The clinical consequences are urticarial rash, angioedema, fever, and arthritis. During the early prodrome ofHBVinfection in these patients,HBsAg in high titer in association with small amounts of anti-HBs leads to the formation of soluble, circulating immune complexes (in antigen excess). Complement components in the serum are depressed during the arthritic phase of the illness and are also detectable in the circulating immune complexes, which also contain HBsAg, anti-HBs, IgG, IgM, IgA, and fibrin.

In patients with chronic hepatitis B, other types of immune-complex disease may be seen. Glomerulonephritis with the nephrotic syndrome is occasionally observed;HBsAg, immunoglobulin, and C3 deposition has been found in the glomerular basement membrane. While polyarteritis nodosa develops in considerably fewer than 1% of patients withHBVinfection, 20 to 30% of patients with polyarteritis nodosa have HBsAg in serum (Chap. 317). In these patients, the affected small and medium-sized arterioles have been shown to contain HBsAg, immunoglobulins, and complement components. Another extrahepatic manifestation of viral hepatitis, essential mixed cryoglobulinemia (EMC), was reported initially to be associated with hepatitis B. The disorder is characterized clinically by arthritis and cutaneous vasculitis (palpable purpura) and serologically by the presence of circulating cryoprecipitable immune complexes of more than one immunoglobulin class (Chap. 275). Many patients with this syndrome have chronic liver disease, but the association with HBV infection is limited; instead, a substantial proportion have chronicHCVinfection. Their circulating immune complexes contain HCV RNA at a concentration that exceeds its serum concentration, favoring a primary role for the virus in the pathogenesis of EMC.

PATHOLOGY

The typical morphologic lesions of all types of viral hepatitis are similar and consist of panlobular infiltration with mononuclear cells, hepatic cell necrosis, hyperplasia of Kupffer cells, and variable degrees of cholestasis. Hepatic cell regeneration is present, as evidenced by numerous mitotic figures, multinucleated cells, and "rosette" or "pseudoacinar" formation. The mononuclear infiltration consists primarily of small lymphocytes, although plasma cells and eosinophils occasionally are present. Liver cell damage consists of hepatic cell degeneration and necrosis, cell dropout, ballooning of cells, and acidophilic degeneration of hepatocytes (forming so-called Councilman bodies). Large hepatocytes with a ground glass appearance of the cytoplasm may be seen in chronic but not in acuteHBVinfection; these cells containHBsAgand can be identified histochemically with orcein or aldehyde fuchsin. In uncomplicated viral hepatitis, the reticulin framework is preserved.

In hepatitis C, the histologic lesion is often remarkable for a relative paucity of inflammation, a marked increase in activation of sinusoidal lining cells, lymphoid aggregates, the presence of fat, and, occasionally, bile duct lesions in which biliary epithelial cells appear to be piled up without interruption of the basement membrane. Occasionally, microvesicular steatosis occurs in hepatitis D. In hepatitis E, a common histologic feature is marked cholestasis. A cholestatic variant of slowly resolving acute hepatitis A also has been described.

A more severe histologic lesion, *bridging hepatic necrosis*, also termed *subacute* or *confluent necrosis*, is occasionally observed in some patients with acute hepatitis. "Bridging" between lobules results from large areas of hepatic cell dropout, with collapse of the reticulin framework. Characteristically, the bridge consists of condensed reticulum, inflammatory debris, and degenerating liver cells that span adjacent portal areas, portal to central veins, or central vein to central vein. This lesion had been thought to have prognostic significance; in many of the originally described patients with this lesion, a subacute course terminated in death within several weeks to months, or severe chronic hepatitis and postnecrotic cirrhosis developed. More recent investigations have failed to uphold the association between bridging necrosis and such a poor prognosis in patients with acute hepatitis. Although the frequency of bridging may be higher among hospitalized patients with severe acute hepatitis, and although cirrhosis, chronic hepatitis, and even death have occurred in this group, the frequency of bridging necrosis in uncomplicated acute viral hepatitis is probably on the order of 1 to 5%. Prospective studies have failed to demonstrate a difference in prognosis between patients with acute hepatitis who have bridging necrosis and those who do not. Therefore, although demonstration of this lesion in patients with chronic hepatitis has prognostic significance (Chap. 297), its demonstration during acute hepatitis is less meaningful, and liver biopsies to identify this lesion are no longer undertaken routinely in patients with acute hepatitis. In *massive hepatic necrosis* (fulminant hepatitis, acute yellow atrophy), the striking feature at postmortem examination is the finding of a small, shrunken, soft liver. Histologic examination reveals massive necrosis and dropout of liver cells of most lobules with extensive collapse and condensation of the reticulin framework.

Immunofluorescence and immunoperoxidase antibody studies have localizedHBsAg to the cytoplasm and plasma membrane of infected liver cells. In contrast,HBcAgpredominates in the nucleus, but occasionally, scant amounts are also seen in the cytoplasm and on the cell membrane. Electron-microscopic studies of liver biopsy material have demonstrated the presence of HBsAg particles in the cytoplasm and HBcAg particles in the nucleus of liver cells during hepatitis B infection. These morphologic observations suggest that DNA is synthesized and packaged within core particles in the nucleus, while the envelope is assembled in the cytoplasm, resulting in the formation of intact hepatitis B virus.HDVantigen is localized to the hepatocyte nucleus, whileHAV,HCV, andHEVantigens are localized to the cytoplasm.

EPIDEMIOLOGY

Before the availability of serologic tests for hepatitis viruses, all viral hepatitis cases were labeled either as "infectious" or "serum" hepatitis. Modes of transmission overlap, however, and *a clear distinction among the different types of viral hepatitis cannot be made solely on the basis of clinical or epidemiologic features* (Table 295-2). The most accurate means to distinguish the various types of viral hepatitis involves specific serologic testing.

Hepatitis A *This agent is transmitted almost exclusively by the fecal-oral route*.

Person-to-person spread ofHAV is enhanced by poor personal hygiene and overcrowding; large outbreaks as well as sporadic cases have been traced to contaminated food, water, milk, frozen raspberries and strawberries, and shellfish. Intrafamily and intrainstitutional spread are also common. Early epidemiologic observations suggested that there is a predilection for hepatitis A to occur in late fall and early winter. In temperate zones, epidemic waves have been recorded every 5 to 20 years as new segments of nonimmune population appeared; however, in developed countries, the incidence of type A hepatitis has been declining, presumably as a function of improved sanitation, and these cyclic patterns are no longer being observed. No HAV carrier state has been identified after acute type A hepatitis; perpetuation of the virus in nature depends presumably on nonepidemic, inapparent subclinical infection.

In the general population, anti-HAV, an excellent marker for previous HAV infection, increases in prevalence as a function of increasing age and of decreasing socioeconomic status. In the 1970s, serologic evidence of prior hepatitis A infection occurred in about 40% of urban populations in the United States, most of whose members never recalled having had a symptomatic case of hepatitis. In subsequent decades, however, the prevalence of anti-HAV has been declining in the United States. In developing countries, exposure, infection, and subsequent immunity are almost universal in childhood. As the frequency of subclinical childhood infections declines in developed countries, a susceptible cohort of adults emerges. Hepatitis A tends to be more symptomatic in adults; therefore, paradoxically, as the frequency of HAV infection declines, the likelihood of clinically apparent, even severe, HAV illnesses increases in the susceptible adult population. Travel to endemic areas is a common source of infection for adults from nonendemic areas. More recently recognized epidemiologic foci of HAV infection include child-care centers, neonatal intensive care units, promiscuous homosexual men, and injection drug users. Although hepatitis A is rarely bloodborne, several outbreaks have been recognized in recipients of clotting factor concentrates.

Hepatitis B Percutaneous inoculation has long been recognized as a major route of hepatitis B transmission, but the outmoded designation "serum hepatitis" is an inaccurate label for the epidemiologic spectrum ofHBVinfection recognized today. As detailed below, most of the hepatitis transmitted by blood transfusion is not caused by HBV; moreover, in approximately two-thirds of patients with acute type B hepatitis, there is no history of an identifiable percutaneous exposure. We now recognize that many cases of type B hepatitis result from less obvious modes of nonpercutaneous or covert percutaneous transmission.HBsAghas been identified in almost every body fluid from infected persons, and at least some of these body fluids -- most notably semen and saliva -- are infectious, albeit less so than serum, when administered percutaneously or nonpercutaneously to experimental animals. Among the nonpercutaneous modes of HBV transmission, oral ingestion has been documented as a potential but inefficient route of exposure. By contrast, the two nonpercutaneous routes considered to have the greatest impact are intimate (especially sexual) contact and perinatal transmission.

In sub-Saharan Africa, intimate contact among toddlers is considered instrumental in contributing to the maintenance of the high frequency of hepatitis B in the population. Perinatal transmission occurs primarily in infants born toHBsAgcarrier mothers or mothers with acute hepatitis B during the third trimester of pregnancy or during the early postpartum period. Perinatal transmission is uncommon in North America and western

Europe but occurs with great frequency and is the most important mode ofHBVperpetuation in the Far East and developing countries. Although the precise mode of perinatal transmission is unknown, and although approximately 10% of infections may be acquired in utero, epidemiologic evidence suggests that most infections occur approximately at the time of delivery and are not related to breast feeding. The likelihood of perinatal transmission of HBV correlates with the presence ofHBeAg; 90% of HBeAg-positive mothers but only 10 to 15% of anti-HBe-positive mothers transmit HBV infection to their offspring. In most cases, acute infection in the neonate is clinically asymptomatic, but the child is very likely to become anHBsAgcarrier.

The more than 350 millionHBsAgcarriers in the world constitute the main reservoir of hepatitis B in human beings. Serum HBsAg is infrequent (0.1 to 0.5%) in normal populations in the United States and western Europe. However, a prevalence of up to 5 to 20% has been found in the Far East and in some tropical countries; in persons with Down's syndrome, lepromatous leprosy, leukemia, Hodgkin's disease, polyarteritis nodosa; in patients with chronic renal disease on hemodialysis; and in injection drug users.

Other groups with high rates ofHBVinfection include spouses of acutely infected persons, sexually promiscuous persons (especially promiscuous homosexual men), health care workers exposed to blood, persons who require repeated transfusions especially with pooled blood product concentrates (e.g., hemophiliacs), residents and staff of custodial institutions for the mentally retarded, prisoners, and, to a lesser extent, family members of chronically infected patients. In volunteer blood donors, the prevalence of anti-HBs, a reflection of previous HBV infection, ranges from 5 to 10%, but the prevalence is higher in lower socioeconomic strata, older age groups, and persons -- including those mentioned above -- exposed to blood products.

Prevalence of infection, modes of transmission, and human behavior conspire to mold geographically different epidemiologic patterns ofHBVinfection. In the Far East and Africa, hepatitis B, a disease of the newborn and young children, is perpetuated by a cycle of maternal-neonatal spread. In North America and western Europe, hepatitis B is primarily a disease of adolescence and early adulthood, the time of life when intimate sexual contact as well as recreational and occupational percutaneous exposures tend to occur.

Hepatitis D Infection withHDV has a worldwide distribution, but two epidemiologic patterns exist. In Mediterranean countries (northern Africa, southern Europe, the Middle East), HDV infection is endemic among those with hepatitis B, and the disease is transmitted predominantly by nonpercutaneous means, especially close personal contact. In nonendemic areas, such as the United States and northern Europe, HDV infection is confined to persons exposed frequently to blood and blood products, primarily injection drug users and hemophiliacs. HDV infection can be introduced into a population through drug users or by migration of persons from endemic to nonendemic areas. Thus, patterns of population migration and human behavior facilitating percutaneous contact play important roles in the introduction and amplification of HDV infection. Occasionally, the migrating epidemiology of hepatitis D is expressed in explosive outbreaks of severe hepatitis, such as those that have occurred in remote

South American villages as well as in urban centers in the United States. Ultimately, such outbreaks of hepatitis D -- either of coinfections with acute hepatitis B or of superinfections in those already infected with HBV -- may blur the distinctions between endemic and nonendemic areas.

Hepatitis C Routine screening of blood donors forHBsAgand the elimination of commercial blood sources in the early 1970s reduced the frequency of, but did not eliminate, transfusion-associated hepatitis. During the 1970s, the likelihood of acquiring hepatitis after transfusion of voluntarily donated, HBsAg-screened blood was approximately 10% per patient (up to 0.9% per unit transfused); 90 to 95% of these cases were classified, based on serologic exclusion of hepatitis A and B, as "non-A, non-B" hepatitis. For patients requiring transfusion of pooled products, such as clotting factor concentrates, the risk was even higher, up to 20 to 30%, while for those receiving such products as albumin and immune globulin, because of prior treatment of these materials by heating to 60°C or cold ethanol fractionation, there was no risk of hepatitis.

During the 1980s, voluntary self-exclusion of blood donors with risk factors for AIDS and then the introduction of donor screening for anti-HIV reduced further the likelihood of transfusion-associated hepatitis to under 5%. During the late 1980s and early 1990s, the introduction first of "surrogate" screening tests for non-A, non-B hepatitis [alanine aminotransferase (ALT) and anti-HBc, both shown to identify blood donors with a higher likelihood of transmitting non-A, non-B hepatitis to recipients] and, subsequently, after the discovery ofHCV, first-generation immunoassays for anti-HCV reduced the frequency of transfusion-associated hepatitis even further. A prospective analysis of transfusion-associated hepatitis conducted between 1986 and 1990 showed that the incidence of transfusion-associated hepatitis at one urban university hospital fell from a baseline of 3.8% per patient (0.45% per unit transfused) to 1.5% per patient (0.19% per unit) after the introduction of surrogate testing and to 0.6% per patient (0.03% per unit) after the introduction of first-generation anti-HCV assays. The introduction of second-generation anti-HCV assays has reduced the frequency of transfusion-associated hepatitis C to almost imperceptible levels, 1 in 100,000.

In addition to being transmitted by transfusion, hepatitis C can be transmitted by other percutaneous routes, such as self-injection with intravenous drugs. In addition, this virus can be transmitted by occupational exposure to blood, and the likelihood of infection is increased in hemodialysis units. Although the frequency of transfusion-associated hepatitis C fell as a result of blood donor screening, the overall frequency of hepatitis C remained the same until the early 1990s, when the overall frequency fell by 80%, in parallel with a reduction in the number of new cases in injection drug users. After the exclusion of anti-HCV-positive plasma units from the donor pool, rare, sporadic instances have occurred of hepatitis C among recipients of immune globulin preparations for intravenous (but not intramuscular) use.

Serologic evidence for HCVinfection occurs in 90% of patients with a history of transfusion-associated hepatitis (almost all occurring before 1992, when second-generation HCV-screening tests were introduced), hemophiliacs and others treated with clotting factors, and injection-drug users; 60 to 70% of patients with sporadic "non-A, non-B" hepatitis who lack identifiable risk factors; 0.5% of volunteer blood donors; and 1.8% of the general population in the United States, which translates into 4 million persons. Comparable frequencies of HCV infection occur in most countries around the world, but extraordinarily high prevalences of HCV infection occur in certain countries, such as Egypt, where more than 20% of the population in some cities is infected. In the United States, African Americans and Mexican Americans have higher frequencies of HCV infection than whites, and 30- to 49-year-old adult males have the highest frequencies of infection. Chronic hepatitis C accounts for 20% of sporadic acute hepatitis and 40% of chronic liver disease, is the most frequent indication for liver transplantation, and is estimated to account for 8000 to 10,000 deaths per year in the United States.

Most asymptomatic blood donors found to have anti-HCVand approximately 40% of persons with reported cases of acute hepatitis C do not fall into a recognized risk group; however, many such blood donors do recall risk-associated behaviors when questioned carefully, and most patients with acute hepatitis C in the absence of clear-cut risk factors tend to be of lower socioeconomic backgrounds. Thorough questioning of anti-HCV-reactive blood donors has identified nasal cocaine inhalation, with shared equipment, as a potential risk factor for acquiring HCV infection.

As a bloodborne infection,HCVpotentially can be transmitted sexually and perinatally; however, both of these modes of transmission are inefficient for hepatitis C. Although 10 to 15% of patients with acute hepatitis C report having potential sexual sources of infection, most studies have failed to identify sexual transmission of this agent. The chances of sexual and perinatal transmission have been estimated to be approximately 5%, well below comparable rates for HIV andHBVinfections. Moreover, sexual transmission appears to be confined to such subgroups as persons with multiple sexual partners and sexually transmitted diseases; transmission of HCV infection is rare between stable, monogamous sexual partners. Breast feeding does not increase the risk of HCV infection between an infected mother and her infant. Infection of health workers is not dramatically higher than among the general population; however, health workers are more likely to acquire HCV infection through accidental needle punctures, the efficiency of which ranges between 3 and 10%. Infection of household contacts is rare as well.

Other groups with an increased frequency ofHCVinfection include patients who require hemodialysis and organ transplantation and those who require transfusions in the setting of cancer chemotherapy. In immunosuppressed individuals, levels of anti-HCV may be undetectable, and a diagnosis may require testing for HCV RNA. Although new acute cases of hepatitis C are rare, newly diagnosed cases are common among otherwise healthy persons who experimented briefly with injection drugs two or three decades earlier. Such instances usually remain unrecognized for years, until unearthed by laboratory screening for routine medical examinations, insurance applications, and attempted blood donation.

Hepatitis E The enteric form of non-A, non-B hepatitis identified in India, Asia, Africa, and Central America resembles hepatitis A in its primarily enteric mode of spread. The commonly recognized cases occur after contamination of water supplies such as after monsoon flooding, but sporadic, isolated cases occur. An epidemiologic feature that distinguishesHEV from other enteric agents is the rarity of secondary person-to-person spread from infected persons to their close contacts. Infections arise in populations that are immune toHAV and favor young adults. It is not known if hepatitis E occurs outside of recognized endemic areas, for example, in the United States, but preliminary studies suggest that HEV does not account for any of the sporadic "non-A, non-B" cases in nonendemic areas. Cases imported from endemic areas have been found in the United States.

CLINICAL AND LABORATORY FEATURES

Symptoms and Signs Acute viral hepatitis occurs after an incubation period that varies according to the responsible agent. Generally, incubation periods for hepatitis A range from 15 to 45 days (mean 4 weeks), for hepatitis B and D from 30 to 180 days (mean 4 to 12 weeks), for hepatitis C from 15 to 160 days (mean 7 weeks), and for hepatitis E from 14 to 60 days (mean 5 to 6 weeks). The *prodromal symptoms* of acute viral hepatitis are systemic and quite variable. Constitutional symptoms of anorexia, nausea and vomiting, fatigue, malaise, arthralgias, myalgias, headache, photophobia, pharyngitis, cough, and coryza may precede the onset of jaundice by 1 to 2 weeks. The nausea, vomiting, and anorexia are frequently associated with alterations in olfaction and taste. A low-grade fever between 38 and 39°C (100 to 102°F) is more often present in hepatitis A and E than in hepatitis B or C, except when hepatitis B is heralded by a serum sicknesslike syndrome; rarely, a fever of 39.5 to 40°C (103 to 104°F) may accompany the constitutional symptoms. Dark urine and clay-colored stools may be noticed by the patient from 1 to 5 days before the onset of clinical jaundice.

With the onset of *clinical jaundice*, the constitutional prodromal symptoms usually diminish, but in some patients mild weight loss (2.5 to 5 kg) is common and may continue during the entire icteric phase. The liver becomes enlarged and tender and may be associated with right upper quadrant pain and discomfort. Infrequently, patients present with a cholestatic picture, suggesting extrahepatic biliary obstruction. Splenomegaly and cervical adenopathy are present in 10 to 20% of patients with acute hepatitis. Rarely, a few spider angiomas appear during the icteric phase and disappear during convalescence. During the *recovery phase*, constitutional symptoms disappear, but usually some liver enlargement and abnormalities in liver biochemical tests are still evident. The duration of the posticteric phase is variable, ranging from 2 to 12 weeks, and usually is more prolonged in acute hepatitis B and C. Complete clinical and biochemical recovery is to be expected 1 to 2 months after all cases of hepatitis A and E and 3 to 4 months after the onset of jaundice in three-quarters of uncomplicated cases of hepatitis B and C. In the remainder, biochemical recovery may be delayed. A substantial proportion of patients with viral hepatitis never become icteric.

Infection withHDV can occur in the presence of acute or chronicHBVinfection; the duration of HBV infection determines the duration of HDV infection. When acute HDV and HBV infection occur simultaneously, clinical and biochemical features may be indistinguishable from those of HBV infection alone, although occasionally they are more severe. As opposed to patients with *acute* HBV infection, patients with *chronic* HBV infection can support HDV replication indefinitely. This can happen when acute HDV infection occurs in the presence of a nonresolving acute HBV infection. More commonly, acute HDV infection becomes chronic when it is superimposed on an underlying chronic HBV infection. In such cases, the HDV superinfection appears as a clinical exacerbation or an episode resembling acute viral hepatitis in someone already chronically infected with HBV. Superinfection with HDV in a patient with chronic hepatitis B often leads to clinical deterioration (see below).

In addition to superinfections with other hepatitis agents, acute hepatitis-like clinical events in persons with chronic hepatitis B may accompany spontaneousHBeAg-to-anti-HBe seroconversion or spontaneous reactivation, i.e., reversion from nonreplicative to replicative infection. Such reactivations can occur as well in therapeutically immunosuppressed patients with chronicHBVinfection when cytotoxic-immunosuppressive drugs are withdrawn; in these cases, restoration of immune competence is thought to allow resumption of previously checked cell-mediated cytolysis of HBV-infected hepatocytes. Occasionally, acute clinical exacerbations of chronic hepatitis B may represent the emergence of a precore mutant (see "Virology and Etiology," above).

Laboratory Features The serum aminotransferases aspartate aminotransferase (AST) andALT(previously designated SGOT and SGPT) show a variable increase during the prodromal phase of acute viral hepatitis and precede the rise in bilirubin level (Figs. 295-2 and295-4). The acute level of these enzymes, however, does not correlate well with the degree of liver cell damage. Peak levels vary from 400 to 4000 IU or more; these levels are usually reached at the time the patient is clinically icteric and diminish progressively during the recovery phase of acute hepatitis. The diagnosis of anicteric hepatitis is difficult and requires a high index of suspicion; it is based on clinical features and on aminotransferase elevations, although mild increases in conjugated bilirubin also may be found.

Jaundice is usually visible in the sclera or skin when the serum bilirubin value exceeds 43 umol/L (2.5 mg/dL). When jaundice appears, the serum bilirubin typically rises to levels ranging from 85 to 340 umol/L (5 to 20 mg/dL). The serum bilirubin may continue to rise despite falling serum aminotransferase levels. In most instances, the total bilirubin is equally divided between the conjugated and unconjugated fractions. Bilirubin levels above 340 umol/L (20 mg/dL) extending and persisting late into the course of viral hepatitis are more likely to be associated with severe disease. In certain patients with underlying hemolytic anemia, however, such as glucose-6-phosphate dehydrogenase deficiency and sickle cell anemia, a high serum bilirubin level is common, resulting from superimposed hemolysis. In such patients, bilirubin levels greater than 513 umol/L (30 mg/dL) have been observed and are not necessarily associated with a poor prognosis.

Neutropenia and lymphopenia are transient and are followed by a relative lymphocytosis. Atypical lymphocytes (varying between 2 and 20%) are common during the acute phase. These atypical lymphocytes are indistinguishable from those seen in infectious mononucleosis. Measurement of the prothrombin time (PT) is important in patients with acute viral hepatitis, for a prolonged value may reflect a severe synthetic defect, signify extensive hepatocellular necrosis, and indicate a worse prognosis. Occasionally, a prolonged PT may occur with only mild increases in the serum bilirubin and aminotransferase levels. Prolonged nausea and vomiting, inadequate carbohydrate intake, and poor hepatic glycogen reserves may contribute to hypoglycemia noted occasionally in patients with severe viral hepatitis. Serum alkaline phosphatase may be normal or only mildly elevated, while a fall in serum albumin is uncommon in uncomplicated acute viral hepatitis. In some patients, mild and transient steatorrhea has been noted as well as slight microscopic hematuria and minimal proteinuria.

A diffuse but mild elevation of the gamma globulin fraction is common during acute viral hepatitis. Serum IgG and IgM levels are elevated in about one-third of patients during the acute phase of viral hepatitis, but the serum IgM level is elevated more characteristically during acute hepatitis A. During the acute phase of viral hepatitis, antibodies to smooth muscle and other cell constituents may be present, and low titers of rheumatoid factor, nuclear antibody, and heterophil antibody also can be found occasionally. In hepatitis C and D, antibodies to liver-kidney microsomes (LKM) may occur; however, the species of LKM antibodies in the two types of hepatitis are different from each other as well as from the LKM antibody species characteristic of autoimmune chronic hepatitis type 2 (Chap. 297). The autoantibodies in viral hepatitis are nonspecific and also can be associated with other viral and systemic diseases. In contrast, virus-specific antibodies, which appear during and after hepatitis virus infection, are serologic markers of diagnostic importance.

As described above, serologic tests are available with which to establish a diagnosis of hepatitis A, B, D, and C. Tests for fecal or serumHAV are not routinely available. Therefore, a diagnosis of type A hepatitis is based on detection of IgM anti-HAV during acute illness (Fig. 295-2). Rheumatoid factor can give rise to false-positive results in this test.

A diagnosis ofHBVinfection can usually be made by detection ofHBsAg in serum. Infrequently, levels of HBsAg are too low to be detected during acute HBV infection, even with the current generation of highly sensitive immunoassays. In such cases, the diagnosis can be established by the presence of IgM anti-HBc.

The titer ofHBsAgbears little relation to the severity of clinical disease. Indeed, there may be an inverse correlation between the serum concentration of HBsAg and the degree of liver cell damage. For example, titers are highest in immunosuppressed patients, lower in patients with chronic liver disease (but higher in mild chronic than in severe chronic hepatitis), and very low in patients with acute fulminant hepatitis. These observations suggest that in hepatitis B the degree of liver cell damage and the clinical course are probably related to variations in the patient's immune response toHBVrather than to the amount of circulatingHBsAg. In immunocompetent persons, however, there is a correlation between markers of HBV *replication* and liver injury (see below).

Another serologic marker that may be of value in patients with hepatitis B isHBeAg. Its principal clinical usefulness is as an indicator of relative infectivity. Because HBeAg is invariably present during early acute hepatitis B, HBeAg testing is indicated primarily during follow-up of chronic infection.

In patients with hepatitis B surface antigenemia of unknown duration, e.g., blood donors found to beHBsAg-positive and referred to a physician for evaluation, testing for IgM anti-HBc may be useful to distinguish between acute or recent infection (IgM anti-HBc-positive) and chronicHBVinfection (IgM anti-HBc-negative, IgG anti-HBc-positive). A false-positive test for IgM anti-HBc may be encountered in patients with high-titer rheumatoid factor.

Anti-HBs is rarely detectable in the presence ofHBsAg in patients with *acute* hepatitis B, but 10 to 20% of persons with *chronic*HBVinfection may harbor low-level anti-HBs. This antibody is directed not against the common group determinant, *a*, but against the heterotypic subtype determinant (e.g., HBsAg of subtype *ad* with anti-HBs of subtype *y*). In most cases, this serologic pattern cannot be attributed to infection with two different HBV subtypes, and the presence of this antibody is not a harbinger of imminent HBsAg clearance. When such antibody is detected, its presence is of no recognized clinical significance (see "Virology and Etiology," above).

After immunization with hepatitis B vaccine, which consists ofHBsAgalone, anti-HBs is the only serologic marker to appear. The commonly encountered serologic patterns of hepatitis B and their interpretations are summarized inTable 295-3. Tests for the detection ofHBV DNA in liver and serum are now available. LikeHBeAg, serum HBV DNA is an indicator of HBV replication, but tests for HBV DNA are more sensitive and quantitative. Hybridization assays for HBV DNA have a sensitivity of approximately 10^5 to 10^6virions/mL, a relative threshold below which infectivity and liver injury are limited and HBeAg is usually undetectable. Currently, testing for HBV DNA has shifted from insensitive hybridization assays to amplification assays, e.g., the polymerase chain reaction-based assay, which can detect as few as 100 or 1000 virions/mL. With increased sensitivity, amplification assays remain reactive well below the threshold for infectivity and liver injury. These markers are useful in following the course of HBV replication in patients with chronic hepatitis B receiving antiviral chemotherapy, e.g., with interferon or lamivudine (Chap. 297). In immunocompetent persons, a general correlation does appear to exist between the level of HBV replication, as reflected by the level of HBV DNA in serum, and the degree of liver injury. High serum HBV DNA levels, increased expression of viral antigens, and necroinflammatory activity in the liver go hand in hand unless immunosuppression interferes with cytolytic T cell responses to virus-infected cells; reduction of HBV replication with antiviral drugs tends to be accompanied by an improvement in liver histology.

In patients with hepatitis C, an episodic pattern of aminotransferase elevation is common. A specific serologic diagnosis of hepatitis C can be made by demonstrating the presence in serum of anti-HCV. When a second- or third-generation immunoassay (that detects antibodies to nonstructural and nucleocapsid proteins) is used, anti-HCV can be detected in acute hepatitis C during the initial phase of elevated aminotransferase activity. This antibody may never become detectable in 5 to 10% of patients with acute hepatitis C, and levels of anti-HCV may become undetectable after recovery from acute hepatitis C. In patients with chronic hepatitis C, anti-HCV is detectable in >95% of cases. Nonspecificity can confound immunoassays for anti-HCV, especially in persons with a low prior probability of infection, such as volunteer blood donors, or in persons with circulating rheumatoid factor, which can bind nonspecifically to assay reagents. A supplementary recombinant immunoblot assay (RIBA), in which serum is incubated with a nitrocellulose strip containing viral protein bands, can be used to establish the specific viral proteins to which anti-HCV is directed (see "Virology and Etiology," above). Such RIBA determinations are used routinely to confirm anti-HCV reactivity in blood donors, but determinations of HCV RNA have supplanted RIBA in many clinical settings. Assays for HCV RNA are the most sensitive tests for HCV infection and represent the "gold standard" in establishing a diagnosis of hepatitis C. HCV RNA can be detected even before acute elevation of aminotransferase activity and before the appearance of anti-HCV in patients with acute hepatitis C. In addition, HCV RNA remains detectable indefinitely, continuously in most but intermittently in some, in patients with chronic hepatitis C (even detectable in some persons with normal liver tests, i.e., asymptomatic carriers). In the small minority of patients with hepatitis C who lack anti-HCV, a diagnosis can be supported by detection of HCV RNA. If all these tests are negative and the patient has a well-characterized case of hepatitis after percutaneous exposure to blood or blood products, a diagnosis of hepatitis caused by another agent, as yet unidentified, can be entertained.

Amplification techniques are required to detectHCV RNA, and two are available. One is a branched-chain complementary DNA (bDNA) assay, in which the detection signal (a colorimetrically detectable enzyme bound to a complementary DNA probe) is amplified. The other is aPCRassay, in which the viral RNA is reverse transcribed to complementary DNA and then amplified by repeated cycles of DNA synthesis and polymerization. Both can be used as quantitative assays and a measurement of relative "viral load"; PCR, with a sensitivity of 10_2 to 10_3virions per milliliter is more sensitive than bDNA, with a sensitivity of 2´10_5. Determination of viral load is not a reliable marker of disease severity or prognosis but is helpful in predicting relative responsiveness to antiviral therapy. The same is true for determinations of HCV genotype (Chap. 297).

A proportion of patients with hepatitis C have isolated anti-HBc in their blood, a reflection of a common risk in certain populations to multiple bloodborne hepatitis agents. The anti-HBc in such cases is almost invariably of the IgG class and usually representsHBVinfection in the remote past, rarely current HBV infection with low-level virus carriage.

The presence ofHDVinfection can be identified by demonstrating intrahepatic HDV antigen or, more practically, an anti-HDV seroconversion (a rise in titer of anti-HDV or de novo appearance of anti-HDV). Circulating HDV antigen, also diagnostic of acute infection, is detectable only briefly, if at all. Because anti-HDV is often undetectable onceHBsAgdisappears, retrospective serodiagnosis of acute self-limited, simultaneousHBV and HDV infection is difficult. Early diagnosis of acute infection may be hampered by a delay of up to 30 to 40 days in the appearance of anti-HDV.

When a patient presents with acute hepatitis and hasHBsAgand anti-HDVin serum, determination of the class of anti-HBc is helpful in establishing the relationship between infection withHBV and HDV. Although IgM anti-HBc does not distinguish *absolutely* between acute and chronic HBV infection, its presence is a reliable indicator of recent infection and its absence a reliable indicator of infection in the remote past. In simultaneous acute HBV and HDV infections, IgM anti-HBc will be detectable, while in acute HDV infection superimposed on chronic HBV infection, anti-HBc will be of the IgG class.

Tests for the presence ofHDV RNA are useful for determining the presence of ongoing HDV replication and relative infectivity. Currently, probes for this marker are restricted to a limited number of research laboratories. Similarly, diagnostic tests for hepatitis E are confined to a small number of research laboratories.

Liver biopsy is rarely necessary or indicated in acute viral hepatitis, except when there is a question about the diagnosis or when there is clinical evidence suggesting a diagnosis of chronic hepatitis.

A diagnostic algorithm can be applied in the evaluation of cases of acute viral hepatitis. A patient with acute hepatitis should undergo four serologic tests,HBsAg, IgM anti-HAV, IgM anti-HBc, and anti-HCV(Table 295-4). The presence of HBsAg, with or without IgM anti-HBc, represents HBV infection. If IgM anti-HBc is present, the HBV infection is considered acute; if IgM anti-HBc is absent, theHBVinfection is considered chronic. A diagnosis of acute hepatitis B can be made in the absence of HBsAg when IgM anti-HBc is detectable. A diagnosis of acute hepatitis A is based on the presence of IgM anti-HAV. If IgM anti-HAV coexists with HBsAg, a diagnosis of simultaneous HAV and HBV infections can be made; if IgM anti-HBc (with or without HBsAg) is detectable, the patient has simultaneous acute hepatitis A and B, and if IgM anti-HBc is undetectable, the patient has acute hepatitis A superimposed on chronic HBV infection. The presence of anti-HCV, if confirmable, supports a diagnosis of acute hepatitis C. Occasionally, testing for HCV RNA or repeat anti-HCV testing later during the illness is necessary to establish the diagnosis. Absence of all serologic markers is consistent with a diagnosis of "non-A, non-B, non-C" hepatitis, if the epidemiologic setting is appropriate.

In patients with chronic hepatitis, initial testing should consist ofHBsAgand anti-HCV. Anti-HCV supports and HCV RNA testing establishes the diagnosis of chronic hepatitis C. If a serologic diagnosis of chronic hepatitis B is made, testing forHBeAgand anti-HBe is indicated to evaluate relative infectivity. Testing forHBV DNA in such patients provides a more quantitative and sensitive measure of the level of virus replication and, therefore, is very helpful during antiviral therapy (Chap. 297). In patients with hepatitis B, testing for anti-HDVis useful under the following circumstances: patients with severe and fulminant diseases, patients with severe chronic disease, patients with chronic hepatitis B who have acute hepatitis-like exacerbations, persons with frequent percutaneous exposures, and persons from areas where HDV infection is endemic.

PROGNOSIS

Virtually all previously healthy patients with hepatitis A recover completely from their illness with no clinical sequelae. Similarly, in acute hepatitis B, 95 to 99% of previously healthy adults have a favorable course and recover completely. There are, however, certain clinical and laboratory features that suggest a more complicated and protracted course. Patients of advanced age and with serious underlying medical disorders may have a prolonged course and are more likely to experience severe hepatitis. Initial presenting features such as ascites, peripheral edema, and symptoms of hepatic encephalopathy suggest a poorer prognosis. In addition, a prolongedPT, low serum albumin level, hypoglycemia, and very high serum bilirubin values suggest severe hepatocellular disease. Patients with these clinical and laboratory features deserve prompt hospital admission. The case-fatality rate in hepatitis A and B is very low (approximately 0.1%) but is increased by advanced age and underlying debilitating disorders. Among patients ill enough to be hospitalized for acute hepatitis B, the fatality rate is 1%. Hepatitis C occurring after transfusion is less severe during the acute phase than hepatitis B and is more likely to be anicteric; fatalities are rare, but the precise case-fatality rate is not known. In outbreaks of waterborne hepatitis E in India and Asia, the case-fatality rate is 1 to 2% and up to 10 to 20% in pregnant women. Patients with simultaneous acute hepatitis B and hepatitis D do not necessarily experience a higher mortality rate than do patients with acute hepatitis B alone; however, in several recent outbreaks of acute simultaneousHBV andHDVinfection among injection drug users, the case-fatality rate has been approximately 5%. In the case of HDV superinfection of a person with chronic hepatitis B, the likelihood of fulminant hepatitis and death is increased substantially. Although the case-fatality rate for hepatitis D has not been defined adequately, in outbreaks of severe HDV superinfection in isolated populations with a high hepatitis B carrier rate, the mortality rate has been recorded in excess of 20%.

COMPLICATIONS AND SEQUELAE

A small proportion of patients with hepatitis A experience *relapsing hepatitis* weeks to months after apparent recovery from acute hepatitis. Relapses are characterized by recurrence of symptoms, aminotransferase elevations, occasionally jaundice, and fecal excretion ofHAV. Another unusual variant of acute hepatitis A is *cholestatic hepatitis*, characterized by protracted cholestatic jaundice and pruritus. Rarely, liver test abnormalities persist for many months, even up to a year. Even when these complications occur, hepatitis A remains self-limited and does not progress to chronic liver disease. During the prodromal phase of acute hepatitis B, a serum sickness-like syndrome characterized by arthralgia or arthritis, rash, angioedema, and rarely hematuria and proteinuria may develop in 5 to 10% of patients. This syndrome occurs before the onset of clinical jaundice, and these patients are often erroneously diagnosed as having rheumatologic diseases. The diagnosis can be established by measuring serum aminotransferase levels, which are almost invariably elevated, and serumHBsAg. As noted above,EMC is an immune-complex disease that can complicate hepatitis C. Attention has been drawn as well to associations between hepatitis C and such cutaneous disorders as porphyria cutanea tarda and lichen planus. A mechanism for these associations is unknown.

The most feared complication of viral hepatitis is *fulminant hepatitis* (massive hepatic necrosis); fortunately, this is a rare event. Fulminant hepatitis is primarily seen in hepatitis B and D, as well as hepatitis E, but rare fulminant cases of hepatitis A occur primarily in older adults and in persons with underlying chronic liver disease. Hepatitis B accounts for more than 50% of fulminant hepatitis cases, a sizable proportion of which are associated withHDVinfection. Participation of HDV can be documented in approximately one-third of patients with acute fulminant hepatitis B and two-thirds of patients with fulminant hepatitis superimposed on chronic hepatitis B. Fulminant hepatitis is seen rarely in hepatitis C, but hepatitis E, as noted above, can be complicated by fatal fulminant hepatitis in 1 to 2% of all cases and in up to 20% of cases occurring in pregnant women. Patients usually present with signs and symptoms of encephalopathy that may evolve to deep coma. The liver is usually small and thePTexcessively prolonged. The combination of rapidly shrinking liver size, rapidly rising bilirubin level, and marked prolongation of the PT, even as aminotransferase levels fall, together with clinical signs of confusion, disorientation, somnolence, ascites, and edema, indicates that the patient has hepatic failure with encephalopathy. Cerebral edema is common; brainstem compression, gastrointestinal bleeding, sepsis, respiratory failure, cardiovascular collapse, and renal failure are terminal events. The mortality rate is exceedingly high (greater than 80% in patients with deep coma), but patients who survive may have a complete biochemical and histologic recovery. If a donor liver can be located in time, liver transplantation may be life-saving in patients with fulminant hepatitis.

It is particularly important to document the disappearance ofHBsAgafter apparent clinical recovery from acute hepatitis B. Before laboratory methods were available to distinguish between acute hepatitis and acute hepatitis-like exacerbations (*spontaneous reactivations*) of chronic hepatitis B, observations suggested that approximately 10% of patients remained HBsAg-positive for longer than 6 months after the onset of clinically apparent acute hepatitis B. Half these persons cleared the antigen from their circulations during the next several years, but the other 5% remained chronically HBsAg-positive. More recent observations suggest that the true rate of chronic infection after clinically apparent acute hepatitis B is as low as 1% in normal, immunocompetent, young adults. Earlier, higher estimates may have been biased by inadvertent inclusion of acute exacerbations in chronically infected patients; these patients, chronically HBsAg-positive before exacerbation, were unlikely to seroconvert to HBsAg-negative thereafter. Whether the rate of chronicity is 10 or 1%, such patients have anti-HBc in serum; anti-HBs is either undetected or detected at low titer against the opposite subtype specificity of the antigen (see "Laboratory Features," above). These patients may (1) be asymptomatic carriers, (2) have low-grade, mild chronic hepatitis, or (3) have moderate to severe chronic hepatitis with or without cirrhosis. The likelihood of becoming an HBsAg carrier after acuteHBVinfection is especially high among neonates, persons with Down's syndrome, chronically hemodialyzed patients, and immunosuppressed patients, including persons with HIV infection.

Chronic hepatitis is an important late complication of acute hepatitis B occurring in a small proportion of patients with acute disease but more common in those who present with chronic infection without having experienced an acute illness (Chap. 297). Certain clinical and laboratory features suggest progression of acute hepatitis to chronic hepatitis: (1) lack of complete resolution of clinical symptoms of anorexia, weight loss, and fatigue and the persistence of hepatomegaly; (2) the presence of bridging or multilobular hepatic necrosis on liver biopsy during protracted, severe acute viral hepatitis; (3) failure of the serum aminotransferase, bilirubin, and globulin levels to return to normal within 6 to 12 months after the acute illness; and (4) the persistence ofHBeAgbeyond 3 months orHBsAgbeyond 6 months after acute hepatitis.

Although acute hepatitis D infection does not increase the likelihood of chronicity of simultaneous acute hepatitis B, hepatitis D has the potential for contributing to the severity of chronic hepatitis B. Hepatitis D superinfection can transform asymptomatic or mild chronic hepatitis B into severe, progressive chronic hepatitis and cirrhosis; it also can accelerate the course of chronic hepatitis B. SomeHDVsuperinfections in patients with chronic hepatitis B lead to fulminant hepatitis. Although HDV andHBVinfections are associated with severe liver disease, mild hepatitis and even asymptomatic carriage have been identified in some patients, and the disease may become indolent beyond the early years of infection. After transfusion-associated acute hepatitis C, at least 50% of patients have abnormal biochemical liver tests for more than a year. In some experiences, the frequency of progression to chronicity after acute hepatitis C is as high as 70%. In most of these patients, liver histology is consistent with moderate to severe chronic hepatitis. Even among those who recover biochemically, the likelihood of retaining circulatingHCV RNA is high. Thus, after acute HCV infection, the likelihood of remaining chronically *infected* approaches 85 to 90%. Although many patients with chronic hepatitis C have no symptoms, cirrhosis may develop in as many as 20% within 10 to 20 years of acute illness; in some series of cases, cirrhosis has been reported in as many as 50% of patients with chronic hepatitis C. Although chronic hepatitis C accounts for at least a quarter of cases of chronic liver disease and a quarter of patients undergoing liver transplantation for end-stage liver disease in the United States and Europe, in the majority of patients with chronic hepatitis C, morbidity and mortality are limited during the initial 20 years after the onset of infection. Progression of chronic hepatitis C may be influenced by hepatitis C genotype, age of acquisition, duration of infection, and immunosuppression, as well as by coexisting excessive alcohol use or other hepatitis virus infection. In contrast, neitherHAV norHEVcauses chronic liver disease.

Rare complications of viral hepatitis include pancreatitis, myocarditis, atypical pneumonia, aplastic anemia, transverse myelitis, and peripheral neuropathy. *Carriers* ofHBsAg, particularly those infected in infancy or early childhood, have an enhanced risk of hepatocellular carcinoma. The risk of hepatocellular carcinoma is increased as well in patients with chronic hepatitis C, almost exclusively in patients with cirrhosis, and almost always after at least several decades, usually after three decades of disease (see Chap. 91). In children, hepatitis B may present rarely with anicteric hepatitis, a nonpruritic papular rash of the face, buttocks, and limbs, and lymphadenopathy (papular acrodermatitis of childhood or Gianotti-Crosti syndrome).

DIFFERENTIAL DIAGNOSIS

Viral diseases such as infectious mononucleosis; those due to cytomegalovirus, herpes simplex, and coxsackieviruses; and toxoplasmosis may share certain clinical features with viral hepatitis and cause elevations in serum aminotransferase and less commonly in serum bilirubin levels. Tests such as the differential heterophile and serologic tests for these agents may be helpful in the differential diagnosis ifHBsAg, anti-HBc, IgM anti-HAV, and anti-HCVdeterminations are negative. Aminotransferase elevations can accompany almost any systemic viral infection; other rare causes of liver injury confused with viral hepatitis are infections with *Leptospira*, *Candida*, *Brucella*, *Mycobacteria*, and *Pneumocystis*. A complete drug history is particularly important, for many drugs and certain anesthetic agents can produce a picture of either acute hepatitis or cholestasis (Chap. 296). Equally important is a past history of unexplained "repeated episodes" of acute hepatitis. This history should alert the physician to the possibility that the underlying disorder is chronic hepatitis. Alcoholic hepatitis also must be considered, but usually the serum aminotransferase levels are not as markedly elevated and other stigmata of alcoholism may be present. The finding on liver biopsy of fatty infiltration, a neutrophilic inflammatory reaction, and "alcoholic hyaline" would be consistent with alcohol-induced rather than viral liver injury. Because acute hepatitis may present with right upper quadrant abdominal pain, nausea and vomiting, fever, and icterus, it is often confused with acute cholecystitis, common duct stone, or ascending cholangitis. Patients with acute viral hepatitis may tolerate surgery poorly; therefore, it is important to exclude this diagnosis, and in confusing cases, a percutaneous liver biopsy may be necessary before laparotomy. Viral hepatitis in the elderly is often misdiagnosed as obstructive jaundice resulting from a common duct stone or carcinoma of the pancreas. Because acute hepatitis in the elderly may be quite severe and the operative mortality high, a thorough evaluation including biochemical tests, radiographic studies of the biliary tree, and even liver biopsy may be necessary to exclude primary parenchymal liver disease. Another clinical constellation that may mimic acute hepatitis is right ventricular failure with passive hepatic congestion or hypoperfusion syndromes, such as those associated with shock, severe hypotension, and severe left ventricular failure. Also included in this general category is any disorder that interferes with venous return to the heart, such as right atrial myxoma, constrictive pericarditis, hepatic vein occlusion (Budd-Chiari syndrome), or venoocclusive disease. Clinical features are usually sufficient to distinguish between these vascular disorders and viral hepatitis. Acute fatty liver of pregnancy, cholestasis of pregnancy, eclampsia, and the HELLP syndrome (hemolysis, elevated liver tests, and low platelets) can be confused with viral hepatitis during pregnancy. Very rarely, malignancies metastatic to the liver can mimic acute or even fulminant viral hepatitis. Occasionally, genetic or metabolic liver disorders (e.g., Wilson's disease,a$_1$-antitrypsin deficiency) are confused with viral hepatitis.

TREATMENT

Treatment of Acute Attack Although therapy has been developed for chronic hepatitis B and C (Chap. 297), opportunities for treating acute hepatitis caused byHBV orHCV are limited. In hepatitis B, among previously healthy adults who present with clinically apparent acute hepatitis, recovery occurs in approximately 99%; therefore, antiviral therapy is not likely to improve the rate of recovery and is not required. In rare instances of severe acute hepatitis B, treatment with a nucleoside analogue, such as lamivudine, at the 100-mg/d oral dose used to treat chronic hepatitis B (Chap. 297), has been attempted successfully. However, clinical trials have not been done to establish the efficacy of this approach, severe acute hepatitis B is not an approved indication for therapy, and the duration of therapy has not been determined. In typical cases of acute hepatitis C, recovery is rare, progression to chronic hepatitis is the rule, occurring in 85 to 90% of patients, and meta-analyses of small clinical trials suggest that antiviral therapy with interferon alpha (3 million units subcutaneously three times a week) is beneficial, reducing the rate of chronicity considerably by inducing sustained responses in 40% of patients. The duration of therapy and whether to add the nucleoside analogue ribavirin remain to be determined, but the most reasonable approach is to follow recommendations for treatment of chronic hepatitis C (Chap 297). Because of the marked reduction over the last two decades in the frequency of acute hepatitis C, opportunities to identify and treat patients with acute hepatitis C are rare indeed. Hospital epidemiologists, however, will encounter health workers who sustain hepatitis C-contaminated needle sticks; when monitoring forALTelevations and HCV RNA after these accidents identifies acute hepatitis C, therapy should be initiated.

Notwithstanding these specific therapeutic considerations, in most cases of typical acute viral hepatitis, specific treatment generally is not necessary. Although hospitalization may be required for clinically severe illness, most patients do not require hospital care. Forced and prolonged bed rest is not essential for full recovery, but many patients will feel better with restricted physical activity. A high-calorie diet is desirable, and because many patients may experience nausea late in the day, the major caloric intake is best tolerated in the morning. Intravenous feeding is necessary in the acute stage if the patient has persistent vomiting and cannot maintain oral intake. Drugs capable of producing adverse reactions such as cholestasis and drugs metabolized by the liver should be avoided. If severe pruritus is present, the use of the bile salt-sequestering resin cholestyramine will usually alleviate this symptom. Glucocorticoid therapy has no value in acute viral hepatitis. Even in severe cases associated with *bridging necrosis*, controlled trials have failed to demonstrate the efficacy of steroids. In fact, such therapy may be hazardous.

Physical isolation of patients with hepatitis to a single room and bathroom is rarely necessary except in the case of fecal incontinence for hepatitis A and E or uncontrolled, voluminous bleeding for hepatitis B (with or without concomitant hepatitis D) and hepatitis C. Because most patients hospitalized with hepatitis A excrete little if anyHAV, the likelihood of HAV transmission from these patients during their hospitalization is low. Therefore, burdensome *enteric precautions are no longer recommended*. Although gloves should be worn when the bedpans or fecal material of patients with hepatitis A are handled, these precautions do not represent a departure from sensible procedure for all hospitalized patients. For patients with hepatitis B and hepatitis C, emphasis should be placed on blood precautions, i.e., avoiding direct, ungloved hand contact with blood and other body fluids. Enteric precautions are unnecessary. The importance of simple hygienic precautions, such as hand washing, cannot be overemphasized. Universal precautions that have been adopted for all patients apply to patients with viral hepatitis.

Hospitalized patients may be discharged when there is substantial symptomatic improvement, a significant downward trend in the serum aminotransferase and bilirubin values, and a return to normal of thePT. Mild aminotransferase elevations should not be considered contraindications to the gradual resumption of normal activity.

In *fulminant hepatitis*, the goal of therapy is to support the patient by maintenance of fluid balance, support of circulation and respiration, control of bleeding, correction of hypoglycemia, and treatment of other complications of the comatose state in anticipation of liver regeneration and repair. Protein intake should be restricted, and oral lactulose or neomycin administered. Glucocorticoid therapy has been shown in controlled trials to be ineffective. Likewise, exchange transfusion, plasmapheresis, human cross-circulation, porcine liver cross-perfusion, and hemoperfusion have not been proven to enhance survival. Meticulous intensive care is the one factor that does appear to improve survival. Orthotopic liver transplantation is resorted to with increasing frequency, with excellent results, in patients with fulminant hepatitis (Chap. 301).

PROPHYLAXIS

Because application of therapy for acute viral hepatitis is limited, and because antiviral therapy for chronic viral hepatitis is effective in only a proportion of patients (Chap. 297), emphasis is placed on prevention through immunization. The prophylactic approach differs for each of the types of viral hepatitis. In the past, immunoprophylaxis relied exclusively on passive immunization with antibody-containing globulin preparations purified by cold ethanol fractionation from the plasma of hundreds of normal donors. Currently, for hepatitis A and B, active immunization with vaccines is available as well.

Hepatitis A Both passive immunization with immune globulin (IG) and active immunization with a killed vaccine are available. All preparations of IG contain anti-HAVconcentrations sufficient to be protective. When administered before exposure or during the early incubation period, IG is effective in preventing clinically apparent hepatitis A. In some cases, IG does not abort infection but, by attenuating it, renders it inapparent. As a result, long-lasting "passive-active" immunity occurs; however, this is now considered to be the exception rather than the rule. For postexposure prophylaxis of intimate contacts (household, institutional) of persons with hepatitis A, the administration of 0.02 mL/kg is recommended as early after exposure as possible; it may be effective even when administered as late as 2 weeks after exposure. Prophylaxis is not necessary for casual contacts (office, factory, school, or hospital), for most elderly persons, who are very likely to be immune, or for those known to have anti-HAV in their serum. In day-care centers, recognition of hepatitis A in children or staff should provide a stimulus for immunoprophylaxis in the center and in the children's family members. By the time most common-source outbreaks of hepatitis A are recognized, it is usually too late in the incubation period for IG to be effective; however, prophylaxis may limit the frequency of secondary cases. For travelers to tropical countries, developing countries, and other areas outside standard tourist routes, IG prophylaxis had been recommended, before a vaccine became available. When such travel lasted less than 3 months, 0.02 mL/kg was given; for longer travel or residence in these areas, a dose of 0.06 mL/kg every 4 to 6 months was recommended. Administration of plasma-derived globulin is safe; it has not been associated with transmission of AIDS to recipients, and the AIDS virus (HIV) is inactivated by 25% alcohol, to which plasma is subjected during the cold ethanol fractionation process.

Formalin-inactivated vaccines made from strains ofHAVattenuated in tissue culture have been shown to be safe, immunogenic, and effective in preventing hepatitis A. Hepatitis A vaccines are approved for use in persons who are at least 2 years old and appear to provide adequate protection 4 weeks after a primary inoculation. If it can be given within 4 weeks of an expected exposure, such as by travel to an endemic area, hepatitis A vaccine is the preferred approach to *preexposure* immunoprophylaxis. If travel is more imminent,IG (0.02 mL/kg) should be administered at a different injection site, along with the first dose of vaccine. Because vaccination provides long-lasting protection (protective levels of anti-HAV should last 20 years after vaccination), persons whose risk will be sustained (e.g., frequent travelers or those remaining in endemic areas for prolonged periods) should be vaccinated, and vaccine should supplant the need for repeated IG injections. Other groups who are candidates for hepatitis A vaccination include military personnel, populations with cyclic outbreaks of hepatitis A (e.g., Alaskan natives), employees of day-care centers, primate handlers, laboratory workers exposed to hepatitis A or fecal specimens, children in communities with a high frequency of hepatitis A, and patients with chronic liver disease. Because of an increased risk of fulminant hepatitis A -- observed in some experiences but not confirmed in others -- among patients with chronic hepatitis C, patients with chronic hepatitis C have been singled out as candidates for hepatitis A vaccination. Other populations whose recognized risk of hepatitis A is increased should be vaccinated, including men who have sex with men, injection drug users, and persons with clotting disorders who require frequent administration of clotting-factor concentrates. Recommendations for dose and frequency differ for the two approved vaccine preparations; all injections are intramuscular. For the hepatitis A vaccine manufactured by SmithKline Beecham (Havrix), adults (older than 18 years) should receive two

1.0-mL injections containing 1440 enzyme-linked immunoassay units (ELU) 6 to 12 months apart. Children age 2 to 18 years should receive three 0.5-mL injections containing 360 ELU at time zero, 6, and 12 months or two 0.5-mL injections containing 720 ELU 6 to 12 months apart. For the hepatitis A vaccine manufactured by Merck (Vaqta), adults (older than 17 years) should receive two 1.0-mL injections containing 50 units 6 months apart; children age 2 to 17 years should receive two 0.5-mL doses containing 25 units 6 to 18 months apart. Hepatitis A vaccine has been reported to be effective in preventing secondary household cases of acute hepatitis A, but its role in other instances of postexposure prophylaxis remains to be demonstrated.

Hepatitis B Until 1982, prevention of hepatitis B was based on *passive* immunoprophylaxis either with standardIG, containing modest levels of anti-HBs, or hepatitis B immune globulin (HBIG), containing high-titer anti-HBs. The efficacy of standard IG has never been established and remains questionable; even the efficacy of HBIG, demonstrated in several clinical trials, has been challenged, and its contribution appears to be in reducing the frequency of clinical *illness*, not in preventing *infection*. The first vaccine for *active* immunization, introduced in 1982, was prepared from purified, noninfectious 22-nm spherical forms ofHBsAgderived from the plasma of healthy HBsAg carriers. In 1987, the plasma-derived vaccine was supplanted by a genetically engineered vaccine derived from recombinant yeast. The latter vaccine consists of HBsAg particles that are nonglycosylated but are otherwise indistinguishable from natural HBsAg; two recombinant vaccines are licensed for use in the United States. Current recommendations can be divided into those for preexposure and postexposure prophylaxis.

For *preexposure* prophylaxis against hepatitis B in settings of frequent exposure (health workers exposed to blood, hemodialysis patients and staff, residents and staff of custodial institutions for the developmentally handicapped, injection drug users, inmates of long-term correctional facilities, promiscuous homosexual men as well as promiscuous heterosexual individuals, persons such as hemophiliacs who require long-term, high-volume therapy with blood derivatives, household and sexual contacts ofHBsAgcarriers, persons living in or traveling extensively in endemic areas, unvaccinated children under the age of 18, and unvaccinated children who are Alaskan natives, Pacific Islanders, or residents in households of first-generation immigrants from endemic countries), three intramuscular (deltoid, not gluteal) injections of hepatitis B vaccine are recommended at 0, 1, and 6 months. Pregnancy is *not* a contraindication to vaccination. In areas of lowHBVendemicity such as the United States, despite the availability of safe and effective hepatitis B vaccines, a strategy of vaccinating persons in high-risk groups has not been effective. The incidence of new hepatitis B cases continued to increase in the United States after introduction of vaccines; fewer than 10% of all targeted persons in high-risk groups have actually been vaccinated, and approximately 30% of persons with sporadic acute hepatitis B do not fall into any high-risk-group category. Therefore, to have an impact on the frequency of HBV infection in an area of low endemicity such as the United States, universal hepatitis B vaccination in childhood has been recommended. For unvaccinated children born after the implementation of universal infant vaccination, vaccination during early adolescence, at age 11 to 12 years, was recommended, and this recommendation has been extended to include all unvaccinated children age 0 to 18 years.

The two available recombinant hepatitis B vaccines are comparable, one containing 10 ug ofHBsAg(Recombivax-HB) and the other containing 20 ug of HBsAg (Engerix-B), and recommended doses for each injection vary for the two preparations. For Recombivax-HB, 2.5 ug is recommended for children<11 years of age born to HBsAg-negative mothers, 5 ug for infants born to HBsAg-positive mothers (see below) and for children and adolescents 11 to 19 years of age; 10 ug for immunocompetent adults; and 40 ug for dialysis patients and other immunosuppressed persons. For Engerix-B, 10 ug is recommended for children aged 10 and under, 20 ug for immunocompetent children older than 10 years of age and adults, and 40 ug for dialysis patients and other immunocompromised persons.

For unvaccinated persons sustaining an exposure toHBV, *postexposure* prophylaxis with a combination ofHBIG(for rapid achievement of high-titer circulating anti-HBs) and hepatitis B vaccine (for achievement of long-lasting immunity as well as its apparent efficacy in attenuating clinical illness after exposure) is recommended. For *perinatal* exposure of infants born toHBsAg-positive mothers, a single dose of HBIG, 0.5 mL, should be administered intramuscularly in the thigh *immediately after birth*, followed by a complete course of three injections of recombinant hepatitis B vaccine (see doses above) to be started within the first 12 h of life. For those experiencing a direct percutaneous inoculation or transmucosal exposure to HBsAg-positive blood or body fluids (e.g., accidental *needle stick*, other mucosal penetration, or ingestion), a single intramuscular dose of HBIG, 0.06 mL/kg, administered as soon after exposure as possible, is followed by a complete course of hepatitis B vaccine to begin within the first week. For those exposed by *sexual* contact to a patient with acute hepatitis B, a single intramuscular dose of HBIG, 0.06 mL/kg, should be given within 14 days of exposure, to be followed by a complete course of hepatitis B vaccine. When both HBIG and hepatitis B vaccine are recommended, they may be given at the same time but at separate sites.

The precise duration of protection afforded by hepatitis B vaccine is unknown; however, approximately 80 to 90% of immunocompetent vaccinees retain protective levels of anti-HBs for at least 5 years, and 60 to 80% for 10 years. Thereafter and even after anti-HBs becomes undetectable, protection persists against clinical hepatitis B, hepatitis B surface antigenemia, and chronicHBVinfection. Currently, *booster* immunizations are not recommended routinely, except in immunosuppressed persons who have lost detectable anti-HBs or immunocompetent persons who sustain percutaneousHBsAg-positive inoculations after losing detectable antibody. Specifically, for hemodialysis patients, annual anti-HBs testing is recommended after vaccination; booster doses are recommended when anti-HBs levels fall below 10 mIU/mL.

Hepatitis D Infection with hepatitis D can be prevented by vaccinating susceptible persons with hepatitis B vaccine. No product is available for immunoprophylaxis to preventHDVsuperinfection inHBsAgcarriers; for them, avoidance of percutaneous exposures and limitation of intimate contact with persons who have HDV infection are recommended.

Hepatitis C IG is ineffective in preventing hepatitis C and is no longer recommended for postexposure prophylaxis in cases of perinatal, needle stick, or sexual exposure. Although a prototype vaccine that induces antibodies toHCVenvelope protein has been developed, currently, hepatitis C vaccination is not feasible practically. Genotype and quasispecies viral heterogeneity, as well as rapid evasion of neutralizing antibodies by this rapidly mutating virus, conspire to render HCV a difficult target for immunoprophylaxis with a vaccine. Prevention of transfusion-associated hepatitis C has been accomplished by the following successively introduced measures: Exclusion of commercial blood donors and reliance on a volunteer blood supply; screening donor blood with surrogate markers such asALT (no longer recommended) and anti-HBc, markers that identify segments of the blood donor population with an increased risk of bloodborne infections; exclusion of blood donors in high-risk groups for AIDS and the introduction of anti-HIV screening tests; and progressively sensitive serologic screening tests for anti-HCV. Chemical and heat treatment of blood products used for large-pool and concentrated blood derivates are being pursued.

In the absence of active or passive immunization, prevention of hepatitis C includes behavior changes and precautions to limit exposures to infected persons. Recommendations designed to identify patients with clinically inapparent hepatitis as candidates for medical management have as a secondary benefit the identification of persons whose contacts could be at risk of becoming infected. A so-called "look-back" program has been recommended to identify persons who were transfused before 1992 with blood from a donor found subsequently to have hepatitis C. In addition, anti-HCVtesting is recommended for anyone who received a blood transfusion or a transplanted organ before the introduction of second-generation screening tests in 1992, people who ever used injection drugs, chronically hemodialyzed patients, persons with clotting disorders who received clotting factors made before 1987 from pooled blood products, persons with elevated aminotransferase levels, health workers exposed to HCV-positive blood or contaminated needles, and children born to HCV-positive mothers.

For stable, monogamous sexual partners, sexual transmission of hepatitis C is unlikely, and sexual barrier precautions are not recommended. For persons with multiple sexual partners or with sexually transmitted diseases, the risk of sexual transmission of hepatitis C is increased, and barrier precautions (latex condoms) are recommended. A person with hepatitis C should avoid sharing such items as razors, toothbrushes, and nail clippers with sexual partners and family members. No special precautions are recommended for babies born to mothers with hepatitis C, and breast feeding does not have to be restricted.

Hepatitis E WhetherIGprevents hepatitis E remains undetermined. Development of a vaccine is in progress.

(Bibliography omitted in Palm version)

Back to Table of Contents

296. TOXIC AND DRUG-INDUCED HEPATITIS - *Jules L. Dienstag*, *Kurt J. Isselbacher*

Liver injury may follow the inhalation, ingestion, or parenteral administration of a number of pharmacologic and chemical agents. These include industrial toxins (e.g., carbon tetrachloride, trichloroethylene, and yellow phosphorus), the heat-stable toxic bicyclic octapeptides of certain species of *Amanita* and *Galerina* (hepatotoxic mushroom poisoning), and, more commonly, pharmacologic agents used in medical therapy. It is essential that any patient presenting with jaundice or altered biochemical liver tests be questioned carefully about exposure to chemicals used in work or at home and drugs taken by prescription or bought "over the counter." Hepatotoxic drugs can injure the hepatocyte directly, e.g., via a free-radical or metabolic intermediate that causes peroxidation of membrane lipids and that results in liver cell injury. Alternatively, the drug or its metabolite can distort cell membranes or other cellular molecules or block biochemical pathways or cellular integrity. Such injuries, in turn, may lead to necrosis of hepatocytes; injure bile ducts, producing cholestasis; or block pathways of lipid movement, inhibit protein synthesis, or impair mitochondrial oxidation of fatty acids, resulting in fat accumulation (steatosis). In general, two major types of chemical hepatotoxicity have been recognized: (1) direct toxic type and (2) idiosyncratic type.

Most drugs, which are water-insoluble, undergo a series of metabolic transformation steps, culminating in a water-soluble form appropriate for renal or biliary excretion. This process begins with oxidation or methylation initially mediated by the mixed-function oxygenases cytochrome P450 (phase I reaction), followed by glucuronidation or sulfation (phase II reaction) or inactivation by glutathione. Most drug hepatotoxicity is mediated by a phase I toxic metabolite, but glutathione depletion, precluding inactivation of harmful compounds by glutathione S-transferase, can contribute as well.

As shown in Table 296-1, direct toxic hepatitis occurs with predictable regularity in individuals exposed to the offending agent and is dose-dependent. The latent period between exposure and liver injury is usually short (often several hours), although clinical manifestations may be delayed for 24 to 48 h. Agents producing toxic hepatitis are generally systemic poisons or are converted in the liver to toxic metabolites. The direct hepatotoxins result in morphologic abnormalities that are reasonably characteristic and reproducible for each toxin. For example, carbon tetrachloride and trichloroethylene characteristically produce a centrilobular zonal necrosis, whereas yellow phosphorus poisoning typically results in periportal injury. The hepatotoxic octapeptides of *Amanita phalloides* usually produce massive hepatic necrosis. The lethal dose of the toxin is about 10 mg, the amount found in a single deathcap mushroom. Tetracycline, when administered in intravenous doses>1.5 g daily, leads to microvesicular fat deposits in the liver. Liver injury, which is often only one facet of the toxicity produced by the direct hepatotoxins, may go unrecognized until jaundice appears.

In idiosyncratic drug reactions the occurrence of hepatitis is usually infrequent and unpredictable, the response is not dose-dependent, and it may occur at any time during or shortly after exposure to the drug. Extrahepatic manifestations of hypersensitivity, such as rash, arthralgias, fever, leukocytosis, and eosinophilia, occur in about one-quarter of patients with idiosyncratic hepatotoxic drug reactions; this observation and the unpredictability of idiosyncratic drug hepatotoxicity contributed to the hypothesis that this category of drug reactions is immunologically mediated. More recent evidence, however, suggests that, in most cases, even idiosyncratic reactions represent direct hepatotoxity but are caused by drug metabolites rather than by the intact compound. Even the prototype of idiosyncratic hepatoxicity reactions, halothane hepatitis, and isoniazid hepatotoxicity, associated frequently with hypersensitivity manifestations, are now recognized to be mediated by toxic metabolites that damage liver cells directly. Currently, most idiosyncratic reactions are thought to result from differences in metabolic reactivity to specific agents; host susceptibility is mediated by the kinetics of toxic metabolite generation, which differs among individuals. Occasionally, however, the clinical features of an allergic reaction (prominent tissue eosinophilia, autoantibodies, etc.) are difficult to ignore. In vitro models have been described in which lymphocyte cytotoxicity can be demonstrated against rabbit hepatocytes altered by incubation with the potential offending drug. Furthermore, several instances of drug hepatotoxicity are associated with the appearance of autoantibodies, including a class of antibodies to liver-kidney microsomes, anti-LKM2, directed against a cytochrome P450 enzyme. Similarly, in selected cases, a drug or its metabolite has been shown to bind to a host cellular component forming a hapten; the immune response to this "neoantigen" is postulated to play a role in the pathogenesis of liver injury. Therefore, some authorities subdivide idiosyncratic drug hepatotoxicity into hypersensitivity (allergic) and "metabolic" categories. Several unusual exceptions notwithstanding, true drug allergy is difficult to support in most cases of idiosyncratic drug-induced liver injury.

Idiosyncratic reactions lead to a morphologic pattern that is more variable than those produced by direct toxins; a single agent is often capable of causing a variety of lesions, although certain patterns tend to predominate. Depending on the agent involved, idiosyncratic hepatitis may result in a clinical and morphologic picture indistinguishable from that of viral hepatitis (e.g., halothane) or may simulate extrahepatic bile duct obstruction clinically with morphologic evidence of cholestasis. Drug-induced cholestasis ranges from mild to increasingly severe: (1) bland cholestasis with limited hepatocellular injury (e.g., estrogens, 17,a-substituted androgens); (2) inflammatory cholestasis (e.g., phenothiazines, amoxicillin-clavulanic acid, oxacillin, erythromcyin estolate); (3) sclerosing cholangitis (e.g., after intrahepatic infusion of the chemotherapeutic agent floxuridine for hepatic metastases from a primary colonic carcinoma); (4) disappearance of bile ducts, "ductopenic" cholestasis, similar to that observed in chronic rejection following liver transplantation (e.g., carbamazine, chlorpromazine, tricyclic antidepressant agents). Morphologic alterations may also include bridging hepatic necrosis (e.g., methyldopa), or, infrequently, hepatic granulomas (e.g., sulfonamides). Some drugs result in macrovesicular or microvesicular steatosis or steatohepatitis, which in some cases has been linked to mitochondrial dysfunciton and lipid peroxidation. Severe hepatotoxicity associated with steatohepatitis, most likely a result of mitochondrial toxicity, is being recognized with increasing frequency among patients receiving antiretroviral therapy with reverse transcriptase inhibitors (e.g., zidovudine, didanosine) or protease inhibitors (e.g., indinavir, ritonavir) for HIV infection.

Not all adverse hepatic drug reactions can be classified as either toxic or idiosyncratic in type. For example, oral contraceptives, which combine estrogenic and progestational compounds, may result in impairment of hepatic tests and occasionally in jaundice. However, they do not produce necrosis or fatty change, manifestations of hypersensitivity are generally absent, and susceptibility to the development of oral contraceptive-induced cholestasis appears to be genetically determined. Other instances of genetically determined drug hepatotoxicity have been identified. For example, approximately 10% of the population have an autosomally recessive trait associated with the absence of cytochrome P450 enzyme 2D6 and have impaired debrisoquine-4-hydroxylase enzyme activity. As a result, they cannot metabolize, and are at increased risk of hepatotoxicity resulting from, certain compounds such as desipramine, propranolol, and quinidine.

Because drug-induced hepatitis is often a presumptive diagnosis and many other disorders produce a similar clinicopathologic picture, evidence of a causal relationship between the use of a drug and subsequent liver injury may be difficult to establish. The relationship is most convincing for the direct hepatotoxins, which lead to a high frequency of hepatic impairment after a short latent period. Idiosyncratic reactions may be reproduced, in some instances, when rechallenge, after an asymptomatic period, results in a recurrence of signs, symptoms, and morphologic and biochemical abnormalities. Rechallenge, however, is often ethically unfeasible, because severe reactions may occur.

TREATMENT

Treatment of toxic and drug-induced hepatic disease is largely supportive, except in acetaminophen hepatotoxicity (see below). In patients with fulminant hepatitis resulting from drug hepatotoxicity, liver transplantation may be life-saving (Chap. 301). Withdrawal of the suspected agent is indicated at the first sign of an adverse reaction. In the case of the direct toxins, liver involvement should not divert attention from renal or other organ involvement, which may also threaten survival.

InTable 296-2, several classes of chemical agents are listed, together with examples of the pattern of liver injury produced by them. Certain drugs appear to be responsible for the development of chronic as well as acute hepatic injury. For example, oxyphenisatin, methyldopa, and isoniazid have been associated with moderate to severe chronic hepatitis, and halothane and methotrexate have been implicated in the development of cirrhosis. A syndrome resembling primary biliary cirrhosis has been described following treatment with chlorpromazine, methyl testosterone, tolbutamide, and other drugs. Portal hypertension in the absence of cirrhosis may result from alterations in hepatic architecture produced by vitamin A or arsenic intoxication, industrial exposure to vinyl chloride, or administration of thorium dioxide. The latter three agents have also been associated with angiosarcoma of the liver. Oral contraceptives have been implicated in the development of hepatic adenoma and, rarely, hepatocellular carcinoma and occlusion of the hepatic vein (Budd-Chiari syndrome). Another unusual lesion, peliosis hepatis (blood cysts of the liver), has been observed in some patients treated with anabolic steroids. The existence of these hepatic disorders expands the spectrum of liver injury induced by chemical agents and emphasizes the need for a thorough drug history in all patients with liver dysfunction.

The following are the patterns of adverse hepatic reactions for some prototypic agents.

ACETAMINOPHEN HEPATOTOXICITY (DIRECT TOXIN)

Acetaminophen has caused severe centrilobular hepatic necrosis when ingested in large amounts in suicide attempts or accidentally by children. A single dose of 10 to 15 g, occasionally less, may produce clinical evidence of liver injury. Fatal fulminant disease is usually (although not invariably) associated with ingestion of 25 g or more. Blood levels of acetaminophen correlate with the severity of hepatic injury (levels >300 ug/mL 4 h after ingestion are predictive of the development of severe damage; levels<150 ug/mL suggest that hepatic injury is highly unlikely). Nausea, vomiting, diarrhea, abdominal pain, and shock are early manifestations occurring 4 to 12 h after ingestion. Then 24 to 48 h later, when these features are abating, hepatic injury becomes apparent. Maximal abnormalities and hepatic failure may not be evident until 4 to 6 days after ingestion, and aminotransferase levels approaching 10,000 units are not uncommon. Renal failure and myocardial injury may be present.

Acetaminophen is metabolized predominantly by a phase II reaction to innocuous sulfate and glucuronide metabolites; however, a small proportion of acetaminophen is metabolized by a phase I reaction to a hepatotoxic metabolite formed from the parent compound by the cytochrome P450 2E1. This metabolite, *N*-acetyl-benzoquinone-imide (NAPQI), is detoxified by binding to "hepatoprotective" glutathione to become harmless, water-soluble mercapturic acid, which undergoes renal excretion. When excessive amounts of NAPQI are formed, or when glutathione levels are low, glutathione levels are depleted and overwhelmed, permitting covalent binding to nucleophilic hepatocyte macromolecules. This process is believed to lead to hepatocyte necrosis; the precise sequence and mechanism are unknown. Hepatic injury may be potentiated by prior administration of alcohol or other drugs, by conditions that stimulate the mixed-function oxidase system, or by conditions such as starvation that reduce hepatic glutathione levels. Cimetidine, which inhibits P450 enzymes, has the potential to reduce generation of the toxic metabolite. Alcohol induces cytochrome P450 2E1; consequently, increased levels of the toxic metabolite NAPQI are produced in chronic alcoholics after acetaminophen ingestion. In addition, alcohol suppresses hepatic glutathione production. Therefore, in chronic alcoholics, the toxic dose of acetaminophen may be as low as 2 g, and alcoholic patients should be warned specifically about the dangers of even standard doses of this commonly used drug. Such "therapeutic misadventures" also occur occasionally in patients with severe, febrile illnesses or pain syndromes; in such a setting, several days of anorexia and near-fasting coupled with regular administration of extra-strength acetaminophen formulations result in a combination of glutathione depletion and relatively high NAPQI levels in the absence of a history of recognized acetaminophen overdose.

TREATMENT

Treatment of acetaminophen overdosage includes gastric lavage, supportive measures, and oral administration of activated charcoal or cholestyramine to prevent absorption of residual drug. Neither of these agents appears to be effective if given more than 30 min after acetaminophen ingestion; if they are used, the stomach lavage should be done before other agents are administered orally. The chances of possible-, probable-, and high-risk hepatotoxicity can be derived from a nomogram plot (see Fig. 396-2), readily available in emergency departments, of acetaminophen plasma levels as a function of hours after ingestion. In patients with high acetaminophen blood levels (>200 ug/mL measured at 4 h or>100 ug/mL at 8 h after ingestion), the administration of sulfhydryl compounds (e.g., cysteamine, cysteine, or *N*-acetylcysteine) appears to reduce the severity of hepatic necrosis. These agents appear to act by providing a reservoir of sulfhydryl groups to bind the toxic metabolites or by stimulating synthesis and repletion of hepatic glutathione. Therapy should be begun within 8 h of ingestion but may be effective even if given as late as 24 to 36 h after overdose. Later administration of sulfhydryl compounds is of uncertain value. Routine use of *N*-acetylcysteine has reduced substantially the occurrence of fatal acetaminophen hepatotoxicity. When given orally, *N*-acetylcysteine is diluted to yield a 5% solution. A loading dose of 140 mg/kg is given, followed by 70 mg/kg every 4 h for 15 to 20 doses. Whenever a patient with potential acetaminophen hepatotoxicity is encountered, a local poison control center should be contacted. Treatment can be stopped when plasma acetominophen levels indicate that the risk of liver damage is low.

Survivors of acute acetaminophen overdose usually have no evidence of hepatic sequelae. In a few patients, prolonged or repeated administration of acetaminophen in therapeutic doses appears to have led to the development of chronic hepatitis and cirrhosis.

HALOTHANE HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

Administration of halothane, a nonexplosive fluorinated hydrocarbon anesthetic agent that is structurally similar to chloroform, results in severe hepatic necrosis in a small number of individuals, many of whom have previously been exposed to this agent. The failure to produce similar hepatic lesions reliably in animals, the rarity of hepatic impairment in human beings, and the delayed appearance of hepatic injury suggest that halothane is not a direct hepatotoxin but rather a sensitizing agent. However, manifestations of hypersensitivity are seen in<25% of cases. A genetic predisposition leading to an idiosyncratic metabolic reactivity has been postulated and appears to be the most likely mechanism of halothane hepatotoxicity. Adults (rather than children), obese people, and women appear to be particularly susceptible. Fever, moderate leukocytosis, and eosinophilia may occur in the first week following halothane administration. Jaundice is usually noted 7 to 10 days after exposure but may occur earlier in previously exposed patients. Nausea and vomiting may precede the onset of jaundice. Hepatomegaly is often mild, but liver tenderness is common. The serum aminotransferase levels are elevated. The pathologic changes at autopsy are indistinguishable from massive hepatic necrosis resulting from viral hepatitis. The case-fatality rate of halothane hepatitis is not known but may vary from 20 to 40% in cases with severe liver involvement. It is strongly suggested that patients in whom unexplained spiking fever, especially delayed fever, or jaundice develops after halothane anesthesia not receive this agent again. Because cross-reactions between halothane and methoxyfluorane have been reported, the latter agent should not be used after halothane reactions. Later-generation halogenated hydrocarbon anesthetics, which have supplanted halothane except in rare instances, are felt to be associated with a lower risk of hepatotoxicity.

METHYLDOPA HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Minor alterations in liver tests are reported in about 5% of patients treated with this antihypertensive agent. These trivial abnormalities typically resolve despite continued drug administration. In <1% of patients, acute liver injury resembling viral or chronic hepatitis or, rarely, a cholestatic reaction is seen 1 to 20 weeks after methyldopa is started. In 50% of cases the interval is <4 weeks. A prodrome of fever, anorexia, and malaise may be noted for a few days before the onset of jaundice. Rash, lymphadenopathy, arthralgia, and eosinophilia are rare. Serologic markers of autoimmunity are detected infrequently, and<5% of patients have a Coombs-positive hemolytic anemia. In about 15% of patients with methyldopa hepatotoxicity, the clinical, biochemical, and histologic features are those of moderate to severe chronic hepatitis, with or without bridging necrosis and macronodular cirrhosis. With discontinuation of the drug, the disorder usually resolves.

ISONIAZID HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

In approximately 10% of adults treated with the antituberculosis agent isoniazid, elevated serum aminotransferase levels develop during the first few weeks of therapy; this appears to represent an adaptive response to a toxic metabolite of the drug. Whether or not isoniazid is continued, these values (usually<200 units) return to normal in a few weeks. In about 1% of treated patients, an illness develops that is indistinguishable from viral hepatitis; approximately half of these cases occur within the first 2 months of treatment, while in the remainder, clinical disease may be delayed for many months. Liver biopsy reveals morphologic changes similar to those of viral hepatitis or bridging hepatic necrosis. The disease may be severe, with a case-fatality rate of 10%. Important liver injury appears to be age-related, increasing substantially after age 35; the highest frequency is in patients over age 50, the lowest under the age of 20. Even for patients>50 years of age monitored carefully during therapy, hepatotoxicity occurs in only approximately 2%, well below the risk estimate derived from earlier experiences. Isoniazid hepatotoxicity is enhanced by alcohol and rifampicin. Fever, rash, eosinophilia, and other manifestations of drug allergy are distinctly unusual. A reactive metabolite of acetylhydrazine, a metabolite of isoniazid, may be responsible for liver injury, and patients who are rapid acetylators would be more prone to such injury. A picture resembling chronic hepatitis has been observed in a few patients.

SODIUM VALPROATE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Sodium valproate, an anticonvulsant useful in the treatment of petit mal and other seizure disorders, has been associated with the development of severe hepatic toxicity and, rarely, fatalities, predominantly in children but also in adults. Asymptomatic elevations of serum aminotransferase levels have been recognized in as many as 45% of treated patients. These "adaptive" changes, however, appear to have no clinical importance, for major hepatotoxicity is not seen in the majority of patients despite continuation of drug therapy. In those rare patients in whom jaundice, encephalopathy, and evidence of hepatic failure are found, examination of liver tissue reveals microvesicular fat and bridging hepatic necrosis, predominantly in the centrilobular zone. Bile duct injury may also be apparent. It seems likely that sodium valproate is not directly hepatotoxic but that its metabolite, 4-pentenoic acid, may be responsible for hepatic injury.

PHENYTOIN HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

Phenytoin, formerly diphenylhydantoin, a mainstay in the treatment of seizure disorders, has been associated in rare instances with the development of severe hepatitis-like liver injury leading to fulminant hepatic failure. In many patients the hepatitis is associated with striking fever, lymphadenopathy, rash (Stevens-Johnson syndrome or exfoliative dermatitis), leukocytosis, and eosinophilia, suggesting an immunologically mediated hypersensitivity mechanism. Despite these observations, there is also evidence that metabolic idiosyncrasy may be responsible for hepatic injury. In the liver, phenytoin is converted by the cytochrome P450 system to metabolites, which include the highly reactive electrophilic arene oxides. These metabolites are normally metabolized further by epoxide hydrolases. A defect (genetic or acquired) in epoxide hydrolase activity could permit covalent binding of arene oxides to hepatic macromolecules, thereby leading to hepatic injury. Regardless of the mechanism, hepatic injury is usually manifest within the first 2 months after beginning phenytoin therapy. With the exception of an abundance of eosinophils in the liver, the clinical, biochemical, and histologic picture resembles that of viral hepatitis. In rare instances, bile duct injury may be the salient feature of phenytoin hepatotoxicity, with striking features of intrahepatic cholestasis. Asymptomatic elevations of aminotransferase and alkaline phosphatase levels have been observed in a sizable proportion of patients receiving long-term phenytoin therapy. These liver changes are believed by some authorities to represent the potent hepatic enzyme-inducing properties of phenytoin and are accompanied histologically by swelling of hepatocytes in the absence of necroinflammatory activity or evidence of chronic liver disease.

CHLORPROMAZINE HEPATOTOXICITY (CHOLESTATIC IDIOSYNCRATIC REACTION)

In about 1% of patients receiving chlorpromazine, intrahepatic cholestasis with jaundice develops after 1 to 4 weeks of treatment. In rare instances, jaundice has been reported after a single exposure. Anicteric reactions are frequent. The onset may be abrupt, with fever, rash, arthralgias, lymphadenopathy, nausea, vomiting, and epigastric or right upper quadrant pain. Pruritus may precede the appearance of jaundice, dark urine, and light stools. Eosinophilia with or without mild leukocytosis may be present, and conjugated hyperbilirubinemia, moderately elevated serum alkaline phosphatase, and mildly elevated serum aminotransferase levels (100 to 200 units) are noted. Liver biopsy reveals cholestasis, bile plugs in dilated bile canaliculi, and a dense portal infiltrate of polymorphonuclear, eosinophilic, and mononuclear leukocytes. Occasionally, scattered foci of hepatic parenchymal necrosis may be evident. Jaundice and pruritus usually subside within 4 to 8 weeks following cessation of therapy, without sequelae, and fatalities are rare. Cholestyramine may be of value in relieving severe pruritus. In a small number of patients, jaundice is prolonged for several months to years; rarely, a disorder resembling but distinct from primary biliary cirrhosis may develop.

AMIODARONE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Therapy with this potent antiarrhythmic drug is accompanied in 15 to 50% of patients by modest elevations of serum aminotransferase levels that may remain stable or diminish despite continuation of the drug. Such abnormalities may appear days to many months after beginning therapy. A proportion of those with elevated aminotransferase levels have detectable hepatomegaly, and clinically important liver disease develops in <5% of patients. Features that represent a direct effect of the drug on the liver and that are common to the majority of long-term recipients are ultrastructural phospholipidosis, unaccompanied by clinical liver disease, and interference with hepatic mixed-function oxidase metabolism of other drugs. The cationic amphiphilic drug and its major metabolite desethylamiodarone accumulate in hepatocyte lysosomes and mitochondria and in bile duct epithelium. The relatively common elevations in aminotransferase levels are also considered a predictable, dose-dependent, direct hepatotoxic effect. On the other hand, in the rare patient with clinically apparent, symptomatic liver disease, liver injury resembling that seen in alcoholic liver disease is observed. The so-called pseudoalcoholic liver injury can range from steatosis, to alcoholic hepatitis-like neutrophilic infiltration and Mallory's hyaline, to cirrhosis. Electron-microscopic demonstration of phospholipid-laden lysosomal lamellar bodies can help to distinguish amiodarone hepatotoxicity from typical alcoholic hepatitis. This category of liver injury appears to be a metabolic idiosyncracy that allows hepatotoxic metabolites to be generated. Rarely, an acute idiosyncratic hepatocellular injury resembling viral hepatitis or cholestatic hepatitis occurs. Hepatic granulomas have occasionally been observed. Because amiodarone has a long half-life, liver injury may persist for months after the drug is stopped.

ERYTHROMYCIN HEPATOTOXICITY (CHOLESTATIC IDIOSYNCRATIC REACTION)

The most important adverse effect associated with erythromycin, more common in children than adults, is the infrequent occurrence of a cholestatic reaction. Although most of these reactions have been associated with erythromycin estolate, other erythromycins may also be responsible. The reaction usually begins during the first 2 or 3 weeks of therapy and includes nausea, vomiting, fever, right upper quadrant abdominal pain, jaundice, leukocytosis, and moderately elevated aminotransferase levels. The clinical picture can resemble acute cholecystitis or bacterial cholangitis. Liver biopsy reveals variable cholestasis; portal inflammation comprising lymphocytes, polymorphonuclear leukocytes, and eosinophils; and scattered foci of hepatocyte necrosis. Symptoms and laboratory findings usually subside within a few days of drug withdrawal, and evidence of chronic liver disease has not been found on follow-up. The precise mechanism remains ill-defined.

ORAL CONTRACEPTIVE HEPATOTOXICITY (CHOLESTATIC REACTION)

The administration of oral contraceptive combinations of estrogenic and progestational steroids leads to intrahepatic cholestasis with pruritus and jaundice in a small number of patients weeks to months after taking these agents. Especially susceptible seem to be patients with recurrent idiopathic jaundice of pregnancy, severe pruritus of pregnancy, or a family history of these disorders. With the exception of liver biochemical tests, laboratory studies are normal, and extrahepatic manifestations of hypersensitivity are absent. Liver biopsy reveals cholestasis with bile plugs in dilated canaliculi and striking bilirubin staining of liver cells. In contrast to chlorpromazine-induced cholestasis, portal inflammation is absent. The lesion is reversible on withdrawal of the agent. The two steroid components appear to act synergistically on hepatic function, although the estrogen may be primarily responsible. Oral contraceptives are contraindicated in patients with a history of recurrent jaundice of pregnancy. Primarily benign, but rarely malignant, neoplasms of the liver, hepatic vein occlusion, and peripheral sinusoidal dilatation have also been associated with oral contraceptive therapy.

17,a-ALKYL-SUBSTITUTED ANABOLIC STEROIDS (CHOLESTATIC REACTION)

In the majority of patients receiving these agents, used therapeutically mainly in the treatment of bone marrow failure but used surreptitiously and without medical indication by athletes to improve their performance, mild hepatic dysfunction develops. Impaired excretory function is the predominant defect, but the precise mechanism is uncertain. Jaundice, which appears to be dose-related, develops in only a minority of patients and may be the sole clinical manifestation of hepatotoxicity, although anorexia, nausea, and malaise may occur. Pruritus is not a prominent feature. Serum aminotransferase levels are usually <100 units, and serum alkaline phosphatase levels are normal, mildly elevated, or, in <5% of patients, three or more times the upper limit of normal. Examination of liver tissue reveals cholestasis without inflammation or necrosis. Hepatic sinusoidal dilatation and peliosis hepatis have been found in a few patients. The cholestatic disorder is usually reversible on cessation of treatment, although fatalities have been linked to peliosis. An association with hepatic adenoma and hepatocellular carcinoma has been reported.

TRIMETHOPRIM-SULFAMETHOXAZOLE HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

This antibiotic combination is used routinely for urinary tract infections in immunocompetent persons and for prophylaxis against and therapy of *Pneumocystis carinii* pneumonia in immunosuppressed persons (transplant recipients, patients with AIDS). With its increasing use, its occasional hepatotoxicity is being recognized with growing frequency. Its likelihood is unpredictable, but when it occurs, trimethoprim-sulfamethoxazole hepatotoxicity follows a relatively uniform latency period of several weeks and is often accompanied by eosinophilia, rash, and other features of a hypersensitivity reaction. Biochemically and histologically, acute hepatocellular necrosis predominates, but cholestatic features are quite frequent. Occasionally, cholestasis without necrosis occurs, and very rarely, a severe cholangiolytic pattern of liver injury is observed. In most cases, liver injury is self-limited, but rare fatalities have been recorded. The hepatotoxicity is attributable to the sulfamethoxazole component of the drug and is similar in features to that seen with other sulfonamides; tissue eosinophilia and granulomas may be seen.

HYDROXYMETHYLGLUTARYL-COENZYME (HMG-COA) REDUCTASE INHIBITORS ("STATINS") (IDIOSYNCRATIC MIXED HEPATOCELLULAR AND CHOLESTATIC REACTION)

Between 1 and 2% of patients taking lovastatin, simvastatin, pravastatin, fluvastatin, or one of the newer "statin" drugs for the treatment of hypercholesterolemia experience asymptomatic, reversible elevations (> threefold) of aminotransferase activity. Acute hepatitis-like histologic changes, centrilobular necrosis, and centrilobular cholestasis have been described in several cases. In a larger proportion, minor aminotrasferase elevations appear during the first several weeks of therapy. Careful laboratory monitoring can distinguish between patients with minor, transitory changes, who may continue therapy, and those with more profound and sustained abnormalities, who should discontinue therapy.

TOTAL PARENTERAL NUTRITION (STEATOSIS, CHOLESTASIS)

Total parenteral nutrition (TPN) is often complicated by cholestatic hepatitis attributable to either steatosis, cholestasis, or gallstones (or gallbladder sludge). Steatosis or steatohepatitis may result from the excess carbohydrate calories in these nutritional supplements and is the predominant form of TPN-associated liver disorder in adults. The frequency of this complication has been reduced substantially by the introduction of balanced TPN formulas that rely on lipid as an alternative caloric source. Cholestasis and cholelithiasis, caused by the absence of stimulation of bile flow and secretion resulting from the lack of oral intake, is the predominant form of TPN-associated liver disease in infants, especially in premature neonates. Often, cholestasis in such neonates is multifactorial, contributed to by other factors such as sepsis, hypoxemia, and hypotension; occasionally, TPN-induced cholestasis in neonates culminates in chronic liver disease and liver failure. When TPN-associated liver test abnormalities occur in adults, balancing the TPN formula with more lipid is the intervention of first recourse. In infants with TPN-associated cholestasis, the addition of oral feeding may ameliorate the problem. Therapeutic interventions suggested, but not yet shown to be of proven benefit, include CCK, ursodeoxycholic acid, *S*-adenosyl methionine, and taurine.

"ALTERNATIVE MEDICINES" (IDIOSYNCRATIC HEPATITIS, STEATOSIS)

The misguided popularity of herbal medications that are of scientifically unproven efficacy and that lack prospective safety oversight by regulatory agencies has resulted in occasional instances of hepatotoxicity. Included among the herbal remedies associated with toxic hepatitis are jin bu huan (Chap. 11), xiao-chai-hu-tang, germander, chaparral, senna, mistletoe, skullcap, gentian, comfrey (containing pyrrolizidine alkaloids), and herbal teas. Recently well characterized are the acute hepatitis-like histologic lesions following jin bu huan use: focal hepatocellular necrosis, mixed mononuclear portal tract infiltration, coagulative necrosis, apoptotic hepatocyte degeneration, tissue eosinophilia, and microvesicular steatosis. Megadoses of vitamin A can injure the liver, as can pyrrolizidine alkaloids, which often contaminate Chinese herbal preparations and can cause a venoocclusive injury leading to sinusoidal hepatic vein obstruction. Given the widespread use of such poorly defined herbal preparations, hepatotoxicity is likely to be encountered with increasing frequency; therefore, a drug history in patients with acute and chronic liver disease should include use of "alternative medicines" and other nonprescription preparations sold in so-called health food stores.

(Bibliography omitted in Palm version)

Back to Table of Contents

297. CHRONIC HEPATITIS - *Jules L. Dienstag*, *Kurt J. Isselbacher*

Chronic hepatitis represents a series of liver disorders of varying causes and severity in which hepatic inflammation and necrosis continue for at least 6 months. Milder forms are nonprogressive or only slowly progressive, while more severe forms may be associated with scarring and architectural reorganization, which, when advanced, lead ultimately to cirrhosis. Several categories of chronic hepatitis have been recognized. These include chronic viral hepatitis (Chap. 295), drug-induced chronic hepatitis (Chap. 296), and autoimmune chronic hepatitis. In many cases, clinical and laboratory features are insufficient to allow assignment into one of these three categories; these "idiopathic" cases are also believed to represent autoimmune chronic hepatitis. Finally, clinical and laboratory features of chronic hepatitis are observed occasionally in patients with such hereditary/metabolic disorders as Wilson's disease (copper overload) and even occasionally in patients with alcoholic liver injury (Chap. 298). Although all types of chronic hepatitis share certain clinical, laboratory, and histopathologic features, chronic viral and chronic autoimmune hepatitis are sufficiently distinct to merit separate discussions.

CLASSIFICATION OF CHRONIC HEPATITIS

Common to all forms of chronic hepatitis are histopathologic distinctions based on localization and extent of liver injury. These vary from the milder forms, previously labeled chronic persistent hepatitis and chronic lobular hepatitis, to the more severe form, formerly called chronic active hepatitis. When first defined, these designations were felt to have prognostic implications, which have been challenged by more recent observations. Compared to the time more than two decades ago when the histologic designations chronic persistent, chronic lobular, and chronic active hepatitis were adopted, much more information is currently available about the causes, natural history, pathogenesis, serologic features, and therapy of chronic hepatitis. Therefore, categorization of chronic hepatitis based primarily upon histopathologic features has been replaced by a more informative classification based upon a combination of clinical, serologic, and histologic variables. Classification of chronic hepatitis is based upon (1) its *cause*, (2) its histologic activity, or *grade*, and (3) its degree of progression, or *stage*. Thus, neither clinical features alone nor histologic features -- requiring liver biopsy -- alone are sufficient to characterize and distinguish among the several categories of chronic hepatitis.

Classification by Cause Clinical and serologic features allow the establishment of a diagnosis of *chronic viral hepatitis*, caused by hepatitis B, hepatitis B plus D, hepatitis C, or potentially other unknown viruses; *autoimmune hepatitis*, including several subcategories, types 1, 2, and 3, based on serologic distinctions; *drug-associated chronic hepatitis*; and a category of unknown cause, or *cryptogenic* chronic hepatitis (Table 297-1). These are addressed in more detail below.

Classification by Grade Grade, a histologic assessment of necroinflammatory activity, is based upon examination of the liver biopsy. An assessment of important histologic features includes the degree of *periportal necrosis* and the disruption of the limiting plate of periportal hepatocytes by inflammatory cells (so-called *piecemeal necrosis* or *interface hepatitis*); the degree of confluent necrosis that links or forms bridges between vascular structures -- between portal tract and portal tract or even more important bridges between portal tract and central vein -- referred to as *bridging necrosis*; the degree of hepatocyte degeneration and focal necrosis within the lobule; and the degree of *portal inflammation*. Several scoring systems that take these histologic features into account have been devised, and the most popular is the numerical histologic activity index (HAI), based on the work of Knodell and Ishak (Table 297-2). Technically, the HAI, which is primarily a measure of *grade*, also includes an assessment of fibrosis, which is currently used to categorize *stage* of the disease, as described below. Such precise HAI scoring tends to be used more in measuring disease activity before and after therapy in clinical studies. In clinical practice, more qualitative grading suffices. Based on the presence and degree of these features of histologic activity, chronic hepatitis can be graded as mild, moderate, or severe.

Classification by Stage The stage of chronic hepatitis, which reflects the level of progression of the disease, is based on the degree of fibrosis. When fibrosis is so extensive that fibrous septa surround parenchymal nodules and alter the normal architecture of the liver lobule, the histologic lesion is defined as cirrhosis. Staging is based on the degree of fibrosis as follows:

0= no fibrosis

1= mild fibrosis

2= moderate fibrosis

3= severe fibrosis, including bridging fibrosis

4 =cirrhosis

Reconciliation between Histologic Classification and New Classification For historical purposes, and to provide the basis for navigating several decades worth of literature on chronic hepatitis, the histologic categories of chronic persistent hepatitis, chronic lobular hepatitis, and chronic active hepatitis are worth reviewing and linking with their new-classification counterparts (Table 297-3).

In *chronic persistent hepatitis*, a mononuclear inflammatory infiltrate expands, but is localized to and contained within, portal tracts. The "limiting plate" of periportal hepatocytes is intact, and there is no extension of the necroinflammatory process into the liver lobule. A "cobblestone" arrangement of liver cells, indicative of hepatic regenerative activity, is a common feature, and although minimal periportal fibrosis may be present, *cirrhosis is absent.* As a general rule, patients with chronic persistent hepatitis are asymptomatic or have relatively mild constitutional symptoms (e.g., fatigue, anorexia, nausea); have normal physical findings, except perhaps for liver enlargement, without the usual stigmata of chronic liver disease (see below); and have modest elevations of aminotransferase activities. Progression to more severe lesions (chronic active hepatitis and cirrhosis) was felt to be very unlikely, especially in patients with autoimmune or idiopathic chronic persistent hepatitis; however, progressive disease occurs in patients with chronic persistent *viral* hepatitis and in those with chronic persistent hepatitis following spontaneous or therapeutic remission of autoimmune hepatitis. In the new nomenclature, chronic persistent hepatitis would be classified by *grade* as minimal or mild chronic hepatitis and by *stage* as absent or mild fibrosis.

In patients with *chronic lobular hepatitis*, in addition to portal inflammation, histologic examination of the liver reveals foci of necrosis and inflammation in the liver lobule. Morphologically, chronic lobular hepatitis resembles slowly resolving acute hepatitis. The limiting plate remains intact, periportal fibrosis is absent or limited, lobular architecture is preserved, and progression to chronic active hepatitis and cirrhosis was felt to be rare. Thus chronic lobular hepatitis can be considered a variant of chronic persistent hepatitis with a lobular component, and clinical/laboratory features are comparable. Occasionally, the clinical activity of chronic lobular hepatitis may increase spontaneously; elevation of aminotransferase activity may resemble that seen in acute hepatitis, and transient histologic deterioration can be documented. The same qualifications in prognostic import mentioned above for chronic persistent hepatitis apply to chronic lobular hepatitis. Chronic lobular hepatitis corresponds in the new nomenclature to a mild or moderate *grade* and a *stage* of absent or minimal fibrosis.

Chronic active hepatitis is characterized clinically by continuing hepatic necrosis, portal/periportal and, to a lesser extent, lobular inflammation, and fibrosis. Varying in severity from mild to severe, chronic active hepatitis was recognized to be a progressive disorder that can lead to cirrhosis, liver failure, and death. Morphologic characteristics of chronic active hepatitis include (1) a dense mononuclear infiltrate of the portal tracts, which are substantially expanded into the liver lobule (in the autoimmune type, plasma cells represent a component of the infiltrate); (2) destruction of the hepatocytes at the periphery of the lobule, with erosion of the limiting plate of hepatocytes surrounding the portal triads (piecemeal necrosis or interface hepatitis); (3) connective tissue septa surrounding portal tracts and extending from the portal zones into the lobule, isolating parenchymal cells into clusters and enveloping bile ducts; and (4) evidence of hepatocellular regeneration -- "rosette" formation, thickened liver cell plates, and regenerative "pseudolobules." This process may be patchy, with individual liver lobules spared, or it may be diffuse. Histologic evidence of single-cell coagulative necrosis, Councilman or acidophilic bodies, appear in the periportal areas. Piecemeal necrosis is the minimal histologic requirement to establish a diagnosis of chronic active hepatitis, but this change is seen even in mild, relatively nonprogressive forms of chronic active hepatitis. A more severe lesion, *bridging hepatic necrosis* (originally termed *subacute hepatic necrosis*), characterizes a more severe and progressive form of chronic active hepatitis. Although bridging necrosis can be seen occasionally in patients with acute hepatitis, in whom it carries no prognostic importance, in chronic active hepatitis this lesion is associated with progression to cirrhosis. Bridging necrosis is characterized by hepatocellular dropout that spans lobules (i.e., between portal tracts -- the periphery of the lobule -- or between portal tracts and central veins -- the centrizonal part of the lobule). Collapse of the reticulin network is a hallmark of bridging necrosis, and bridging fibrosis follows, leading ultimately to architectural reorganization by nodular regeneration, i.e., cirrhosis. A more extensive and ominous variant of bridging necrosis is multilobular collapse, in which bridging necrosis is widespread throughout the liver and which is associated clinically with rapid deterioration and even acute liver failure.

Although progression to cirrhosis is difficult to demonstrate in patients with chronic active hepatitis who have isolated piecemeal necrosis, in more severe forms of chronic active hepatitis, progression to cirrhosis is common. Among patients with chronic active hepatitis on liver biopsy, 20 to 50% also have cirrhosis, even early during the course of the disease. Ordinarily, chronic active hepatitis is more severe clinically than chronic persistent and lobular hepatitis. Although a sizable proportion of patients with chronic active hepatitis are asymptomatic, the majority tend to have mild to severe constitutional symptoms, especially fatigue. Generally, physical findings associated with chronic liver disease and portal hypertension are more common, aminotransferase levels tend to be higher, and jaundice and hyperbilirubinemia are more frequent in this form of chronic hepatitis.

In the new nomenclature for chronic hepatitis, what used to be called chronic active hepatitis spans the entire spectrum of activity *grade* from minimal, to mild, to severe chronic hepatitis, based on the degree of periportal and piecemeal necrosis, on the degree of lobular inflammation and injury, and on the degree of portal inflammation. Similarly, *stage* in chronic active hepatitis can translate to mild, moderate, or severe fibrosis as well as to cirrhosis.

CHRONIC VIRAL HEPATITIS

Both the enterically transmitted forms of viral hepatitis, hepatitis A and E, are self-limited and do not cause chronic hepatitis (rare reports notwithstanding in which acute hepatitis A serves as a trigger for the onset of autoimmune hepatitis in genetically susceptible patients). In contrast, the entire clinicopathologic spectrum of chronic hepatitis occurs in patients with chronic viral hepatitis B and C as well as in patients with chronic hepatitis D superimposed on chronic hepatitis B.

Chronic Hepatitis B The likelihood of chronicity after acute hepatitis B varies as a function of age. Infection at birth is associated with a clinically silent acute infection but a 90% chance of chronic infection, while infection in young adulthood in immunocompetent persons is typically associated with clinically apparent acute hepatitis but a risk of chronicity of only approximately 1%. Most cases of chronic hepatitis B among adults, however, occur in patients who never had a recognized episode of clinically apparent acute viral hepatitis. The degree of liver injury (grade) in patients with chronic hepatitis B is variable, ranging from none in asymptomatic carriers, to mild, to severe. Among adults with chronic hepatitis B, histologic features are of prognostic importance. In one long-term study of patients with chronic hepatitis B, investigators found a 5-year survival of 97% for patients with chronic persistent hepatitis (mild chronic hepatitis), of 86% for patients with chronic active hepatitis (moderate to severe chronic hepatitis), and of only 55% for patients with chronic active hepatitis and postnecrotic cirrhosis. The 15-year survival in these cohorts were 77, 66, and 40%, respectively. On the other hand, more recent observations do not allow us to be so sanguine about the prognosis in patients with mild chronic hepatitis; among patients with what used to be labeled chronic persistent hepatitis followed for 1 to 13 years, progression to more severe chronic hepatitis and cirrhosis has been observed in more than a quarter of cases.

Probably more important to consider than histology alone in patients with chronic hepatitis B is the degree of hepatitis B virus (HBV) replication. As reviewed inChap. 295, chronic hepatitis B can be divided into two phases based on the relative level of HBV replication. The relatively *replicative phase* is characterized by the presence in the serum of markers of HBV replication [hepatitis B e antigen (HBeAg) HBV DNA], by the presence in the liver of detectable intrahepatocyte nucleocapsid antigens [primarily hepatitis B core antigen (HBcAg)], by high infectivity, and by accompanying liver injury; HBV DNA can be detected in the liver but is extrachromosomal. In contrast, the relatively *nonreplicative phase* is characterized by the absence of conventional markers of HBV replication (HBeAg and HBV DNA detectable by hybridization) but an association with anti-HBe, the absence of intrahepatocytic HBcAg, limited infectivity, and minimal liver injury; HBV DNA can be detected in the liver but is integrated into the host genome. Those in the replicative phase tend to have more severe chronic hepatitis, while those in the nonreplicative phase tend to have minimal or mild chronic hepatitis or to be asymptomatic hepatitis B carriers; however, distinctions in HBV replication and in histologic category do not always coincide. The likelihood of converting spontaneously from relatively replicative to nonreplicative chronic HBV infection is approximately 10 to 15% per year. As noted in Chap. 295, the conversion from replicative to nonreplicative chronic hepatitis B is associated with a transient elevation in aminotransferase activity resembling acute hepatitis; occasionally, spontaneous resumptions of replicative activity occur in nonreplicative infection; and occasionally, HBV variants occur in which serologic markers of replication (HBeAg) are absent, despite the presence of replicative infection. Chronic HBV infection, especially when acquired at birth or in early childhood, is associated with an increased risk of hepatocellular carcinoma (Chap. 91).**A discussion of the pathogenesis of liver injury in patients with chronic hepatitis B appears in Chap. 295.*

The spectrum of *clinical features* of chronic hepatitis B is broad, ranging from asymptomatic infection to debilitating disease or even end-stage, fatal hepatic failure. As noted above, the onset of the disease tends to be insidious in most patients, with the exception of the very few in whom chronic disease follows failure of resolution of clinically apparent acute hepatitis B. The clinical and laboratory features associated with progression from acute to chronic hepatitis B are discussed in Chap. 295. *Fatigue* is a common symptom, and persistent or intermittent *jaundice* is a common feature in severe or advanced cases. Intermittent deepening of jaundice and recurrence of malaise and anorexia, as well as worsening fatigue, are reminiscent of acute hepatitis; such exacerbations may occur spontaneously, often coinciding with evidence of virologic reactivation, may lead to progressive liver injury, and, when superimposed on well-established cirrhosis, may cause hepatic decompensation. Complications of cirrhosis occur in end-stage chronic hepatitis and include ascites, edema, bleeding gastroesophageal varices, hepatic encephalopathy, coagulopathy, or hypersplenism. Occasionally, these complications bring the patient to initial clinical attention. Extrahepatic complications of chronic hepatitis B, similar to those seen during the prodromal phase of acute hepatitis B, are associated with deposition of circulating hepatitis B antigen-antibody immune complexes. These include arthralgias and arthritis, which are common, and the more rare purpuric cutaneous lesions (leukocytoclastic vasculitis), immune-complex glomerulonephritis, and generalized vasculitis (polyarteritis nodosa) (Chaps. 295 and317).

Laboratory features of chronic hepatitis B do not distinguish adequately between histologically mild and severe hepatitis. Aminotransferase elevations tend to be modest for chronic hepatitis B but may fluctuate in the range of 100 to 1000 units. As is true for acute viral hepatitis B, alanine aminotransferase (ALT, or SGPT) tends to be more elevated than aspartate aminotransferase (AST, or SGOT); however, once cirrhosis is established, AST tends to exceed ALT. Levels of alkaline phosphatase activity tend to be normal or only marginally elevated. In severe cases, moderate elevations in serum bilirubin [51.3 to 171 umol/L (3 to 10 mg/dL)] occur. Hypoalbuminemia and prolongation of the prothrombin time occur in severe or end-stage cases. Hyperglobulinemia and detectable circulating autoantibodies are distinctly absent in chronic hepatitis B (in contrast to autoimmune hepatitis).**Viral markers of chronic HBV infection are discussed in Chap. 295*.

TREATMENT

Management of chronic hepatitis B depends on the level of virus replication. Although progression to cirrhosis is more likely in severe chronic than in mild or moderate chronic hepatitis B, all forms of chronic viral hepatitis can be progressive. Interferon a(IFN-a) was the first approved therapy for chronic hepatitis B, but the recently approved dideoxynucleoside lamivudine expands the options for treatment. The most common indication for treatment is chronic "replicative" hepatitis B, with detectableHBeAgand HBVDNA (by hybridization assay), elevated ALT activity, and histologic evidence of chronic hepatitis on liver biopsy in an immunocompetent adult. A 16-week course of INF-a given by subcutaneous injection at a daily dose of 5 million units, or three times a week at a dose of 10 million units, results in seroconversion from "replicative" (detectable HBeAg and HBV DNA) to "nonreplicative" (undetectable HBeAg and HBV DNA by hybridization assay) HBV infection in approximately 35% of patients, with a concomitant improvement in liver histologic features. As a result of INF-atherapy, approximately 20% of patients acquire anti-HBe, and in early trials, approximately 8% lost hepatitis B surface antigen (HBsAg). Successful interferon therapy and seroconversion is often accompanied by an acute hepatitis-like elevation in aminotransferase activity, which has been postulated to result from enhanced cytolytic T cell clearance of HBV-infected hepatocytes. Relapse after successful therapy is rare (1 or 2%). The likelihood of responding to interferon is higher in patients with lower levels of HBV DNA and substantial elevations ofALT. Although children can respond as well as adults, interferon therapy has not been effective in very young children infected at birth. Similarly, interferon therapy has not been effective in immunosuppressed persons, Asian patients with minimal-to-mild ALT elevations, patients with pre-core mutant HBV infection (Chap. 295), or in patients with decompensated chronic hepatitis B (in whom such therapy can actually be detrimental, sometimes precipitating decompensation, often associated with severe adverse effects). Among patients with HBeAg loss during therapy, long-term follow-up has demonstrated that 80% experience eventual loss of HBsAg, i.e., all serologic markers of infection, and normalization of ALT over a 9-year posttreatment period. In addition, improved long-term and complication-free survival as well as a reduction in the frequency of hepatocellular carcinoma have been documented among interferon responders, supporting the conclusion that successful interferon therapy improves the natural history of chronic hepatitis B. Indications for interferon therapy in patients with chronic hepatitis B are summarized in Table 297-4.

Complications of interferon therapy include systemic "flulike" symptoms, marrow suppression, emotional lability (irritability commonly, depression rarely), autoimmune reactions (especially autoimmune thyroiditis), and miscellaneous side effects such as alopecia, rashes, diarrhea, and numbness and tingling of the extremities. With the possible exception of autoimmune thyroiditis, all these side effects are reversible upon dose lowering or cessation of therapy.

In patients with chronic hepatitis B, long-term therapy with glucocorticoids is not only ineffective but also detrimental. Short-term glucocorticoid therapy, however, has been advocated as a potential antiviral approach. Glucocorticoids increaseHBVreplication and expression in hepatocytes and depress cytolytic T cells. When glucocorticoids are administered for a brief time and then withdrawn abruptly, cytolytic T cells, suppressed while HBV replication was enhanced by the drug, resume their presteroid function. These restored cytolytic T cells attack hepatocytes, the HBV expression of which had been enhanced by the brief pulse of glucocorticoid therapy. An acute hepatitis-like flare of aminotransferase activity follows and may be accompanied by a dramatic drop, or even loss of, HBV replication. Such glucocorticoid "priming" prior to interferon therapy has not been shown to be more effective than interferon alone and has been abandoned.

Several nucleoside analogues active againstHBV are being evaluated and developed. Famciclovir and ganciclovir have only limited activity against hepatitis B; however, lamivudine, which inhibits reverse transcriptase activity of both HIV and HBV, is a potent and effective agent for patients with chronic hepatitis B. Lamivudine suppresses HBV DNA by a median of four orders of magnitude at oral daily doses of 100 mg. In clinical trials conducted in Asia, North America, Europe, and Australia, lamivudine therapy for 12 months was associated with almost universal suppression of HBV DNA detectable by hybridization assays; loss ofHBeAg in 32 to 33%; HBeAg seroconversion (i.e., conversion from HBeAg-reactive to anti-HBe-reactive) in 16 to 20%; normalization ofALT in approximately 40%; improvement in histology in over 50%; and retardation in fibrosis in 20%. Among patients who experienced HBeAg responses during therapy, 70 to 80% maintained the response over longer than a year of follow-up monitoring. Because maintenance of the response to lamivudine occurs in almost all patients with an HBeAg response, the achievement of an HBeAg response may be a viable stopping point in therapy. If HBeAg is unaffected by lamivudine therapy, the current approach is to continue therapy until an HBeAg response occurs, but long-term therapy may be required to suppress HBV replication and, in turn, limit liver injury. Preliminary observations indicate that HBeAg seroconversions can increase to a level of 27% after 2 years and 44% after 3 years of therapy.

Losses ofHBsAghave been few during lamivudine therapy, and this observation had been cited as an advantage of interferon over lamivudine; however, in head-to-head comparisons between interferon and lamivudine monotherapy, HBsAg losses were rare in both groups. Trials in which lamivudine and interferon were administered in combination failed to show a benefit of combination therapy over lamivudine monotherapy for either treatment-naive patients or prior interferon nonresponders.

Among patients withHBeAgand HBVDNA but with normalALTactivity, lamivudine suppresses liver injury during therapy but rarely achieves an HBeAg response. In patients with pre-core HBV mutations, who lack HBeAg but who have detectable HBV DNA and liver injury, lamivudine suppresses HBV DNA and normalizes ALT in 65% and improves liver histology in 60%. When therapy is discontinued, reactivation is common, and these patients require long-term therapy.

Clinical and laboratory side effects of lamivudine are negligible, indistinguishable from those observed in placebo recipients. During lamivudine therapy, transientALTelevations, resembling those seen during interferon therapy and during spontaneousHBeAg-to-anti-HBe seroconversions, occur in a quarter of patients. These ALT elevations may result from restored cytolytic T cell activation permitted by suppression of HBV replication. Similar ALT elevations, however, occur at an identical frequency in placebo recipients, but ALT elevations associated with HBeAg seroconversion are confined to lamivudine-treated patients. When therapy is stopped after a year of therapy, 2- to 3-fold ALT elevations occur in 20 to 30% of lamivudine-treated patients, representing renewed liver-cell injury as HBV replication returns. Although these posttreatment flares are almost always transient and mild, rare severe exacerbations have been observed, mandating close and careful clinical and virologic monitoring after discontinuation of treatment.

Long-term monotherapy with lamivudine is associated with methionine-to-valine or methionine-to-isoleucine mutations in the YMDD (tyrosine-methionine-aspartate-aspartate) motif ofHBV DNA polymerase, analogous to mutations that occur in patients with HIV infection treated with this drug. During a year of therapy, YMDD mutations occur in 15 to 30% of patients; the frequency increases at year two to 38% and at year three to almost 50%. Although transient elevations inALT and HBV DNA levels occur when such variants emerge, YMDD-variant HBV appears to be less replicatively competent and a less robust pathogen. Even after YMDD mutations occur, HBV DNA and ALT levels as well as histologic scores tend to remain lower than baseline levels in immunocompetent patients. In immunosuppressed patients, a proportion of patients with YMDD mutations experience hepatic decompensation. Until other antivirals are developed, the approach to YMDD variants emerging during lamivudine treatment is to continue therapy. Other antiviral drugs, such as the experimental agent adefovir dipivoxil, inhibit YMDD-variant HBV. In the future, combination antiviral therapy will almost invariably become the norm as new agents are introduced.

Because lamivudine monotherapy can result universally in the rapid emergence ofYMDDvariants in persons with HIV infection, patients with chronic hepatitis B should be tested for anti-HIV prior to therapy; if HIV infection is identified, lamivudine monotherapy at theHBVdaily dose of 100 mg is contraindicated. These patients should be treated with triple-drug antiretroviral therapy, including a lamivudine daily dose of 300 mg (Chap. 309). The safety of lamivudine during pregnancy has not been established.

No treatment is indicated or available for asymptomatic "nonreplicative" hepatitis B carriers. Whereas patients with decompensated chronic hepatitis B are not candidates for interferon therapy, they may respond to lamivudine, with reversal of the signs of decompensation.

Table 297-4 summarizes the indications in patients with chronic hepatitis B for antiviral therapy with lamivudine, as compared with interferon. Both drugs are quite comparable in efficacy as first-line therapy for chronic hepatitis B (Table 297-5). Interferon requires only brief-duration therapy, too limited in duration to support viral variants, but requires subcutaneous injections and is associated with a high level of intolerability. Lamivudine requires long-term therapy in most patients and, when used alone, fosters the emergence of viral variants. On the other hand, lamivudine is taken orally, is very well tolerated, leads to improved histology even in the absence of HBeAgresponses, and is effective even in patients who fail to respond to interferon. Although some prefer to begin with interferon, most physicians and patients prefer lamivudine as first-line therapy.

For patients with end-stage chronic hepatitis B, liver transplantation is the only potential lifesaving intervention. Reinfection of the new liver is almost universal; however, the likelihood of liver injury associated with hepatitis B in the new liver is variable. The majority of patients become high-level viremic carriers with minimal liver injury. Unfortunately, an unpredictable proportion experience severe hepatitis B-related liver injury, sometimes a fulminant-like hepatitis, sometimes a rapid recapitulation of the original severe chronic hepatitis B (Chap. 301). Prevention of recurrent hepatitis B after liver transplantation has been achieved by *prophylaxis* with hepatitis B immune globulin and with nucleoside analogues such as lamivudine; in addition, nucleoside analogues have been used successfully to *reverse* posttransplantation liver injury associated with recurrent hepatitis B (Chap. 301).

Chronic Hepatitis D (Delta Hepatitis) The clinical and laboratory features of chronic hepatitis D virus (HDV) infection are summarized inChap. 295. Chronic hepatitis D may follow acute coinfection with HBV but at a rate no higher than the rate of chronicity of hepatitis B. That is, although HDV coinfection can increase the severity of acute hepatitis B,HDV does not increase the likelihood of progression to chronic hepatitis B. However, when HDV superinfection occurs in a person who is already chronically infected with HBV, long-term HDV infection is the rule and a worsening of the liver disease the expected consequence. Except for severity, chronic hepatitis B plus D has similar clinical and laboratory features to those seen in chronic hepatitis B alone. Relatively severe chronic hepatitis, with or without cirrhosis, is the rule, and mild chronic hepatitis the exception. A distinguishing serologic feature of chronic hepatitis D is the presence in the circulation of antibodies to liver-kidney microsomes (anti-LKM); however, the anti-LKM seen in hepatitis D are designated anti-LKM3, are directed against uridine diphosphate glucuronosyltransferase, and are distinct from anti-LKM1 seen in patients with autoimmune hepatitis and in a subset of patients with chronic hepatitis C (see below).

TREATMENT

Management is not well defined. Glucocorticoids are ineffective and are not used. Preliminary experimental trials of interferon asuggested that conventional doses and durations of therapy lower levels ofHDV RNA and aminotransferase activity only transiently during treatment but have no impact on the natural history of the disease. Although high-doseIFN-a(9 million units) three times a week for 12 months may be associated with a sustained loss of HDV replication and clinical improvement in up to 50% of patients, ultimately recurrent HDV replication becomes universal after cessation of therapy. Antiviral therapy for chronic hepatitis D remains the subject of experimental trials; early observations suggest that lamivudine is not effective. In patients with end-stage liver disease secondary to chronic hepatitis D, liver transplantation has been effective. If hepatitis D recurs in the new liver without the expression of hepatitis B (an unusual serologic profile in immunocompetent persons, but common in transplant patients), liver injury is limited. In fact, the outcome of transplantation for chronic hepatitis D is superior to that for chronic hepatitis B (Chap. 301).

Chronic Hepatitis C Regardless of the epidemiologic mode of acquisition of hepatitis C virus (HCV) infection, chronic hepatitis follows acute hepatitis C in 50 to 70% of cases; even in those with a return to normal in aminotransferase levels after acute hepatitis C, chronic infection is common, adding up to an 85 to 90% likelihood of chronic HCV infection after acute hepatitis C. Furthermore, in patients with chronic transfusion-associated hepatitis followed for 10 to 20 years, progression to cirrhosis occurs in about 20%. Such is the case even for patients with relatively clinically mild chronic hepatitis, including those without symptoms, with only modest elevations of aminotransferase activity, and with mild chronic hepatitis on liver biopsy. Even in cohorts of well-compensated patients with chronic hepatitis C (no complications of chronic liver disease and with normal hepatic synthetic function), the prevalence of cirrhosis may be as high as 50%. Many cases of hepatitis C are identified in asymptomatic patients who have no history of acute hepatitis C, e.g., those discovered while attempting to donate blood or as a result of routine laboratory screening tests. The source of HCV infection in most of these cases is not defined, although a long-forgotten percutaneous exposure in the remote past can be elicited in a substantial proportion. The natural history of chronic hepatitis C identified under these circumstances remains to be determined. Among asymptomatic persons with anti-HCV, even when aminotransferase levels are normal, between a third and a half have been reported to have chronic hepatitis on liver biopsy, although mild in most cases. In these asymptomatic persons with normal aminotransferase levels, the presence of detectable circulatingHCV RNA appears to distinguish those with chronic hepatitis on biopsy from those with normal liver histology.

Despite this substantial rate of progression of chronic hepatitis C, and despite the fact that liver failure can result from end-stage chronic hepatitis C, the long-term prognosis for chronic hepatitis C in a majority of patients is relatively benign. Mortality over 10 to 20 years among patients with transfusion-associated chronic hepatitis C has been shown not to differ from mortality in a matched population of transfused patients in whom hepatitis C did not develop. Although death in the hepatitis group is more likely to result from liver failure, and although hepatic decompensation may occur in approximately 15% of such patients over the course of a decade, the majority (almost 60%) of patients remain asymptomatic and well compensated, with no clinical sequelae of chronic liver disease. Overall, then, chronic hepatitis C tends to be very slowly and insidiously progressive, if at all, in the vast majority of patients, while in approximately a quarter of cases, chronic hepatitis C will progress eventually to end-stage cirrhosis. Referral bias may account for the more severe outcomes described in cohorts of patients reported from tertiary-care centers versus the more benign outcomes in cohorts of patients monitored from initial blood-product-associated acute hepatitis. Still unexplained, however, are the wide ranges in reported progression to cirrhosis, from 2% over 17 years in a population of women with hepatitis C infection acquired from contaminated anti-D immune globulin to 30% over £11 years in recipients of contaminated intravenous immune globulin.

Progression of liver disease in patients with chronic hepatitis C has been reported to be more likely in patients with older age, longer duration of infection, advanced histologic stage and grade, genotype 1 (especially type 1b), more complex quasispecies diversity, and increased hepatic iron. Among these variables, however, duration of infection appears to be the most important, and many of the others probably reflect disease duration to some extent (e.g., quasispecies diversity, hepatic iron accumulation).

Perhaps the best prognostic indicator in chronic hepatitis C is liver histology. Patients with mild necrosis and inflammation as well as those with limited fibrosis have an excellent prognosis and limited progression to cirrhosis. In contrast, among patients with moderate to severe necroinflammatory activity or fibrosis, including septal or bridging fibrosis, progression to cirrhosis is highly likely over the course of 10 to 20 years. Among patients with compensated cirrhosis associated with hepatitis C, the 10-year survival is close to 80 percent; mortality occurs at a rate of 2 to 6% per year, decompensation at a rate of 4 to 5% per year, and hepatocellular carcinoma at a rate of 1 to 3% per year.

In addition, severity of chronic hepatitis is greater and progression of chronic liver disease is more accelerated in patients who have chronic hepatitis C as well as other liver processes, including alcoholic liver disease, chronic hepatitis B, hemochromatosis, anda$_1$-antitrypsin deficiency. No other epidemiologic or clinical features of chronic hepatitis C (e.g., severity of acute hepatitis, level of aminotransferase activity, level ofHCV RNA, presence or absence of jaundice) are predictive of eventual outcome. Despite the relative benignity of chronic hepatitis C over time, cirrhosis following chronic hepatitis C has been associated with the late development, after several decades, of hepatocellular carcinoma (HCC) (Chap. 91). As noted above, the annual rate of HCC in cirrhotic patients with hepatitis C is 1 to 3%.

Clinical features of chronic hepatitis C are similar to those described above for chronic hepatitis B. Generally, *fatigue* is the most common symptom; jaundice is rare. Immune-complex mediated extrahepatic complications of chronic hepatitis C are less common than in chronic hepatitis B, with the exception of essential mixed cryoglobulinemia (Chap. 295). This is the case despite the fact that assays for immune-complex-like activity are often positive in patients with chronic hepatitis C. In addition, chronic hepatitis C has been associated with extrahepatic complications unrelated to immune-complex injury. These include Sjogren's syndrome, lichen planus, and porphyria cutanea tarda. *Laboratory features* of chronic hepatitis C are similar to those in patients with chronic hepatitis B, but aminotransferase levels tend to fluctuate more (the characteristic episodic pattern of aminotransferase activity) and to be lower, especially in patients with long-standing disease. An interesting and occasionally confusing finding in patients with chronic hepatitis C is the presence of autoantibodies. Rarely, patients with autoimmune hepatitis (see below) and hyperglobulinemia have false-positive enzyme immunoassays for anti-HCV. On the other hand, some patients with serologically confirmable chronic hepatitis C have circulatinganti-LKM. These antibodies are anti-LKM1, as seen in patients with autoimmune hepatitis *type 2* (see below), and are directed against a 33-amino-acid sequence of P450 IID6. The occurrence of anti-LKM1 in some patients with chronic hepatitis C may result from the partial sequence homology between the epitope recognized by anti-LKM1 and two segments of the HCV polyprotein. In addition, the presence of this autoantibody in some patients with chronic hepatitis C suggests that autoimmunity may be playing a role in the pathogenesis of chronic hepatitis C.**Histopathologic features of chronic hepatitis C, especially those that distinguish hepatitis C from hepatitis B, are described in Chap. 295.*

TREATMENT

Two approaches to antiviral therapy of chronic hepatitis C have been approved: *monotherapy* with interferon and *combination therapy* with interferon plus ribavirin. According to a National Institutes of Health Consensus Development Conference in March 1997, responses measured at the end of treatment are referred to as end-treatment responses, and responses sustained for at least 6 months after discontinuation of therapy are referred to as sustained responses.

Interferon Monotherapy Interferon a, administered by subcutaneous injection three times a week for 6 months yields end-treatment biochemical responses (return to normal ofALTlevels) as high as approximately 50% and virologic responses (undetectableHCV RNA) by polymerase chain reaction (PCR) of approximately 30%. Unfortunately, because of a relapse rate as high as 90% in end-treatment responders, these responses are not maintained after discontinuation of therapy except in a small minority of patients; after 6 months of interferon monotherapy, the likelihood of a sustained biochemical and virologic response is only approximately 10%. Even in the absence of a biochemical/virologic response, however, end-treatment histologic responses -- primarily reductions in periportal and lobular activity -- occur in three-fourths of treated patients. Unlike the case in hepatitis B, in chronic hepatitis C successful responses to therapy are not accompanied by transient, acute-hepatitis-like elevations in aminotransferase activity; instead, ALT levels fall precipitously. Between 85 to 90% of responses occur within the first 3 months of therapy; responses thereafter are rare.

In a proportion of cases, markers ofHCVreplication can be eradicated by interferon therapy, and durable responses with normalALT, improved histology, and absence of HCV RNA in serum and liver have been documented many years after successful therapy. A small proportion of patients, approximately 10%, experience biochemical "breakthrough" *during* interferon therapy and are classified as nonresponders. In general, they remain refractory to retreatment thereafter; some such breakthroughs are associated with interferon antibodies, while others may reflect mutations in the HCV genome that render HCV nonresponsive to interferon.

Levels ofHCV RNA fall in tandem withALTlevels during interferon therapy, but loss of detectable HCV RNA does not preclude relapse. When a patient experiences an apparently sustained biochemical response after discontinuing interferon but continues to remain viremic, as reflected by the persistence of detectable HCV RNA, future biochemical relapse is likely. Patient variables that tend to correlate with *sustained* responsiveness to interferon include a low baseline level of HCV RNA and histologically mild hepatitis. Patients with cirrhosis can respond, but they are less likely to do so and especially unlikely to have a *sustained* response. Patients with HCV genotype 1 are less likely to respond than patients with other genotypes. Other variables reported to correlate with increased responsiveness include brief duration of infection, low HCV quasispecies diversity, immunocompetence, and low liver iron levels. High levels of HCV RNA, more histologically advanced liver disease, and high quasispecies diversity all go hand in hand with advanced duration of infection, which may be the single most important variable determining interferon responsiveness. The ironic fact, then, is that patients whose disease is *least* likely to progress are the ones *most* likely to respond to interferon and vice versa. Finally, among patients with genotype 1b, responsiveness to interferon is enhanced in those with amino-acid-substitution mutations in the nonstructural protein 5A gene.

The most effective approach to increasing responsiveness to interferon monotherapy is to increase the duration of therapy to 12 months or longer, a regimen associated with a sustained biochemical and virologic response of approximately 20%. Higher doses of interferon (e.g., 5 to 10 million units) or daily injections increase response rates only marginally and at a substantial cost in intolerability. Thus, if interferon monotherapy is selected, the consensus is that 3 million units for at least 12 months is the preferred regimen. Currently, three types ofIFN-aare approved in the United States; for the two recombinant products, the recommended dose is 3 million units, and for the one synthetic consensus interferon (synthesized to represent the amino acids at each position that occur most frequently among the multiple, natural interferona subspecies), the dose is 9ug. Several other types of INF-a, including lymphoblastoid interferon, are available in Europe and Asia. A review of the different types of INF-a during the NIH Consensus Development Conference in 1997 led to the conclusion that they are all equivalent in efficacy.

Studies of viral kinetics have shown that despite a virion half life in serum of only 2 to 3 h, the level ofHCV is maintained by a high replication rate of 10_{12}hepatitis C virions per day. Interferon a blocks virion production or release with an efficacy that increases with increasing drug doses; moreover, the calculated death rate for infected cells during interferon therapy is inversely related to viral load; patients with the most rapid death rate of infected hepatocytes are more likely to achieve undetectable HCV RNA at 3 months; achieving this landmark is predictive of a subsequent sustained response. Therefore, to achieve rapid viral clearance from serum and the liver, *high-dose induction therapy* has been advocated. In practice, high-dose induction therapy has not yielded higher sustained response rates. Other approaches that have been suggested include tapering therapy slowly, rather than discontinuing therapy abruptly, and, because high liver iron levels are associated with nonresponsiveness, the addition of phlebotomy to interferon therapy. None of these approaches has been shown to be effective.

Long-acting interferons bound to polyethylene glycol (PEG) have several advantages. Such "pegylated" interferons, with elimination times seven-fold longer than standard interferons, achieve prolonged concentration peaks and can be administered once, rather than three times, a week. Instead of the frequent drug peaks and troughs associated with frequent administration of short-acting interferons, administration of pegylated interferons results in drug concentrations that are more stable and sustained over time. Preliminary studies suggest that once-a-week injections of pegylated interferons are at least as effective as standard interferons given three times a week and may result in sustained responses comparable to those achieved with combination interferon-ribavirin therapy (see below).

If a patient relapses after a course of interferon monotherapy, repeating a course of interferon monotherapy is unlikely to achieve a sustained response unless the dose or preferably the duration of therapy is increased. Under these circumstances, sustained response rates as high as 40% can be realized. Although a small proportion of interferon nonresponders can respond to a repeat course of interferon monotherapy, and although a 13% sustained response rate has been reported for prior interferon nonresponders treated with high-dose (15 ug) consensus interferon, the likelihood of responding is not increased substantially by retreating interferon nonresponders with interferon monotherapy.

Combination Interferon-Ribavirin Therapy The most effective way to increase the efficacy of interferon therapy is to add ribavirin, an oral guanoside nucleoside. When used as monotherapy, ribavirin is ineffective and does not reduceHCV RNA levels. In contrast, the combination of interferon at standard doses with ribavirin at doses of 1000 mg (for patients weighing<75 kg) to 1200 mg (for patients weighing³75 kg) per day increases both end-treatment responses and sustained responses in previously untreated patients. Large, international, multicenter trials have shown that end-treatment responses at 6 months or 12 months exceed 50% and sustained responses as high as 33% at 6 months and 41% at 12 months have been achieved. Thus, a full year of combination therapy is twice as effective as a year of interferon monotherapy. Sustained responses were more likely in patients with low viral loads (below 2 million copies/mL), genotypes other than 1, minimal fibrosis, age<40, and females. In patients with low viral loads and non-1 genotypes, sustained response rates can be as high as 95%, and combination therapy for 24 weeks suffices, achieving the same end as continuing therapy for a full year. Therefore, for patients with low viral loads and non-1 genotypes, therapy need last only 6 months. Unless contraindications to the use of combination therapy exist (see below), combination interferon-ribavirin is the treatment of choice for chronic hepatitis C (Table 297-6).

For those who relapse after a 6-month course of interferon monotherapy, a 6-month course of combination therapy results in a sustained response rate of 50%, and retreatment of relapsers is another approved indication for combination therapy. Unfortunately, combination therapy has been disappointing in interferon nonresponders.

Side effects of combination therapy are similar to those of interferon monotherapy; however, ribavirin causes hemolysis; a reduction in hemoglobin of up to 2 to 3 gm or in hematocrit of up 5 to 10% can be anticipated. A small, unpredictable proportion of patients will experience profound, brisk hemolysis, resulting in symptomatic anemia. Therefore, close monitoring of blood counts is crucial, and combination therapy should be avoided in patients with anemia or hemoglobinopathies and in patients with coronary artery disease or cerebrovascular disease, in whom anemia can precipitate an ischemic event. Ribavirin, which is renally excreted, should not be used by patients with renal insufficiency; the drug is teratogenic, precluding its use during pregnancy and mandating the use of efficient contraception during therapy.

Ribavirin therapy has also been characterized by nasal congestion, pruritus, and precipitation of gout; the combination is more difficult to tolerate than interferon monotherapy. In one large clinical trial of combination therapy versus monotherapy among patients treated for a year, 21% of the combination group (but only 14% of the monotherapy group) had to discontinue treatment, while 26% of the combination group (but only 9% of the monotherapy group) required dose reductions.

Indications for Antiviral Therapy Patients with chronic hepatitis C who have elevatedALTlevels, detectableHCV RNA, and chronic hepatitis of at least moderate grade and stage are candidates for antiviral therapy with interferon and ribavirin, unless ribavirin is contraindicated (Table 297-6). Preliminary retrospective analyses have shown that interferon treatment improves survival and complication-free survival. One year of combination therapy is standard, but 6 months suffice for patients with non-1 genotypes and low viral loads. For patients treated with interferon monotherapy, 12 months is the standard duration in all cases, regardless of genotype and viral load. According to the NIH Consensus Development Conference in 1997, therapy should be discontinued in patients who have not achieved a normal ALT and an undetectable HCV RNA by month three. Although the vast majority of patients treated with combination therapy who become sustained responders will have achieved an early biochemical and virologic response, a proportion of sustained responders experienced late viral clearance. In addition, even in biochemical and virologic nonresponders, histologic improvement is common. Therefore, recommendations for early cessation of therapy based on interim assessments of biochemical and virologic responsiveness require reevaluation. Although response rates are lower in patients with certain pretreatment variables, selection for treatment should not be based on symptoms, genotype, viral load, or the mode of acquisition of infection.

Patients who have relapsed after an initial course of interferon monotherapy are candidates for a 6-month course of combination interferon-ribavirin therapy; if they cannot tolerate ribavirin, they should be retreated with interferon monotherapy, but the course should be longer. It remains to be determined whether long-term (even indefinite) maintenance therapy will be necessary or effective in patients who relapse repeatedly whenever therapy is discontinued. For interferon nonresponders, retreatment with interferon monotherapy or combination therapy is unlikely to achieve a sustained response. Clinical trials are in progress to determine whether long-term suppression of virus-induced liver injury with antiviral therapy will be of benefit in this population.

In patients with acute hepatitis C, a course of interferon has been shown to reduce the likelihood of chronicity by one-half (Chapter 295). In patients with normalALTlevels, long-term monitoring studies have shown absence of histologic progression, and clinical trials of antiviral therapy have shown no benefit; therefore, treatment of such patients is not recommended. Because hepatitis C can reactivate in patients with normal ALT levels, laboratory monitoring several times a year should be done, and therapy should be considered for sustained elevations in ALT levels. Patients with mild hepatitis on liver biopsy are not routine candidates for antiviral therapy, but treatment decisions should be individualized between physician and patient. Most authorities would recommend a pretreatment liver biopsy to help in the decision-making about therapy.

Patients with compensated cirrhosis can respond to therapy, although their likelihood of a sustained response is lower than in noncirrhotics. Combination therapy brings sustained response rates in cirrhotics up to the level achieved with interferon monotherapy in noncirrhotics. Retrospective analyses generally have not demonstrated an improvement in survival among interferon-treated cirrhotic patients. Similarly, several studies have suggested that treatment of cirrhotics with hepatitis C reduces the frequency ofHCC; however, logistic regression analyses have shown that patient characteristics at the time of therapy (e.g., less advanced disease), not treatment itself, accounted for the reduced frequency of HCC observed in the treated cohort. Patients with decompensated cirrhosis are not candidates for antiviral therapy but should be referred for liver transplantation. After liver transplantation, recurrent hepatitis C is the rule. Most patients who undergo liver transplantation for chronic hepatitis C experience little, if any, morbidity, allograft loss, or mortality associated with recurrent hepatitis C during the early postoperative years (Chapter 301); studies are in progress to determine how best to treat hepatitis C after liver transplantation. The cutaneous and renal vasculitis ofHCV-associated essential mixed cryoglobulinemia (Chap. 295) may respond to interferon, but sustained responses are rare after discontinuation of therapy; therefore, prolonged, perhaps indefinite, therapy is recommended in this group.

Anecdotal reports suggest that antiviral therapy may be effective in porphyria cutanea tarda or lichen planus associated with hepatitis C. In patients with HIV infection, responses similar to those seen in other groups have been reported in patients with normal CD4 counts.

AUTOIMMUNE HEPATITIS

Definition Autoimmune hepatitis (formerly called autoimmune chronic active hepatitis) is a chronic disorder characterized by continuing hepatocellular necrosis and inflammation, usually with fibrosis, which tends to progress to cirrhosis and liver failure. When fulfilling criteria of severity, this type of chronic hepatitis may have a 6-month mortality of as high as 40%. The prominence of extrahepatic features of autoimmunity as well as seroimmunologic abnormalities in this disorder supports an autoimmune process in its pathogenesis; this concept is reflected in the labels "lupoid," plasma cell, or autoimmune hepatitis. Because autoantibodies and other typical features of autoimmunity do not occur in all cases, however, a broader, more appropriate designation for this type of chronic hepatitis is "idiopathic" or cryptogenic. Cases in which hepatotropic viruses, metabolic/genetic derangements, and hepatotoxic drugs have been excluded merit this designation and probably include a spectrum of heterogeneous liver disorders of unknown cause, a proportion of which have characteristic autoimmune features.

Immunopathogenesis The weight of evidence suggests that the progressive liver injury in patients with idiopathic/autoimmune hepatitis is the result of a cell-mediated immunologic attack directed against liver cells; in all likelihood, predisposition to autoimmunity is inherited, while the liver specificity of this injury is triggered by environmental (e.g., chemical or viral) factors. For example, patients have been described in whom apparently self-limited cases of acute hepatitis A or B led to autoimmune hepatitis, presumably because of genetic susceptibility or predisposition. Evidence to support an autoimmune pathogenesis in this type of hepatitis includes the following: (1) In the liver, the histopathologic lesions are composed predominantly of cytotoxic T cells and plasma cells; (2) circulating autoantibodies (nuclear, smooth muscle, thyroid, etc.; see below), rheumatoid factor, and hyperglobulinemia are common; (3) other autoimmune disorders -- such as thyroiditis, rheumatoid arthritis, autoimmune hemolytic anemia, ulcerative colitis, proliferative glomerulonephritis, juvenile diabetes mellitus, and Sjogren's syndrome -- occur with increased frequency in patients who have autoimmune hepatitis and in their relatives; (4) histocompatibility haplotypes associated with autoimmune diseases, such as HLA-B1, -B8, -DR3, and -DR4, are common in patients with autoimmune hepatitis; and (5) this type of chronic hepatitis is responsive to glucocorticoid/immunosuppressive therapy, effective in a variety of autoimmune disorders.

Cellular immune mechanisms appear to be important in the pathogenesis of autoimmune hepatitis. In vitro studies have suggested that in patients with this disorder, lymphocytes are capable of becoming sensitized to hepatocyte membrane proteins and of destroying liver cells. Abnormalities of immunoregulatory control over cytotoxic lymphocytes (impaired suppressor cell influences) may play a role as well. Studies of genetic predisposition to autoimmune hepatitis demonstrate that certain haplotypes are associated with the disorder, as enumerated above. The precise triggering factors, genetic influences, and cytotoxic and immunoregulatory mechanisms involved in this type of liver injury remain poorly defined.

Intriguing clues into the pathogenesis of autoimmune hepatitis come from the observation that circulating autoantibodies are prevalent in patients with this disorder. Among the autoantibodies described in these patients are antibodies to nuclei [so-called antinuclear antibodies (ANA), primarily in a homogeneous pattern] and smooth muscle (so-called anti-smooth-muscle antibodies, directed at actin),anti-LKM(see below), antibodies to "soluble liver antigen" (directed at a member of the glutathione S-transferase gene family), as well as antibodies to the liver-specific asialoglycoprotein receptor (or "hepatic lectin") and other hepatocyte membrane proteins. Although some of these provide helpful diagnostic markers, their involvement in the pathogenesis of autoimmune hepatitis has not been established.

Humoral immune mechanisms have been shown to play a role in the extrahepatic manifestations of autoimmune/idiopathic hepatitis. Arthralgias, arthritis, cutaneous vasculitis, and glomerulonephritis occurring in patients with autoimmune hepatitis appear to be mediated by the deposition in affected tissue vessels of circulating immune complexes, followed by complement activation, inflammation, and tissue injury. While specific viral antigen-antibody complexes can be identified in acute and chronic viral hepatitis, the nature of the immune complexes in autoimmune hepatitis has not been defined.

Many of the *clinical features* of autoimmune hepatitis are similar to those described for chronic viral hepatitis. The onset of disease may be insidious or abrupt; the disease may present initially like, and be confused with, acute viral hepatitis; a history of recurrent bouts of what had been labeled acute hepatitis is not uncommon. A subset of patients with autoimmune hepatitis has distinct features. Such patients are predominantly young to middle-aged women with marked hyperglobulinemia and high-titer circulatingANA. This is the group with positive LE preparations (initially labeled "lupoid" hepatitis) in whom other autoimmune features are common. Fatigue, malaise, anorexia, amenorrhea, acne, arthralgias, and jaundice are common. Occasionally, arthritis, maculopapular eruptions (including cutaneous vasculitis), erythema nodosum, colitis, pleurisy, pericarditis, anemia, azotemia, and sicca syndrome (keratoconjunctivitis, xerostomia) occur. In some patients, complications of cirrhosis, such as ascites and edema (associated with hypoalbuminemia), encephalopathy, hypersplenism, coagulopathy, or variceal bleeding may bring the patient to initial medical attention.

The course of autoimmune hepatitis may be variable. In those with mild disease or limited histologic lesions (e.g., piecemeal necrosis without bridging), progression to cirrhosis is limited. In those with severe symptomatic autoimmune hepatitis (aminotransferase levels >10 times normal, marked hyperglobulinemia, "aggressive" histologic lesions -- bridging necrosis or multilobular collapse, cirrhosis), the 6-month mortality without therapy may be as high as 40%. Such severe disease accounts for only 20% of cases; the natural history of milder disease is variable, often accentuated by spontaneous remissions and exacerbations. Especially poor prognostic signs include multilobular collapse at the time of initial presentation and failure of the bilirubin to improve after 2 weeks of therapy. Death may result from hepatic failure, hepatic coma, other complications of cirrhosis (e.g., variceal hemorrhage), and intercurrent infection. In patients with established cirrhosis, hepatocellular carcinoma may be a late complication (Chap. 91).

Laboratory features of autoimmune hepatitis are similar to those seen in chronic viral hepatitis. Liver biochemical tests are invariably abnormal but may not correlate with the clinical severity or histopathologic features in individual cases. Many patients with autoimmune hepatitis have normal serum bilirubin, alkaline phosphatase, and globulin levels with only minimal aminotransferase elevations. SerumAST andALTlevels are increased and fluctuate in the range of 100 to 1000 units. In severe cases, the serum bilirubin level is moderately elevated [51 to 171 umol/L (3 to 10 mg/dL)]. Hypoalbuminemia occurs in patients with very active or advanced disease. Serum alkaline phosphatase levels may be moderately elevated or near normal. In a small proportion of patients, marked elevations of alkaline phosphatase activity occur; in such patients, clinical and laboratory features overlap with those of primary biliary cirrhosis (Chap. 299). The prothrombin time is often prolonged, particularly late in the disease or during active phases.

Hypergammaglobulinemia (>2.5 g/dL) is common in autoimmune hepatitis. Rheumatoid factor is common as well. As noted above, circulating autoantibodies are also common. The most characteristic areANA in a homogeneous staining pattern. Smooth-muscle antibodies are less specific, seen just as frequently in chronic viral hepatitis. Because of the high levels of globulins achieved in the circulation of some patients with autoimmune hepatitis, occasionally the globulins may bind nonspecifically in solid-phase binding immunoassays for viral antibodies. This has been recognized most commonly in tests for antibodies to hepatitis C virus, as noted above. In fact, studies of autoantibodies in autoimmune hepatitis have led to the recognition of new categories of autoimmune hepatitis. *Type I autoimmune hepatitis* is the classic syndrome occurring in young women, associated with marked hyperglobulinemia, lupoid features, and circulating ANA. *Type II autoimmune hepatitis*, often seen in children and more common in Mediterranean populations, is associated not with ANA but withanti-LKM. Actually, anti-LKM represent a heterogeneous group of antibodies. In type II autoimmune hepatitis, the antibody is anti-LKM1, directed against P450 IID6. This is the same anti-LKM seen in some patients with chronic hepatitis C. Anti-LKM2 is seen in drug-induced hepatitis, and anti-LKM3 is seen in patients with chronic hepatitis D. Type II autoimmune hepatitis has been subdivided by some authorities into two categories, one more typically autoimmune and the other associated with viral hepatitis type C. Autoimmune hepatitis type IIa is felt to be autoimmune, is more likely to occur in young women, is associated with hyperglobulinemia, is associated with high-titer anti-LKM1, responds to glucocorticoid therapy, and is seen commonly in western Europe and the United Kingdom. Type IIb autoimmune hepatitis is associated with hepatitis C virus infection, tends to occur in older men, is associated with normal globulin levels and low-titer anti-LKM1, responds to interferon, and occurs most commonly in Mediterranean countries. In addition, another type of autoimmune hepatitis has been recognized, *autoimmune hepatitis type III*. These patients lack ANA and anti-LKM1 and have circulating antibodies to soluble liver antigen, which are directed at hepatocyte cytoplasmic cytokeratins 8 and 18. Most of these patients are women and have clinical features similar to those of patients with type I autoimmune hepatitis.

TREATMENT

The mainstay of management in autoimmune or idiopathic (nonviral) hepatitis is glucocorticoid therapy. Several controlled clinical trials have documented that such therapy leads to symptomatic, clinical, biochemical, and histologic improvement as well as increased survival. A therapeutic response can be expected in up to 80% of patients. Unfortunately, therapy has not been shown to prevent ultimate progression to cirrhosis. Although some advocate the use of prednisolone (the hepatic metabolite of prednisone), prednisone is just as effective and is favored by most authorities. Therapy may be initiated at 20 mg/d, but a popular regimen in the United States relies on an initiation dose of 60 mg/d. This high dose is tapered successively over the course of a month down to a maintenance level of 20 mg/d. An alternative but equally effective approach is to begin with half the prednisone dose (30 mg/d) along with azathioprine (50 mg/d). With azathioprine maintained at 50 mg/d, the prednisone dose is tapered over the course of a month down to a maintenance level of 10 mg/d. The advantage of the combination approach is a reduction, over the span of an 18-month course of therapy, in serious, life-threatening complications of steroid therapy from 66% down to under 20%. Azathioprine alone, however, is not effective in achieving remission, nor is alternate-day glucocorticoid therapy. Although therapy has been shown to be effective for severe autoimmune hepatitis, therapy is not indicated for mild forms of chronic hepatitis (which used to be labeled chronic persistent hepatitis or chronic lobular hepatitis), and the efficacy of therapy in mild or asymptomatic autoimmune hepatitis has not been established.

Improvement of fatigue, anorexia, malaise, and jaundice tends to occur within days to several weeks; biochemical improvement occurs over the course of several weeks to months, with a fall in serum bilirubin and globulin levels and an increase in serum albumin. Serum aminotransferase levels usually drop promptly, but improvements inAST andALTalone do not appear to be a reliable marker of recovery in individual patients; histologic improvement, characterized by a decrease in mononuclear infiltration and in hepatocellular necrosis may be delayed for 6 to 24 months. Still, if interpreted cautiously, aminotransferase levels are valuable indicators of relative disease activity, and many authorities do *not* advocate serial liver biopsies to assess therapeutic success or to guide decisions to alter or stop therapy. Therapy should continue for at least 12 to 18 months. After tapering and cessation of therapy, the likelihood of relapse is at least 50%, even if posttreatment histology has improved to show mild chronic hepatitis, and the majority of patients require therapy at maintenance doses indefinitely. Continuing azathioprine alone after cessation of prednisone therapy may reduce the frequency of relapse.

If medical therapy fails, or when chronic hepatitis progresses to cirrhosis and is associated with life-threatening complications of liver decompensation, liver transplantation is the only recourse (Chap. 301). Recurrence of autoimmune hepatitis in the new liver occurs rarely, if at all.

DIFFERENTIAL DIAGNOSIS

Early during the course of chronic hepatitis, the disease may resemble typical *acute viral hepatitis*. Without histologic assessment, severe chronic hepatitis cannot be readily distinguished based on clinical or biochemical criteria from mild chronic hepatitis. In adolescence, *Wilson's disease* may present with features of chronic hepatitis long before neurologic manifestations become apparent and before the formation of Kayser-Fleischer rings; in this age group, serum ceruloplasmin and serum and urinary copper determinations plus measurement of liver copper levels will establish the correct diagnosis. *Postnecrotic* or *cryptogenic cirrhosis* and *primary biliary cirrhosis* share clinical features with autoimmune hepatitis; biochemical, serologic, and histologic assessments are usually sufficient to allow these entities to be distinguished from autoimmune hepatitis. Of course, the distinction between autoimmune ("idiopathic") and chronic viral hepatitis is not always straightforward, especially when viral antibodies occur in patients with autoimmune disease or when autoantibodies occur in patients with viral disease. Finally, the presence of extrahepatic features such as arthritis, cutaneous vasculitis, or pleuritis -- not to mention the presence of circulating autoantibodies -- may cause confusion with *rheumatologic disorders* such as rheumatoid arthritis and systemic lupus erythematosus. The existence of clinical and biochemical features of progressive necroinflammatory liver disease distinguishes chronic hepatitis from these other disorders, which are not associated with severe liver disease.

(Bibliography omitted in Palm version)

Back to Table of Contents

298. ALCOHOLIC LIVER DISEASE - *Mark E. Mailliard*, *Michael F. Sorrell*

Chronic and excessive alcohol ingestion is one of the major causes of liver disease in the western world. Classically, alcoholic liver injury comprises three major forms: (1) fatty liver, (2) alcoholic hepatitis, and (3) cirrhosis. Although cirrhosis is discussed in Chap. 299, it is important to emphasize that rarely does a pure form of liver injury exist by itself. Fatty liver is present in over 90% of binge and heavy drinkers. A much smaller percentage of drinkers progress to alcoholic hepatitis, thought to be a precursor to cirrhosis. Although alcohol is considered a direct hepatotoxin, only 10 to 20% of alcoholics develop alcoholic hepatitis. The explanation for this apparent paradox is unclear but involves complex factors such as gender and heredity.

ETIOLOGY AND PATHOGENESIS

Quantity and duration of alcohol intake are the most important risk factors involved in the development of alcoholic liver disease (Table 298-1). The roles of beverage type and pattern of drinking are less clear. Progress of the hepatic injury beyond the fatty liver stage seems to require additional risk factors that remain incompletely defined. Women are more susceptible to alcoholic liver injury than men; they develop advanced liver disease with substantially less alcohol intake. In general, the time it takes to develop liver disease is directly related to the amount of alcohol consumed. It is useful in estimating alcohol consumption to understand that one beer, four ounces of wine, or one ounce of 80% spirits all contain approximately 12 g of alcohol. The threshold for developing severe alcoholic liver disease in men is an intake of >60 to 80 g/d of alcohol for 10 years, while women are at increased risk for developing similar degrees of liver injury by consuming 20 to 40 g/d. Gender-dependent differences in the gastric and hepatic metabolism of alcohol, in addition to poorly understood hormonal factors, likely contribute to the increased susceptibility of women to alcohol-induced liver injury. Social, nutritional, immunologic, and host factors have all been postulated to play a part in the development of the pathogenic process.

Chronic infection with hepatitis C (HCV) (Chap. 297) is an important risk factor in the progression and acceleration of alcoholic liver disease. The presence of HCV in patients with severe alcoholic liver disease is increased five- to tenfold above that in a matched control alcoholic population. Patients with both alcoholic liver injury and HCV develop decompensated liver disease at a younger age and have poorer overall survival rates. As a consequence of the overlapping injurious processes secondary to alcohol abuse and HCV infection, patients can develop an increased liver iron burden and rarely, porphyria cutanea tarda.

Our understanding of the pathogenesis of alcoholic liver injury is incomplete. Alcohol is a direct hepatotoxin, but ingestion of alcohol initiates a variety of metabolic responses that influence the final hepatotoxic response. The initial concept of malnutrition as the major pathogenic mechanism has given way to the present understanding that the metabolism of alcohol by the hepatocyte initiates a cascade of events involving production of protein-aldehyde adducts, lipid peroxidation, immunologic events, and cytokine release (Fig. 298-1). The production of cytokines is in large measure responsible for the systemic manifestations of alcoholic hepatitis, e.g., fever, leukocytosis, and anorexia. The degree of fibrosis stimulated by these complex events determines the extent of architectural derangement of the liver after chronic alcohol ingestion.

PATHOLOGY

The liver has a limited repertoire in response to injury. Fatty liver is the initial and most common histologic response to increased alcohol ingestion. The accumulation of fat in the perivenular hepatocytes coincides with the location of alcohol dehydrogenase, the major enzyme responsible for alcohol metabolism. Continuing alcohol ingestion results in fat accumulation throughout the entire hepatic lobule. Despite extensive fatty changes and distortion of the hepatocytes with macrovesicular fat, the cessation of drinking results in normalization of hepatic architecture and fat content in the liver. Alcoholic fatty liver has traditionally been regarded as entirely benign; but similar to the spectrum of non-alcoholic steatohepatitis, certain pathologic features such as giant mitochondria, perivenular fibrosis, and macrovesicular fat may be associated with progressive liver injury.

The transition between fatty liver and the development of alcoholic hepatitis is blurred. The hallmark of alcoholic hepatitis is hepatocyte injury characterized by ballooning degeneration, spotty necrosis, polymorphonuclear infiltration, and fibrosis in the perivenular and perisinusoidal space of Disse. Mallory bodies are often present in florid cases but are neither specific nor necessary to establishing the diagnosis. Alcoholic hepatitis is thought to be a precursor to the development of cirrhosis. However, like fatty liver, it is potentially reversible with cessation of drinking. Cirrhosis is present in up to 50% of patients with biopsy-proven alcoholic hepatitis.

CLINICAL FEATURES

The clinical manifestations of alcoholic fatty liver are subtle and characteristically detected as a consequence of the patient's visit for a seemingly unrelated matter. Previously unsuspected hepatomegaly is often the only clinical finding. Occasionally, patients with fatty liver present with right upper quadrant discomfort, tender hepatomegaly, nausea, and jaundice. Differentiation of alcoholic fatty liver from non-alcoholic fatty liver is difficult unless an accurate drinking history is verified. Alcoholism does not respect social and economic class. In every instance where liver disease is present, a thoughtful and sensitive drinking history should be obtained. Alcoholic hepatitis is associated with a wide gamut of clinical features. Fever, spider nevi, jaundice, and abdominal pain simulating an acute abdomen represent the extreme end of the spectrum; but many patients are entirely asymptomatic. Recognition of the clinical features of alcoholic hepatitis is central to the initiation of an effective and appropriate diagnostic and therapeutic strategy.

LABORATORY FEATURES

Patients with alcoholic fatty liver are often identified through routine screening tests. The typical laboratory abnormalities are nonspecific and include modest elevations of the aspartate aminotransferase (AST) and alanine aminotransferase (ALT) accompanied by hypertriglyceridemia, hypercholesterolemia, and, occasionally, hyperbilirubinemia. In alcoholic hepatitis and in contrast to other causes of fatty liver, the AST and ALT are usually elevated two- to sevenfold. They rarely are above 400 IU, and the AST/ALT ratio is >1 (Table 298-2). Hyperbilirubinemia is common and is accompanied by modest increases in the alkaline phosphatase. Derangement in hepatocyte synthetic function indicates more serious disease. Hypoalbuminemia and coagulopathy are common in advanced liver injury. The mean corpuscular volume (MCV) and uric acid level are commonly elevated in chronic alcohol abuse. Measurement of the carbohydrate-deficient transferrin (CDT) is superior to the measurement of the gamma-glutamyl transpeptidase (GGTP) or MCV in identifying excessive drinking. Ultrasonography is useful in detecting fatty infiltration of the liver and determining liver size. The demonstration by ultrasound of portal vein flow reversal, ascites, and intra-abdominal collaterals indicates serious liver injury with less potential for complete reversal of liver disease.

PROGNOSIS

Critically ill patients with alcoholic hepatitis have short-term mortality rates approaching 70%. Severe alcoholic hepatitis is heralded by coagulopathy (prothrombin time >5 s), anemia, serum albumin concentrations below 2.5 mg/dL, serum bilirubin levels>8 mg/dL, renal failure, and ascites. A discriminant function calculated as 4.6 ´[prothrombin time - control(seconds)] + serum bilirubin (mg/dL) can identify patients with a poor prognosis (discriminant function >32). The presence of ascites, variceal hemorrhage, deep encephalopathy, or hepatorenal syndrome predicts a dismal prognosis. The pathologic stage of the injury can be helpful in predicting prognosis. Liver biopsy should be performed whenever possible to confirm the diagnosis, to establish potential reversibility of the liver disease, and to guide the therapeutic decisions.

TREATMENT

Complete abstinence from alcohol is the cornerstone in the treatment of alcoholic liver disease. Improved survival rates and the potential for reversal of histologic injury regardless of the initial clinical presentation are associated with total avoidance of alcoholic ingestion. Referral of patients to experienced alcohol counselors and/or alcohol treatment programs should be routine in the management of patients with alcoholic liver disease. Attention should be directed to the nutritional and psychosocial states during the evaluation and treatment periods. Because of data suggesting that the pathogenic mechanisms in alcoholic hepatitis involve cytokine release and the perpetuation of injury by immunologic processes, glucocorticoids have been extensively evaluated in the treatment of alcoholic hepatitis. Patients with severe alcoholic hepatitis, defined as a discriminant function >32, were given prednisone, 40 mg/d, or prednisolone, 32 mg/d, for 4 weeks followed by a steroid taper (Fig. 298-2). Exclusion criteria included active gastrointestinal bleeding, sepsis, renal failure, or pancreatitis. Because of inordinate surgical mortality rates and the high rates of recidivism after transplantation, patients with alcoholic hepatitis are not candidates for immediate liver transplantation. The transplant candidacy of these patients should be reevaluated after a defined period of sobriety.

(Bibliography omitted in Palm version)

Back to Table of Contents

299. CIRRHOSIS AND ITS COMPLICATIONS - *Raymond T. Chung*, *Daniel K. Podolsky*

Cirrhosis is a pathologically defined entity that is associated with a spectrum of characteristic clinical manifestations. The cardinal pathologic features reflect irreversible chronic injury of the hepatic parenchyma and include extensive fibrosis in association with the formation of regenerative nodules. These features result from hepatocyte necrosis, collapse of the supporting reticulin network with subsequent connective tissue deposition, distortion of the vascular bed, and nodular regeneration of remaining liver parenchyma. The central event leading to hepatic fibrosis is activation of the hepatic stellate cell. Upon activation by factors released by hepatocytes and Kupffer cells, the stellate cell assumes a myofibroblast-like conformation and, under the influence of cytokines such as transforming growth factor b(TGF-b), produces fibril-forming type I collagen. The precise point at which fibrosis becomes irreversible is unclear. The pathologic process should be viewed as a final common pathway of many types of chronic liver injury. Clinical features of cirrhosis derive from the morphologic alterations and often reflect the severity of hepatic damage rather than the etiology of the underlying liver disease. Loss of functioning hepatocellular mass may lead to jaundice, edema, coagulopathy, and a variety of metabolic abnormalities; fibrosis and distorted vasculature lead to portal hypertension and its sequelae, including gastroesophageal varices and splenomegaly. Ascites and hepatic encephalopathy result from both hepatocellular insufficiency and portal hypertension.

Classification of the various types of cirrhosis based on either etiology or morphology alone is unsatisfactory. A single pathologic pattern may result from a variety of insults, while the same insult may produce several morphologic patterns. Nevertheless, most types of cirrhosis may be usefully classified by a mixture of etiologically and morphologically defined entities as follows: (1) alcoholic; (2) cryptogenic and posthepatitic; (3) biliary; (4) cardiac; and (5) metabolic, inherited, and drug-related. This chapter considers the various types of cirrhosis and their complications.

ALCOHOLIC CIRRHOSIS

Definition *Alcoholic cirrhosis* is only one of many consequences resulting from chronic alcohol ingestion, and it often accompanies other forms of alcohol-induced liver injury, including alcoholic fatty liver and alcoholic hepatitis (Chap. 298). Alcoholic cirrhosis, historically referred to as *Laennec's cirrhosis*, is the most common type of cirrhosis encountered in North America and many parts of western Europe and South America. It is characterized by diffuse fine scarring, fairly uniform loss of liver cells, and small regenerative nodules, and therefore it is sometimes referred to as *micronodular cirrhosis*. However, micronodular cirrhosis may also result from other types of liver injury (e.g., following jejunoileal bypass), and thus alcoholic cirrhosis and micronodular cirrhosis are not necessarily synonymous. Conversely, alcoholic cirrhosis may progress to macronodular cirrhosis with time.

Etiology SeeChap. 298, "Alcoholic Liver Disease."

Pathology and Pathogenesis With continued alcohol intake and destruction of hepatocytes, fibroblasts (including activated hepatic stellate cells that have transformed into myofibroblasts with contractile properties) appear at the site of injury and deposit collagen. Weblike septa of connective tissue appear in periportal and pericentral zones and eventually connect portal triads and central veins. This fine connective tissue network surrounds small masses of remaining liver cells, which regenerate and form nodules. Although regeneration occurs within the small remnants of parenchyma, cell loss generally exceeds replacement. With continuing hepatocyte destruction and collagen deposition, the liver shrinks in size, acquires a nodular appearance, and becomes hard as "end-stage" cirrhosis develops. Although alcoholic cirrhosis is usually a progressive disease, appropriate therapy and strict avoidance of alcohol may arrest the disease at most stages and permit functional improvement. In addition, there is strong evidence that concomitant chronic hepatitis C virus (HCV) infection significantly accelerates development of alcoholic cirrhosis.

Clinical Features

Signs and Symptoms Alcoholic cirrhosis may be clinically silent, and many cases (10 to 40%) are discovered incidentally at laparotomy or autopsy. In many cases symptoms are insidious in onset, occurring usually after 10 or more years of excessive alcohol use and progressing slowly over subsequent weeks and months. Anorexia and malnutrition lead to weight loss and a reduction in skeletal muscle mass. The patient may experience easy bruising, increasing weakness, and fatigue. Eventually the clinical manifestations of hepatocellular dysfunction and portal hypertension ensue, including progressive jaundice, bleeding from gastroesophageal varices, ascites, and encephalopathy. The abrupt onset of one of these complications may be the first event prompting the patient to seek medical attention. In other cases, cirrhosis first becomes evident when the patient requires treatment of symptoms related to alcoholic hepatitis.

A firm, nodular liver may be an early sign of disease; the liver may be either enlarged, normal, or decreased in size. Other frequent findings include jaundice, palmar erythema, spider angiomas, parotid and lacrimal gland enlargement, clubbing of fingers, splenomegaly, muscle wasting, and ascites with or without peripheral edema. Men may have decreased body hair and/or gynecomastia and testicular atrophy, which, like the cutaneous findings, result from disturbances in hormonal metabolism, including increased peripheral formation of estrogen due to diminished hepatic clearance of the precursor androstenedione. Testicular atrophy may reflect hormonal abnormalities or the toxic effect of alcohol on the testes. In women, signs of virilization or menstrual irregularities may occasionally be encountered. Dupuytren's contractures resulting from fibrosis of the palmar fascia with resulting flexion contracture of the digits are associated with alcoholism but are not specifically related to cirrhosis.

Although the cirrhotic patient may stabilize if drinking is discontinued, over a period of years, the patient may become emaciated, weak, and chronically jaundiced. Ascites and other signs of portal hypertension may become increasingly prominent. Ultimately, most patients with advanced cirrhosis die in hepatic coma, commonly precipitated by hemorrhage from esophageal varices or intercurrent infection. Progressive renal dysfunction often complicates the terminal phase of the illness.

Laboratory Findings In advanced alcoholic liver disease, abnormalities of laboratory tests are more common. Anemia may result from acute and chronic gastrointestinal blood loss, coexistent nutritional deficiency (notably of folic acid and vitamin B_{12}), hypersplenism, and a direct suppressive effect of alcohol on the bone marrow. Hemolytic anemia, presumably due to effects of hypercholesterolemia or erythrocyte membranes resulting in unusual spurlike projections (acanthocytosis), has been described in some alcoholics with cirrhosis. Mild or pronounced hyperbilirubinemia may be found, usually in association with varying elevations of serum alkaline phosphatase levels. Levels of serum AST (asparate aminotransferase) are frequently elevated, but levels>5ukat (300 units) are unusual and should prompt one to look for other coincident or complicating factors. In contrast to viral hepatitis, the serum AST is usually disproportionately elevated relative to ALT (alanine aminotransferase), i.e., AST/ALT ratio >2. This discrepancy in alcoholic liver disease may result from the proportionally greater inhibition of ALT synthesis by ethanol, which may be partially reversed by pyridoxal phosphate.

The serum prothrombin time is frequently prolonged, reflecting reduced synthesis of clotting proteins, most notably the vitamin K-dependent factors (see "Coagulopathy," below). The serum albumin level is usually depressed, while serum globulins are increased. Hypoalbuminemia reflects in part overall impairment in hepatic protein synthesis, while hyperglobulinemia is thought to result from nonspecific stimulation of the reticuloendothelial system. Elevated blood ammonia levels in patients with hepatic encephalopathy reflect diminished hepatic clearance because of impaired liver function and shunting of portal venous blood around the cirrhotic liver into the systemic circulation (see "Hepatic Encephalopathy," below).

A variety of metabolic disturbances may be detected. Glucose intolerance due to endogenous insulin resistance may be present; however, clinical diabetes is uncommon. Central hyperventilation may lead to respiratory alkalosis in patients with cirrhosis. Dietary deficiency and increased urinary losses lead to hypomagnesemia and hypophosphatemia. In patients with ascites and dilutional hyponatremia, hypokalemia may occur from increased urinary potassium losses due in part to hyperaldosteronism. Prerenal azotemia is also observed in such patients.

Diagnosis Alcoholic cirrhosis should be strongly suspected in patients with a history of prolonged or excessive alcohol intake and physical signs of chronic liver disease. However, since only 10 to 15% of individuals with excessive alcohol intake develop cirrhosis, other causes and types of liver disease may have to be excluded. The clinical features and laboratory findings are usually sufficient to provide reasonable indication of the presence and extent of hepatic injury. Although a percutaneous needle biopsy of the liver is not usually necessary to confirm the typical findings of alcoholic hepatitis or cirrhosis, it may be helpful in distinguishing patients with less advanced liver disease from those with cirrhosis and in excluding other forms of liver injury such as viral hepatitis. Biopsy may also be helpful as a diagnostic tool in evaluating patients with clinical findings suggestive of alcoholic liver disease who deny alcohol intake. In patients with features of cholestasis, ultrasonography may be appropriate to exclude the presence of extrahepatic biliary obstruction. When the clinical status of an otherwise stable cirrhotic patient deteriorates without an obvious explanation, complicating conditions, such as infection, portal vein thrombosis, and hepatocellular carcinoma, should be sought.

Prognosis Abstinence from alcohol as well as early and appropriate medical care can decrease long-term morbidity and mortality and delay or prevent the appearance of further complications. Patients who have had a major complication of cirrhosis and who continue to drink have a 5-year survival of less than 50%. However, those patients who remain abstinent have a substantially better prognosis. In general, the overall outlook in patients with advanced liver disease remains poor; most of these patients eventually die as a result of massive variceal hemorrhage and/or profound hepatic encephalopathy.

TREATMENT

Alcoholic cirrhosis is a serious illness that requires long-term medical supervision and careful management. Therapy of the underlying liver disease is largely supportive. Specific treatment is directed at particular complications such as variceal bleeding and ascites (see below). While some studies suggest that administration of glucocorticoids in moderately large doses for 4 weeks is helpful in patients with severe alcoholic hepatitis and encephalopathy, these drugs have no role in the treatment of established alcoholic cirrhosis. While one study suggested a mortality benefit for the antifibrotic agent colchicine in alcoholic cirrhosis, it has not yet been reproduced; thus colchicine cannot be routinely recommended.

The patient should be made to realize that there is no medication that will protect the liver against the effects of further alcohol ingestion. Therefore, alcohol should be absolutely forbidden. An important component of the complete care of such patients is encouragement to become involved in an appropriate alcohol counseling program.

All medicines must be administered with caution in the patient with cirrhosis, especially those eliminated or modified through hepatic metabolism or biliary pathways. In particular, care must be taken to avoid overzealous use of drugs that may directly or indirectly precipitate complications of cirrhosis. For example, vigorous treatment of ascites with diuretics may result in electrolyte abnormalities or hypovolemia, which can lead to coma. Similarly, even modest doses of sedatives can lead to deepening encephalopathy. Aspirin should be avoided in patients with cirrhosis because of its effects on coagulation and gastric mucosa. Acetaminophen should be used with caution and in doses of less than 2 g/day. Patients who drink alcohol are more sensitive to the hepatotoxic effects of acetaminophen, probably due to increased metabolism of the drug to toxic intermediates and decreased glutathione levels.

POSTHEPATITIC AND CRYPTOGENIC CIRRHOSIS

Definition Posthepatitic or postnecrotic cirrhosis represents the final common pathway of many types of chronic liver disease. *Coarsely nodular* and *multilobular cirrhosis* are terms synonymous with posthepatitic cirrhosis. The term *cryptogenic cirrhosis* has been used interchangeably with postpathepatitic cirrhosis, but this designation should be reserved for those cases in which the etiology of cirrhosis is unknown (approximately 10% of all patients with cirrhosis).

Etiology *Posthepatitic cirrhosis* is a morphologic term referring to a defined stage of advanced chronic liver injury of both specific and unknown (cryptogenic) causes. Epidemiologic and serologic evidence suggest that viral hepatitis (hepatitis B or hepatitis

C) may be an antecedent factor in from one-fourth to three-fourths of cases of apparently cryptogenic posthepatitic cirrhosis. In areas where hepatitis B virus (HBV) infection is endemic (e.g., Southeast Asia, sub-Saharan Africa), up to 15% of the population may acquire the infection in early childhood, and cirrhosis may ultimately develop in one-fourth of these chronic carriers. Although HBV infection is much less prevalent in the United States, it is relatively common among certain high-risk groups (e.g., persons with multiple sexual partners, especially men who have sex with men, injection drug users) and contributes to an increased incidence of cirrhosis. In the United States,HCVinfection accounts for many cases of cirrhosis following blood transfusions. Before routine screening of blood donors was introduced, hepatitis C occurred in 5 to 10% of blood recipients. Following infection, cirrhosis may ultimately develop in more than 20% of individuals after 20 years. More than half of patients who would previously have been designated as having cryptogenic chronic liver disease have evidence of HCV infection. Increasing recognition of the progressive nature of nonalcoholic steatohepatitis has revealed that a large portion of cases previously designated cryptogenic cirrhosis may be attributable to this disorder (Chap. 300). Posthepatitic cirrhosis may also develop in patients with autoimmune hepatitis (Chap. 297).

The most common causes of cirrhosis in the United States, which ultimately lead to liver transplantation, include chronicHCVinfection, alcohol, primary biliary cirrhosis, primary sclerosing cholangitis, and nonalcoholic steatohepatitis (NASH). Less common causes of posthepatitic cirrhosis, including drugs and toxins, are listed inTable 299-1.

Pathology The posthepatitic liver is typically shrunken in size, distorted in shape, and composed of nodules of liver cells separated by dense and broad bands of fibrosis. The microscopic picture is consistent with the gross impression. Posthepatitic cirrhosis is characterized morphologically by (1) extensive confluent loss of liver cells, (2) stromal collapse and fibrosis resulting in broad bands of connective tissue containing the remains of many portal triads, and (3) irregular nodules of regenerating hepatocytes, varying in size from microscopic to several centimeters in diameter.

Clinical Features In patients with cirrhosis of known etiology in whom there is progression to a posthepatitic stage, the clinical manifestations are an extension of those resulting from the initial disease process. Usually clinical symptoms are related to portal hypertension and its sequelae, such as ascites, splenomegaly, hypersplenism, encephalopathy, and bleeding gastroesophageal varices. The hematologic and liver function abnormalities resemble those seen with other types of cirrhosis. In a few patients with posthepatitic cirrhosis, the diagnosis may be made incidentally at operation, at postmortem, or by a needle biopsy of the liver performed to investigate abnormal liver function tests or hepatomegaly.

Diagnosis and Prognosis Posthepatitic cirrhosis should be suspected in patients with signs and symptoms of cirrhosis or portal hypertension. Needle or operative liver biopsies confirm the diagnosis, although nonuniformity of the pathologic process may result in sampling errors. The diagnosis of cryptogenic cirrhosis is reserved for those patients in whom no known etiology can be demonstrated. About 75% of patients have progressive disease despite supportive therapy and die within 1 to 5 years from complications, including variceal hemorrhage, hepatic encephalopathy, or superimposed hepatocellular carcinoma.

TREATMENT

Management is usually limited to treatment of the complications of portal hypertension, including control of ascites, avoidance of drugs or excessive protein intake that may induce hepatic coma, and prompt treatment of infections (see below). In patients with asymptomatic cirrhosis, expectant management alone is appropriate. In those patients in whom posthepatitic cirrhosis has developed as a result of a treatable condition, therapy directed at the primary disorder may limit further progression (e.g., Wilson's disease, hemochromatosis).

BILIARY CIRRHOSIS

Biliary cirrhosis results from injury to or prolonged obstruction of either the intrahepatic or extrahepatic biliary system. It is associated with impaired biliary excretion, destruction of hepatic parenchyma, and progressive fibrosis. Primary biliary cirrhosis (PBC) is characterized by chronic inflammation and fibrous obliteration of intrahepatic bile ductules. Secondary biliary cirrhosis (SBC) is the result of long-standing obstruction of the larger extrahepatic ducts. Although primary and secondary biliary cirrhosis are separate pathophysiologic entities with respect to the initial insult, many clinical features are similar.

PRIMARY BILIARY CIRRHOSIS

Etiology and Pathogenesis The cause ofPBCremains unknown. Several observations suggest that a disordered immune response may be involved. PBC is frequently associated with a variety of disorders presumed to be autoimmune in nature, such as the syndrome of *c*alcinosis, *R*aynaud's phenomenon, *e*sophageal dysmotility, *s*clerodactyly, *t*elangiectasia (CREST); the sicca syndrome (dry eyes and dry mouth); autoimmune thyroiditis; type 1 diabetes mellitus; and IgA deficiency.

Most important, a circulating IgG antimitochondrial antibody (AMA) is detected in more than 90% of patients withPBC and only rarely in other forms of liver disease. It has been demonstrated that these autoantibodies recognize three to five inner mitochondrial membrane proteins identified as enzymes of the pyruvate dehydrogenase complex (PDC), the branched chain-ketoacid dehydrogenase complex (BCKDC), and thea-ketoglutarate dehydrogenase complex (KGDC). The major autoantigen in PBC (found in 90% of patients) has been identified as the 74-kDa E2 component of the PDC, dihydrolipoamide acetyltransferase. The antibodies are directed to a region essential for binding of a lipoic acid cofactor and inhibit the overall enzymatic activity of the PDC. Other AMA autoantibodies in PNC patients are directed to similar constituents of BCKDC and KGDC and also inhibit their enzymatic function. It remains unclear whether these properties have a direct pathogenetic role in the development of PBC. In addition to AMA, elevated serum levels of IgM and cryoproteins consisting of immune complexes capable of activating the alternative complement pathway are found in 80 to 90% of patients. Aberrant expression of major histocompatibility complex class II molecules has been found on biliary epithelium in association with PBC, suggesting that these cells may serve as antigen-presenting cells in this setting. Lymphocytes are prominent in the portal regions and surround damaged bile ducts. These histologic findings resemble those noted in graft-versus-host disease following bone marrow transplantation and suggest that damage to bile ducts may be immunologically mediated, perhaps reflecting a defect in a suppressor cell population.

Pathology PBC is divided into four stages based on morphologic findings. The earliest recognizable lesion (stage I), termed *chronic nonsuppurative destructive cholangitis*, is a necrotizing inflammatory process of the portal triads. It is characterized by destruction of medium and small bile ducts, a dense infiltrate of acute and chronic inflammatory cells, mild fibrosis, and occasionally, bile stasis. At times, periductal granulomas and lymph follicles are found adjacent to affected bile ducts. Subsequently, the inflammatory infiltrate becomes less prominent, the number of bile ducts is reduced, and smaller bile ductules proliferate (stage II). Progression over a period of months to years leads to a decrease in interlobular ducts, loss of liver cells, and expansion of periportal fibrosis into a network of connective tissue scars (stage III). Ultimately, cirrhosis, which may be micronodular or macronodular, develops (stage IV).

Clinical Features

Signs and Symptoms Many patients withPBC are asymptomatic, and the disease is initially detected on the basis of elevated serum alkaline phosphatase levels during routine screening. The majority of such patients remain asymptomatic for prolonged periods, although most ultimately develop progressive liver injury.

Among patients with symptomatic disease, 90% are women age 35 to 60. Often the earliest symptom is pruritus, which may be either generalized or limited initially to the palms and soles. In addition, fatigue is commonly a prominent early symptom. After several months or years, jaundice and gradual darkening of the exposed areas of the skin (melanosis) may ensue. Other early clinical manifestations ofPBCreflect impaired bile excretion. These include steatorrhea and the malabsorption of lipid-soluble vitamins. Protracted elevation of serum lipids, especially cholesterol, leads to subcutaneous lipid deposition around the eyes (xanthelasmas) and over joints and tendons (xanthomas). Over a period of months to years, the itching, jaundice, and hyperpigmentation slowly worsen. Eventually, signs of hepatocellular failure and portal hypertension develop and ascites appears. Progression may be quite variable. Whereas a proportion of asymptomatic patients may show no signs of progression for a decade or longer, in others, death due to hepatic insufficiency may occur within 5 to 10 years after the first signs of the illness. Such decompensation is often precipitated by uncontrolled variceal hemorrhage or infection.

Physical examination may be entirely normal in the early phase of the disease, when patients are asymptomatic or pruritus is the sole complaint. Later, there may be jaundice of varying intensity, hyperpigmentation of the exposed skin areas, xanthelasmas and tendinous and planar xanthomas, moderate to striking hepatomegaly, splenomegaly, and clubbing of the fingers. Bone tenderness, signs of vertebral compression, ecchymoses, glossitis, and dermatitis may all be noted. Clinical evidence of the sicca syndrome can be found in as many as 75% of patients, and serologic evidence of autoimmune thyroid disease in 25%. Other conditions encountered with increased frequency include rheumatoid arthritis,CRESTsyndrome, keratoconjunctivitis sicca, IgA deficiency, type 1 diabetes mellitus, scleroderma, pernicious anemia, and renal tubular acidosis. Bone disease is often a significant problem encountered over the course of the disease. While osteomalacia occurs due to diminished vitamin D absorption, accelerated osteoporosis in this patient population (the majority of whom are postmenopausal women) is even more common.

Laboratory Findings PBC is increasingly diagnosed at a presymptomatic stage, prompted by the finding of a twofold or greater elevation of the serum alkaline phosphatase during routine screening. Serum 5¢-nucleotidase activity andg-glutamyl transpeptidase levels are also elevated. In this setting, serum bilirubin is usually normal and aminotransferase levels minimally increased. The diagnosis is supported by a positiveAMA test (titer > 1:40). The latter is both relatively specific and sensitive; a positive test is found in over 90% of symptomatic patients and is present in fewer than 5% of patients with other liver diseases. As the disease evolves, the serum bilirubin level rises progressively and may reach 510 umol/L (30 mg/dL) or more in the final stages. Serum aminotransferase values rarely exceed 2.5 to 3.3 ukat (150 to 200 units). Hyperlipidemia is common, and a striking increase of the serum unesterified cholesterol is often noted. An abnormal serum lipoprotein (lipoprotein X) may be present in PBC but is not specific and appears in other cholestatic conditions. A deficiency of bile salts in the intestine leads to moderate steatorrhea and impaired absorption of the fat-soluble vitamins and hypoprothrombinemia. Patients with PBC have elevated liver copper levels, but this finding is not specific and is found in all disorders in which there is prolonged cholestasis.

Diagnosis PBCshould be considered in middle-aged women with unexplained pruritus or an elevated serum alkaline phosphatase and in whom there may be other clinical or laboratory features of protracted impairment of biliary excretion. Although a positive serumAMAdetermination provides important diagnostic evidence, false-positive results do occur; therefore, liver biopsy should be performed to confirm the diagnosis. Rarely, the AMA test may be negative in patients with histologic features of PBC. Frequently, patients have antibodies to the E2 protein in tests using these specific antigens. In some cases with histologic features of PBC and a negative AMA, antinuclear or smooth-muscle antibodies are present (as in autoimmune hepatitis), and the designation *autoimmune cholangitis* is applied. The natural history of this entity, however, appears to resemble that of PBC. If the AMA test is negative, the biliary tract should be evaluated to exclude primary sclerosing cholangitis and remediable extrahepatic biliary tract obstruction, especially in view of the frequent presence of coexisting cholelithiasis.

TREATMENT

While there is no specific therapy for PBC, ursodiol has been shown to improve biochemical and histologic features and might improve survival, particularly liver transplantation-free survival (although this remains unproven). Ursodiol should be given in doses of 10 to 15 mg/kg per day, but lower doses are sometimes just as effective in reducing serum alkaline phosphatase and aminotransferase levels. Ursodiol should be given with food and can be taken in a single dose daily. Side effects are rare: gastrointestinal intolerance (bloating, indigestion) and skin rashes occur but are uncommon. Isolated instances of severe exacerbation of pruritus have been reported in patients with advanced disease. Ursodiol probably works by replacing the endogenously produced hydrophobic bile acids with urosdeoxycholate, a hydrophilic and relatively nontoxic bile acid.

Unfortunately, ursodiol does not prevent ultimate progression ofPBC, and the only established "cure" is liver transplantation. Results of liver transplantation for PBC are excellent, survival exceeding that for patients receiving transplantation for most other forms of end-stage liver disease. Recurrence of PBC after liver transplantation has been reported but is uncommon, and the recurrent disease is only slowly progressive. Most patients remainAMApositive after transplantation, and as many as 25% will have histologic features of PBC on liver biopsy after 5 years. Other therapies such as glucocorticoids, colchicine, methotrexate, azathioprine, cyclosporine, and tacrolimus have been reported as effective in small cases series, but none have shown to be effective in adequately controlled trials.

Relief of symptoms is also an important part of management ofPBC. As noted, ursodiol may be helpful in controlling symptoms and improving the patient's sense of well-being. Although the mechanism of the protracted pruritus is not entirely clear, cholestyramine, an oral bile salt-sequestering resin, may be helpful in doses of 8 to 12 g/d to decrease both pruritus and hypercholesterolemia. Rifampin, opiate antagonists, ondansetron, plasmapheresis, and ultraviolet light have all been tried for control of pruritus, with varying results. Steatorrhea can be reduced by a low-fat diet and substituting medium-chain triglycerides for dietary long-chain triglycerides. Fat-soluble vitamins A and K should be given by parenteral injection at regular intervals to prevent or correct night blindness and hypoprothrombinemia, respectively. Zinc supplementation may be necessary if night blindness is refractory to vitamin A therapy. An important part of management of PBC and any cholestatic liver disease is assessment and treatment of osteoporosis and osteomalacia. Patients should be screened periodically by bone densitometry and treated as needed with calcium supplements, estrogen, and/or the newer bisphosphonate agents (e.g., alendronate). Progression of PBC leads to the typical complications of advanced liver disease (see below).

SECONDARY BILIARY CIRRHOSIS

Etiology SBCresults from prolonged partial or total obstruction of the common bile duct or its major branches. In adults, obstruction is most frequently caused by postoperative strictures or gallstones, usually with superimposed infectious cholangitis. Chronic pancreatitis may lead to biliary stricture and secondary cirrhosis. SBC is also an important complication of primary sclerosing cholangitis, a progressive immunologic disorder of the intrahepatic and extrahepatic biliary tree (Chap. 302). Patients with malignant tumors of the common bile duct or pancreas rarely survive long enough to develop SBC. In children, congenital biliary atresia and cystic fibrosis are common causes of SBC. Choledochal cysts, if unrecognized, may also be a rare cause of SBC.

Pathology and Pathogenesis Unrelieved obstruction of the extrahepatic bile ducts leads to (1) bile stasis and focal areas of centrilobular necrosis followed by periportal necrosis, (2) proliferation and dilatation of the portal bile ducts and ductules, (3) sterile or infected cholangitis with accumulation of polymorphonuclear infiltrates around bile ducts, and (4) progressive expansion of portal tracts by edema and fibrosis.

Extravasation of bile from ruptured interlobular bile ducts into areas of periportal necrosis leads to the formation of "bile lakes" surrounded by cholesterol-rich pseudoxanthomatous cells. As in other forms of cirrhosis, injury is accompanied by regeneration in residual parenchyma. These changes gradually lead to a finely nodular cirrhosis. In general, at least 3 to 12 months is required for biliary obstruction to result in cirrhosis. Relief of the obstruction is frequently accompanied by biochemical and morphologic improvement.

Clinical Features The symptoms, signs, and biochemical findings ofSBC are similar to those ofPBC. Jaundice and pruritus are usually the most prominent features. In addition, fever and/or right upper quadrant pain, reflecting bouts of cholangitis or biliary colic, are typical. The manifestations of portal hypertension are found only in advanced cases. SBC should be considered in any patient with clinical and laboratory evidence of prolonged obstruction to bile flow, especially when there is a history of previous biliary tract surgery or gallstones, bouts of ascending cholangitis, or right upper quadrant pain. Cholangiography (either percutaneous or endoscopic) usually demonstrates the underlying pathologic process. Liver biopsy, although not always necessary from a clinical standpoint, can document the development of cirrhosis.

TREATMENT

Relief of obstruction to bile flow, by either endoscopic or surgical means, is the most important step in the prevention and therapy ofSBC. Effective decompression of the biliary tract results in a significant improvement in both symptoms and survival, even in patients with established cirrhosis. When obstruction cannot be relieved, as in sclerosing cholangitis, antibiotics may be helpful acutely in controlling superimposed infection or, when administered on a chronic basis, as prophylactic therapy in suppressing recurring episodes of ascending cholangitis. Without relief of obstruction, there is a steady progression to end-stage cirrhosis and its terminal manifestations.

CARDIAC CIRRHOSIS

Definition Prolonged, severe right-sided congestive heart failure may lead to chronic liver injury and cardiac cirrhosis. The characteristic pathologic features of fibrosis and regenerative nodules distinguish cardiac cirrhosis from both reversible passive congestion of the liver due to acute heart failure and acute hepatocellular necrosis ("ischemic hepatitis" or "shock liver") resulting from systemic hypotension and hypoperfusion of the liver.

Etiology and Pathology In right-sided heart failure, retrograde transmission of elevated venous pressure via the inferior vena cava and hepatic veins leads to congestion of the liver. Hepatic sinusoids become dilated and engorged with blood, and the liver becomes tensely swollen. With prolonged passive congestion and ischemia from poor perfusion secondary to reduced cardiac output, necrosis of centrilobular hepatocytes ensues and leads to fibrosis in these central areas. Ultimately, centrilobular fibrosis develops, with collagen extending outward in a characteristic stellate pattern from the central vein. Gross examination of the liver shows alternating red (congested) and pale (fibrotic) areas, a pattern often referred to as "nutmeg liver." Improvement in management of cardiac disorders, particularly advances in surgical treatment, has reduced the frequency of cardiac cirrhosis.

Clinical Features A range of abnormalities of liver function tests may be found, though none is uniformly present. The serum bilirubin is usually only mildly increased and may be predominantly either conjugated or unconjugated. Mild to moderate elevation in alkaline phosphatase level and prothrombin time prolongation are sometimes present. TheASTlevel is typically mildly elevated but may be transiently very high following a period of marked systemic hypotension (shock liver), when the clinical picture can mimic acute viral or drug-induced hepatitis. In cases of tricuspid insufficiency the liver may be pulsatile, but this finding disappears as cirrhosis develops. With prolonged right-sided heart failure the liver becomes enlarged, firm, and usually nontender. The signs and symptoms of heart failure usually overshadow the liver disease. Bleeding from esophageal varices is rare, but chronic encephalopathy may be prominent, with a waxing and waning course reflecting variations in the severity of right-sided heart failure. Ascites and peripheral edema, often primarily related to the underlying cardiac dysfunction, may be worsened by the superimposed liver disease.

Diagnosis The presence of a firm, enlarged liver with signs of chronic liver disease in a patient with valvular heart disease, constrictive pericarditis, or cor pulmonale of long duration (>10 years) should suggest cardiac cirrhosis. Liver biopsy can confirm the diagnosis but is usually contraindicated because of coagulopathy or ascites. Coexistent chronic heart and liver disease should also raise the possibility of hemochromatosis (Chap. 345), amyloidosis (Chap. 319), or other infiltrative diseases.

Budd-Chiari syndrome resulting from the occlusion of the hepatic veins or inferior vena cava may be confused with acute congestive hepatomegaly. In this condition the liver is grossly enlarged and tender, and severe intractable ascites is present. However, signs and symptoms of heart failure are notably absent. The most common cause is thrombosis of the hepatic veins, often in the setting of polycythemia rubra vera, myeloproliferative syndromes, paroxysmal nocturnal hemoglobinuria, oral contraceptive use, or other hypercoagulable states; it may also result from invasion of the inferior vena cava by tumor, such as renal cell or hepatocellular carcinoma. Idiopathic membranous obstruction of the inferior vena cava is the most common cause of this syndrome in Japan. Hepatic venography or liver biopsy showing centrilobular congestion and sinusoidal dilatation in the absence of right-sided heart failure establishes the diagnosis of Budd-Chiari syndrome. Venocclusive disease affecting the sublobular branches of the hepatic veins and the hepatic venues may result from hepatic irradiation, treatment with certain antineoplastic agents, or ingestion of pyrrolidizine alkaloids present in some herbal teas ("bush tea disease") and can mimic congestive hepatomegaly.

TREATMENT

Prevention or treatment of cardiac cirrhosis depends on the diagnosis and therapy of the underlying cardiovascular disorder. Improvement in cardiac function frequently results in improvement of liver function and stabilization of the liver disease.

METABOLIC, HEREDITARY, DRUG-RELATED, AND OTHER TYPES OF CIRRHOSIS (See Table 299-1)

Cirrhosis or hepatitis may result from a wide variety of other processes encompassing the spectrum of etiologic factors listed in Table 299-2. Although some of these disorders have distinctive clinical or morphologic features, the manifestations of cirrhosis are largely independent of the underlying pathogenic mechanism.

NONCIRRHOTIC FIBROSIS OF THE LIVER

Several diseases, either congenital or acquired, may be associated with localized or generalized hepatic fibrosis. They are distinguished from cirrhosis by the absence of hepatocellular damage and the lack of nodular regenerative activity. The clinical manifestations in such cases are largely secondary to portal hypertension. The different types of these disorders are indicated inTable 299-2; with the exception of schistosomiasis in some regions of the world, all these conditions are relatively rare.

MAJOR COMPLICATIONS OF CIRRHOSIS

The clinical course of patients with advanced cirrhosis is often complicated by a number of important sequelae that are independent of the etiology of the underlying liver disease. These include portal hypertension and its consequences (e.g., gastroesophageal varices and splenomegaly), ascites, hepatic encephalopathy, spontaneous bacterial peritonitis, hepatorenal syndrome, and hepatocellular carcinoma.

PORTAL HYPERTENSION

Definition and Pathogenesis Normal pressure in the portal vein is low (5 to 10 mmHg) because vascular resistance in the hepatic sinusoids is minimal. Portal hypertension (>10 mmHg) most commonly results from increased resistance to portal blood flow. Because the portal venous system lacks valves, resistance at any level between the right side of the heart and splanchnic vessels results in retrograde transmission of an elevated pressure. Increased resistance can occur at three levels relative to the hepatic sinusoids: (1) presinusoidal, (2) sinusoidal, and (3) postsinudoidal. Obstruction in the *presinusoidal* venous compartment may be anatomically outside the liver (e.g., portal vein thrombosis) or within the liver itself but at a functional level proximal to the hepatic sinusoids so that the liver parenchyma is not exposed to the elevated venous pressure (e.g., schistosomiasis).

Postsinusoidal obstruction may also occur outside the liver at the level of the hepatic veins (e.g., Budd-Chiari syndrome), the inferior vena cava, or, less commonly, within the liver (e.g., venocclusive disease). When cirrhosis is complicated by portal hypertension, the increased resistance is usually *sinusoidal*. While distinctions between pre-, post-, and sinusoidal processes are conceptually appealing, functional resistance to portal flow in a given patient may occur at more than one level. Portal hypertension may also arise from increased blood flow (e.g., massive splenomegaly or arteriovenous fistulas), but the low outflow resistance of the normal liver makes this a rare clinical problem.

Cirrhosis is the most common cause of portal hypertension in the United States. Clinically significant portal hypertension is present in >60% of patients with cirrhosis. *Portal vein obstruction* is the second most common cause; it may be idiopathic or occur in association with cirrhosis, infection, pancreatitis, or abdominal trauma. Portal vein thrombosis may develop in a variety of hypercoagulable states including polycythemia vera; essential thrombocythemia; deficiencies of protein C, protein S, or antithrombin III; resistance to activated protein C (factor V Leiden); and a mutation of the prothrombin gene (G20210A). Portal vein thrombosis may be idiopathic, though some of these patients may have a subclinical myeloproliferative disorder. Hepatic vein thrombosis (Budd-Chiari syndrome) and hepatic venoocclusive disease are relatively infrequent causes of portal hypertension (see above). Portal vein occlusion may result in massive hematemesis from gastroesophageal varices, but ascites is usually found only when cirrhosis is present. Noncirrhotic portal fibrosis (Table 299-2) accounts for only a few cases of portal hypertension.

Clinical Features The major clinical manifestations of portal hypertension include hemorrhage from gastroesophageal varices, splenomegaly with hypersplenism, ascites, and acute and chronic hepatic encephalopathy. These are related, at least in part, to the development of portal-systemic collateral channels. The absence of valves in the portal venous system facilitates retrograde (hepatofugal) blood flow from the high-pressure portal venous system to the lower-pressure systemic venous circulation. Major sites of collateral flow involve the veins around cardioesophageal junction (esophagogastric varices), the rectum (hemorrhoids), retroperitoneal space, and the falciform ligament of the liver (periumbilical or abdominal wall collaterals). Abdominal wall collaterals appear as tortuous epigastric vessels that radiate from the umbilicus toward the xiphoid and rib margins (caput medusae).

A frequent marker of the presence of cirrhosis in a patient being followed for chronic liver disease is a progressive decrease in platelet count. A low-normal platelet count can be the first clue to progression to cirrhosis. Ultimately, a marked decrease in platelets (to 30,000 to 60,000/uL) and white blood cells can occur.

Diagnosis In patients with known liver disease, the development of portal hypertension is usually revealed by the appearance of splenomegaly, ascites, encephalopathy, and/or esophageal varices. Conversely, the finding of any of these features should prompt evaluation of the patient for the presence of underlying portal hypertension and liver disease. Varices are most reliably documented by fiberoptic esophagoscopy; their presence lends indirect support to the diagnosis of portal hypertension. Although rarely necessary, portal venous pressure may be measured directly by percutaneous transhepatic "skinny needle" catheterization or indirectly through transjugular cannulation of the hepatic veins. Both free and wedged hepatic vein pressure should be measured. While the latter is elevated in sinusoidal and postsinusoidal portal hypertension, including cirrhosis, this measurement is usually normal in presinusoidal portal hypertension. In patients in whom additional information is necessary (e.g., preoperative evaluation before portal-systemic shunt surgery) or when percutaneous catheterization is not feasible, mesenteric and hepatic angiography may be helpful. Particular attention should be directed to the venous phase to assess the patency of the portal vein and the direction of portal blood flow.

TREATMENT

Although treatment is usually directed toward a specific complication of portal hypertension, attempts are sometimes made to reduce the pressure in the portal venous system. Surgical decompression procedures have been used for many years to lower portal pressure in patients with bleeding esophageal varices (see below). However, portal-systemic shunt surgery does not result in improved survival rates in patients with cirrhosis. Decompression can now be accomplished without surgery through the percutaneous placement of a portal-systemic shunt, termed a *transjugular intrahepatic portosystemic shunt* (TIPS).b-Adrenergic blockade with propranolol or nadolol reduces portal pressure through vasodilatory effects on both the splanchnic arterial bed and the portal venous system in combination with reduced cardiac output. Such therapy has been shown to be effective in preventing both a first variceal bleed and subsequent episodes after an initial bleed. Treatment of patients with clinically significant sequelae of portal hypertension, especially variceal bleeding, with doses of propranolol titrated to reduce the resting pulse by 25% is reasonable if no contraindications exist.

Vigorous treatment of patients with alcoholic hepatitis and cirrhosis, chronic active hepatitis, or other liver diseases may lead to a fall in portal pressure and to a reduction in variceal size. In general, however, portal hypertension due to cirrhosis is not reversible. In appropriately selected patients, hepatic transplantation will be beneficial.

VARICEAL BLEEDING

Pathogenesis While vigorous hemorrhage may arise from any portal-systemic venous collaterals, bleeding is most common from varices in the region of the gastroesophageal junction. The factors contributing to bleeding from gastroesophageal varices are not entirely understood but include the degree of portal hypertension (>12 mmHg) and the size of the varices.

Clinical Features and Diagnosis Variceal bleeding often occurs without obvious precipitating factors and usually presents with painless but massive hematemesis with or without melena. Associated signs range from mild postural tachycardia to profound shock, depending on the extent of blood loss and degree of hypovolemia. Because patients with varices may bleed just as frequently from other gastrointestinal lesions (e.g., peptic ulcer, gastritis), exclusion of other bleeding sources is important even in patients with prior variceal hemorrhage. Endoscopy is the best approach to evaluate upper gastrointestinal hemorrhage in patients with known or suspected portal hypertension.

TREATMENT

(SeeFig. 299-1) Variceal bleeding is a life-threatening emergency. Prompt estimation and vigorous replacement of blood loss to maintain intravascular volume are essential and take precedence over diagnostic studies and more specific intervention to stop the bleeding. However, excessive fluid administration can increase portal pressure with resultant further bleeding and should therefore be avoided. Replacement of clotting factors with fresh-frozen plasma is important in patients with coagulopathy. Patients are best managed in an intensive care unit and require close monitoring of central venous or pulmonary capillary wedge pressures, urine output, and mental status. Only when the patient is hemodynamically stable should attention be directed toward specific diagnostic studies (especially endoscopy) and other therapeutic modalities to prevent further or recurrent bleeding.

About half of all episodes of variceal hemorrhage cease without intervention, although the risk of rebleeding is very high. The medical management of acute variceal hemorrhage includes the use of vasoconstrictors (somatostatin/octreotide or vasopressin), balloon tamponade, and endoscopic banding of varices or endoscopic sclerosis of varices (sclerotherapy). Intravenous infusion of *vasopressin* at a rate of 0.1 to 0.4 U/min results in generalized vasoconstriction leading to diminished blood flow in the portal venous system. Intravenous infusion of vasopressin is as effective as selective intraarterial administration. Control of bleeding can be achieved in up to 80% of cases, but bleeding recurs in more than half after the vasopressin is tapered and discontinued. Furthermore, a number of serious side effects, including cardiac and gastrointestinal tract ischemia, acute renal failure, and hyponatremia, may be associated with vasopressin therapy. Concurrent use of venodilators such as nitroglycerin as an intravenous infusion or isosorbide dinitrate sublingually may enhance the effectiveness of vasopressin and reduce complications. *Somatostatin* and its analogue, *octreotide*, are direct splanchnic vasoconstrictors. In some studies somatostatin, given as an initial 250-ug bolus followed by constant infusion (250 ug/h), has been found to be as effective as vasopressin. Octreotide at doses of 50 to 100 ug/h is also effective. These agents are preferable to vasopressin, offering equivalent efficacy with fewer complications. If bleeding is too vigorous or endoscopy is not available, *balloon tamponade* of the bleeding varices may be accomplished with a triple-lumen (Sengstaken-Blakemore) or four-lumen (Minnesota) tube with esophageal and gastric balloons. Because of the high risk of aspiration, endotracheal intubation should be performed prior to placing one of these tubes. After the tube is introduced into the stomach, the gastric balloon is inflated and pulled back into the cardia of the stomach. If bleeding does not stop, the esophageal balloon is inflated for additional tamponade. Complications occur in 15% or more of patients and include aspiration pneumonitis as well as esophageal rupture.

Where available, *endoscopic intervention* should be employed as the first line of treatment to control bleeding acutely (Chaps. 44 and283). Over the past 18 years, endoscopic sclerosis of esophageal varices has been extensively employed. In this procedure, the varices are injected with one of several sclerosing agents via a needle-tipped catheter passed through the endoscope. After endoscopic identification of varices as the presumed source of bleeding, sclerotherapy controls acute bleeding in up to 90% of cases. In addition, repeated sclerotherapy can be performed until obliteration of all varices is accomplished in an effort to prevent recurrent bleeding. While available data support the efficacy of sclerotherapy in controlling bleeding acutely and in decreasing rebleeding rates, repeated sclerotherapy has not been documented to prolong survival. Mucosal ulceration resulting from injection of the caustic sclerosant may occur and result in further hemorrhage or stenosis. More recently, endoscopic band ligation, in which esophageal varices are ligated and strangulated with endoscopically placed small elastic O-rings, has gained favor. Band ligation has proven to be at least as effective as sclerotherapy in controlling acute variceal bleeding and preventing rebleeding. Because it has been associated with fewer treatment-related complications, band ligation is recommended for long-term obliteration of varices that have bled. Although prophylactic sclerosis or banding of esophageal varices in the absence of proven bleeding cannot yet be recommended, one report suggests that banding may be more effective than beta-blockade in primary prevention of variceal bleeding in high-risk patients.

The effectiveness of *nonselective b-adrenergic blocking agents* (e.g., propranolol) in the management of acute variceal bleeding is limited due to concomitant hypotension resulting from hypovolemia. However, a number of studies suggest they may be of value in secondary prevention of recurrent variceal hemorrhage. Moreover, prophylactic treatment with nonselective beta blockers (propranolol or nadolol) in patients with large ("high-risk") varices that have never bled appears to decrease the incidence of bleeding and prolong survival. Thus, endoscopic screening for varices in patients with cirrhosis is desirable; some have suggested this should be repeated every other year. Patients with portal hypertension without specific contraindications should be given propranolol in doses that produce a 25% reduction in the resting heart rate or the hepatic venous pressure gradient (HVPG), where available. Propranolol may also prevent recurrent bleeding from severe portal hypertensive gastropathy in patients with cirrhosis. The optimal combination of endoscopic and pharmacologic therapy for prevention of recurrent hemorrhage remains to be established and is the subject of ongoing trials.

Surgical treatment of portal hypertension and variceal bleeding involves the creation of a portal-systemic shunt to permit decompression of the portal system. Two types of portal systemic shunts have been used: *nonselective shunts*, to decompress the entire portal system, and *selective shunts*, intended to decompress only the varices while maintaining blood flow to the liver itself. Nonselective shunts include end-to-side or side-to-side portacaval and proximal splenorenal anastomoses; selective shunts include the distal splenorenal shunt. Nonselective shunts are more likely to be complicated by encephalopathy than selective shunts. Emergency portal-systemic nonselective shunts may control acute hemorrhage, but such surgery is usually used only as a last resort because early operative mortality can be high. The role of portal-systemic shunt surgery after initial control of bleeding by nonoperative means is also uncertain. Surgically created shunts effectively reduce the risk of recurrent hemorrhage, but the overall mortality of patients undergoing such surgery is comparable to that of unoperated patients. Although patients who have undergone portal-systemic surgery succumb to recurrent bleeding less commonly than unoperated patients, this improvement is counterbalanced by increased morbidity from encephalopathy and death from progressive liver failure. Increasingly, therapeutic portal-systemic shunts have been reserved for patients who experience further bleeding despite serial endoscopic sclerotherapy or band ligation.

InTIPS, a technique developed to create a portal-systemic shunt by a percutaneous approach, an expandable metal stent is advanced to the hepatic veins under angiographic guidance and then through the substance of the liver to create a direct portacaval channel. This technique offers an alternative to surgery for refractory bleeding due to portal hypertension. However, stents frequently undergo stenosis or occlude over a period of months, prompting the need for a second TIPS or an alternative approach. Encephalopathy may be encountered after TIPS just as in the surgical shunts and is especially problematic in the elderly and those patients with preexisting encephalopathy. TIPS should be reserved for those individuals who fail endoscopic or medical management and are poor surgical risks. TIPS may have a useful role as a "bridge" for those patients with end-stage cirrhosis awaiting liver transplantation. Procedures such as esophageal transection have also been advocated for the management of acute variceal bleeding, but their efficacy remains unproven. Even though recent trials found that esophageal transection was as effective as endoscopic sclerotherapy, transection is usually considered a last resort.

The management of bleeding gastric fundal varices, either alone or in conjunction with esophageal varices, is more problematic, since sclerotherapy and banding are generally not effective. Vasoactive pharmacologic therapy should be instituted, butTIPS or shunt surgery should be considered because of high failure and rebleeding rates. For isolated gastric varices, splenic vein thrombosis should be specifically sought, since splenectomy is curative.

Portal Hypertensive Gastropathy Although variceal hemorrhage is the most commonly encountered bleeding complication of portal hypertension, many patients will develop a congestive gastropathy due to the venous hypertension. In this condition, identified by endoscopic examination, the mucosa appears engorged and friable. Indolent mucosal bleeding occurs rather than the brisk hemorrhage typical of a variceal source. b-Adrenergic blockade with propranolol (reducing splanchnic arterial pressure as well as portal pressure) is sometimes effective in ameliorating this condition. H_2receptor antagonists or other agents useful in the treatment of peptic disease are usually not helpful.

SPLENOMEGALY

Definition and Pathogenesis Congestive splenomegaly is common in patients with severe portal hypertension. Rarely, massive splenomegaly from nonhepatic disease leads to portal hypertension due to increased blood flow in the splenic vein.

Clinical Features Although usually asymptomatic, splenomegaly may be massive and contribute to the thrombocytopenia or pancytopenia of cirrhosis. In the absence of cirrhosis, splenomegaly in association with variceal hemorrhage should suggest the possibility of splenic vein thrombosis.

TREATMENT

Splenomegaly usually requires no specific treatment, although massive enlargement of the spleen may occasionally necessitate splenectomy at the time of shunt surgery. However, it should be noted that splenectomy without an accompanying shunt may actually increase portal pressure, and portal vein thrombosis may result from splenectomy. Splenectomy may also be indicated if splenomegaly is the cause rather than the result of portal hypertension (as in splenic vein thrombosis). Thrombcytopenia alone is rarely severe enough to necessitate removal of the spleen. Splenectomy should be avoided in a patient eligible for liver transplantation.

ASCITES

Definition Ascites is the accumulation of excess fluid within the peritoneal cavity. It is most frequently encountered in patients with cirrhosis and other forms of severe liver disease, but a number of other disorders may lead to either transudative or exudative ascites (Chap. 46).

Pathogenesis The accumulation of ascitic fluid represents a state of total-body sodium and water excess, but the event that initiates this imbalance is unclear. Three theories have been proposed (Fig. 299-2). The "underfilling" theory suggests that the primary abnormality is inappropriate sequestration of fluid within the splanchnic vascular bed due to portal hypertension and a consequent decrease in effective circulating blood volume. According to this theory, an apparent decrease in intravascular volume (underfilling) is sensed by the kidney, which responds by retaining salt and water. The "overflow" theory suggests that the primary abnormality is inappropriate renal retention of salt and water in the absence of volume depletion. A third and more recent theory, the peripheral arterial vasodilation hypothesis, may unify the earlier theories and accounts for the constellation of arterial hypotension and increased cardiac output in association with high levels of vasoconstrictor substances that are routinely found in patients with cirrhosis and ascites. Again, sodium retention is considered secondary to arterial vascular underfilling and the result of a disproportionate increase of the vascular compartment due to arteriolar vasodilation rather than from decreased intravascular volume. According to this theory, portal hypertension results in splanchnic arteriolar vasodilation, mediated by nitric oxide, and leading to underfilling of the arterial vascular space and baroreceptor-mediated stimulation of renin-angiotensin, sympathetic output, and antidiuretic hormone release.

Regardless of the initiating event, a number of factors contribute to accumulation of fluid in the abdominal cavity (Fig. 299-2). Elevated levels of serum epinephrine and norepinephrine have been well documented. *Increased central sympathetic outflow* is found in patients with cirrhosis and ascites but not in those with cirrhosis alone. Increased sympathetic output results in diminished natriuresis by activation of the renin-angiotensin system and diminished sensitivity to atrial natriuretic peptide. *Portal hypertension* plays an important role in the formation of ascites by raising hydrostatic pressure within the splanchnic capillary bed. *Hypoalbuminemia* and *reduced plasma oncotic pressure* also favor the extravasation of fluid from plasma to the peritoneal cavity, and thus ascites is infrequent in patients with cirrhosis unless both portal hypertension and hypoalbuminemia are present. Hepatic lymph may weep freely from the surface of the cirrhotic liver due to distortion and obstruction of hepatic sinusoids and lymphatics and contributes to ascites formation. In contrast to the contribution of transudative fluid from the portal vascular bed, hepatic lymph may weep into the peritoneal cavity even in the absence of marked hypoproteinemia because the endothelial lining of the hepatic sinusoids is discontinuous. This mechanism may account for the high protein concentration present in the ascitic fluid of some patients with venoocclusive disease or the Budd-Chiari syndrome.

Renal factors also play an important role in perpetuating ascites. Patients with ascites fail to excrete a water load in a normal fashion. They have increased renal sodium reabsorption by both proximal and distal tubules, the latter due largely to increased plasma renin activity and secondary hyperaldosteronism. Insensitivity to circulating atrial natriuretic peptide, often present in elevated concentrations in patients with cirrhosis and ascites, may be an important contributory factor in many patients. This insensitivity has been documented in those patients with the most severely impaired sodium excretion, who typically also exhibit low arterial pressure and marked overactivity of the renin-aldosterone axis. Renal vasoconstriction, perhaps resulting from increased serum prostaglandin or catecholamine levels, may also contribute to sodium retention. Recently a role for endothelin, a potent vasoconstrictor peptide, has been proposed. While elevated levels have been reported by some, this has not been observed by others.

As discussed inChap. 46, ascites may arise in a number of clinical settings in addition to cirrhosis and portal hypertension. Although historically ascites was classified as either transudative or exudative, similar to the characterization of pleural fluids, this schema has limitations. Instead, the serum-ascites albumin gradient (SAAG) provides a better classification than total protein content or other parameters. In cirrhosis, the serum albumin concentration is usually at least 10 g/L (1 g/dL) higher than that of the ascitic fluid, thus yielding a high SAAG [³11 g/L (³1.1 g/dL)], reflecting indirectly the abnormally high hydrostatic pressure gradient between the portal bed and the ascitic compartment. Conversely, the presence of a low SAAG [<11 g/L (<1.1 g/dL)] will usually exclude cirrhosis and portal hypertension.

Clinical Features and Diagnosis Usually ascites is first noticed by the patient because of increasing abdominal girth. More pronounced accumulation of fluid may cause shortness of breath because of elevation of the diaphragm. When peritoneal fluid accumulation exceeds 500 mL, ascites may be demonstrated on physical examination by the presence of shifting dullness, a fluid wave, or bulging flanks. Ultrasound examination, preferably with a Doppler study, can detect smaller quantities of ascites and should be performed when physical examination is equivocal or when the cause of the recent onset of ascites is not clear (e.g., exclude Budd-Chiari syndrome or portal vein thrombosis).

TREATMENT

(SeeFig. 299-3) A thorough search should be made for precipitating factors in the patient with recent onset of or worsening ascites, e.g., excessive salt intake, medication noncompliance, superimposed infection, worsening liver disease, portal vein thrombosis, or development of hepatocellular carcinoma. When ascites develops in the setting of severe, acute liver disease, resolution of ascites is likely to follow improvement in liver function. More commonly, ascites develops in patients with stable or steadily worsening liver function. Paracentesis should usually be performed with a small-gauge needle at the time of initial evaluation or at the time of any clinical deterioration of a cirrhotic patient with ascites. A small amount of fluid (<200 mL) should be obtained and examined for evidence of infection, tumor, or other possible causes and complications of ascites. Therapeutic intervention is indicated both to prevent potential complications and to control progressive increase in ascites, which may become pronounced enough to cause physical discomfort. For the patient with a modest accumulation, therapy can be undertaken as an outpatient and should be gentle and incremental (see below). The goal is the loss of no more than 1.0 kg/d if both ascites and peripheral edema are present and no more than 0.5 kg/d in patients with ascites alone. In some patients, particularly those with a large accumulation of fluid, it may be desirable to hospitalize the patient so that daily weights and frequent serum electrolyte levels can be monitored and compliance ensured. Although abdominal girth measurements are frequently used as an index of fluid loss, they tend to be unreliable.

Salt restriction is the cornerstone of therapy. A diet containing 800 mg sodium (2 g NaCl) is often adequate to induce a negative sodium balance and permit diuresis. Response to salt restriction alone is more likely to occur if the ascites is of recent onset, the underlying liver disease is reversible, a precipitating factor can be corrected, or the patient has a high urinary sodium excretion (>25 mmol/d) and normal renal function. Fluid restriction of approximately 1000 mL/d does little to enhance diuresis but may be necessary to correct hyponatremia. If sodium restriction alone fails to result in diuresis and weight loss, diuretics should be prescribed. Because of the role of hyperaldosteronism in sustaining salt retention, spironolactone or other distal tubule-acting diuretics (triamterene, amiloride) are the drugs of choice. These agents are also preferred because of their gentle action and specific potassium-sparing properties. Spironolactone is initially given in a dose of 100 mg a day and is increased as needed by 100 mg/d every several days to a maximum dose that should rarely exceed 400 mg/d. An indication of the minimum effective dose of spironolactone may be obtained by monitoring urinary electrolyte concentrations for a rise in sodium and fall in potassium levels, reflecting effective competitive inhibition of aldosterone. Conversely, the development of azotemia or hyperkalemia may be dose-limiting or even warrant a reduction in the amount of this medication. In some patients, diuresis cannot be initiated despite maximal doses of distal tubule-acting agents (e.g., 400 mg spironolactone) because of avid proximal tubular sodium absorption. More potent and proximally acting diuretics (furosemide, thiazide, or ethacrynic acid) may then be added cautiously to the regimen. Spironolactone plus furosemide, 40 or 80 mg/d, is usually sufficient to initiate a diuresis in most patients. However, such aggressive therapy must be used with great caution to avoid plasma volume depletion, azotemia, and hypokalemia, which may lead to encephalopathy.

In patients with pronounced ascites, particularly those requiring hospitalization, large-volume paracentesis has proven to be an effective and less costly approach to initial management than prolonged bed rest and conventional diuretic treatment. In this approach, ascitic fluid is removed by peritoneal cannula using strict aseptic techniques and monitoring hemodynamic and renal function. This can be safely accomplished in a single session. The need for concomitant albumin replacement by intravenous infusion remains controversial but may be prudent in the patient without peripheral edema, to avoid depleting the intravascular space and precipitating hypotension. Maintenance diuretic therapy in conjunction with sodium restriction may then be instituted to avoid recurrent ascites.

A minority of patients with advanced cirrhosis has "refractory ascites" or rapidly reaccumulate fluid after control by paracentesis. In some patients, a side-to-side *portacaval shunt* may result in improvement in ascites, although generally these patients are extremely poor surgical risks. In the past, intractable ascites has also been treated with the surgical implantation of a plastic *peritoneovenous shunt*, which has a pressure-sensitive, one-way valve allowing ascitic fluid to flow from the abdominal cavity to the superior vena cava. However, the usefulness of this technique is limited by a high rate of complications such as infection, disseminated intravascular coagulation, and thrombosis of the shunt. More recently, in selected patientsTIPS has been used effectively to control refractory ascites, although portal decompression, while mobilizing ascitic fluid, has precipitated severe hepatic encephalopathy in some patients. TIPS remains a promising but unproven treatment for refractory ascites. None of these shunts has been shown to extend life expectancy.

SPONTANEOUS BACTERIAL PERITONITIS (SBP)

Patients with ascites and cirrhosis may develop acute bacterial peritonitis without an obvious primary source of infection. Patients with very advanced liver disease are particularly susceptible to SBP. The ascitic fluid in these patients typically has especially low concentrations of albumin and other so-called opsonic proteins, which normally may provide some protection against bacteria. Although key steps in the pathogenesis of SBP remain to be elucidated, it is clear that most bacteria contributing to SBP derive from the bowel and eventually are spread to ascitic fluid by the hematogenous route after transmigration through the bowel wall and transversing the lymphatics. Clinical features can include abrupt onset of fever, chills, generalized abdominal pain, and, rarely, rebound abdominal tenderness. However, the clinical symptoms *may be minimal*, and some patients manifest only worsening jaundice or encephalopathy in the absence of localizing abdominal complaints. The diagnosis is based on careful examination of the ascitic fluid. An ascitic fluid leukocyte count of>500 cells/L (with a proportion of polymorphonuclear leukocytes of ³50%) or more than 250 polymorphonuclear leukocytes should suggest the possibility of bacterial peritonitis while results of bacterial cultures of ascitic fluid are pending. Other measurements such as fluid pH or determination of gradients between serum and fluid pH or lactate are generally not necessary. The presence of more than 10,000 leukocytes per liter, multiple organisms, or failure to improve after standard therapy for 48 h suggest that the peritonitis may be secondary to an infection elsewhere in the body.

A variant ofSBP, designated *monomicrobial nonneutrocytic bacterascites*, is sometimes seen. In these patients, culture of ascitic fluid yields bacteria, but the neutrophil count is less than 250/L. These patients often have less severe liver disease than those found initially to have typical SBP. While many patients with this variant have cleared the bacterascites at the time of a subsequent paracentesis, nearly 40% will develop typical SBP; thus follow-up paracentesis is usually warranted in this setting.

TREATMENT

Empirical therapy with cefotaxime or ampicillin and an aminoglycoside should be initiated when the diagnosis is first suspected because enteric gram-negative bacilli are found in the majority of cases; less frequently, the infection is caused by pneumococci and other gram-positive bacteria. Cefotaxime is preferable due to the lower rate of renal toxicity. Specific antibiotic therapy can be selected once the specific organism is identified. Therapy is usually administered for 10 to 14 days, although one controlled study has suggested that a 5-day course of intravenous antibiotics may be as effective when repeat paracentesis at 48 h demonstrates a decline in the ascitic polymorphonuclear leukocytes count by more than 50% and negative cultures.

While appropriate antibiotic therapy is usually effective in the treatment of an episode ofSBP, recurrent episodes are relatively common; as many as 70% of patients will experience at least one recurrence within a year of the first episode. The risk of recurrence likely reflects the predisposing role of the underlying advanced liver disease that contributed to the development of the first episode of SBP. Recent trials have demonstrated that prophylactic maintenance therapy with norfloxacin (400 mg/d) can reduce the frequency of recurrent SBP. This agent presumably causes selective decontamination of the intestine, eliminating many aerobic gram-negative bacilli. Trimethoprim-sulfamethoxazole given for 5 days a week has also proven effective. Antibiotics may be administered as infrequently as once a week (e.g., ciprofoxacin, 750 mg once weekly). While maintenance therapy reduces the frequency of SBP and need for hospitalization, it is unclear whether this is associated with prolonged survival. Primary prevention of SBP in a subset of high-risk cirrhotic patients [ascitic fluid protein<10 g/L (<1.0 g/dL)] also appears to be warranted, as is prophylaxis for SBP during variceal hemorrhage.

HEPATORENAL SYNDROME

Definition and Pathogenesis Hepatorenal syndrome is a serious complication in the patient with cirrhosis and ascites and is characterized by worsening azotemia with avid sodium retention and oliguria in the absence of identifiable specific causes of renal dysfunction. The exact basis for this syndrome is not clear, but altered renal hemodynamics appear to be involved. The kidneys are structurally intact; urinalysis and pyelography are usually normal. Renal biopsy, although rarely needed, is also normal, and in fact, kidneys from such patients have been used successfully for renal transplantation. There are indications that an imbalance in certain metabolites of arachidonic acid (prostaglandins and thromboxane) may play a pathogenetic role.

Clinical Features and Diagnosis Worsening azotemia, hyponatremia, progressive oliguria, and hypotension are the hallmarks of the hepatorenal syndrome. This syndrome, which is distinct from prerenal azotemia, may be precipitated by severe gastrointestinal bleeding, sepsis, or overly vigorous attempts at diuresis or paracentesis; it may also occur without an obvious cause. It is essential to exclude other causes of renal impairment often seen in these patients. These include prerenal azotemia or acute tubular necrosis due to hypovolemia (e.g., secondary to gastrointestinal bleeding or diuretic therapy) or an increased nitrogen load such as that seen as a result of bleeding. Drug nephrotoxicity is also often a consideration, particularly in the patient who has received agents such as aminoglycosides or contrast dye. The diagnosis rests on the finding of an elevated serum creatinine level [>133 umol/L (>1.5 g/dL)] that fails to improve with volume expansion or withdrawal of diuretics, together with an unremarkable urine sediment. The diagnosis is supported by the demonstration of avid urinary sodium retention. Typically, the urine sodium concentration is<5 mmol/L, a concentration lower than that generally found in uncomplicated prerenal azotemia.

TREATMENT

Treatment is usually unsuccessful. Although some patients with hypotension and decreased plasma volume may respond to infusions of salt-poor albumin, volume expansion must be undertaken with caution to avoid precipitating variceal bleeding. Vasodilator therapy, including intravenous infusions of low dose dopamine, is not effective. WhileTIPS has been reported to improve renal function in some patients, its use can not be recommended. In appropriate candidates, the treatment of choice for hepatorenal syndrome is liver transplantation.

HEPATIC ENCEPHALOPATHY

Definition Hepatic (portal-systemic) encephalopathy is a complex neuropsychiatric syndrome characterized by disturbances in consciousness and behavior, personality changes, fluctuating neurologic signs, asterixis or "flapping tremor," and distinctive electroencephalographic changes. Encephalopathy may be *acute* and reversible or *chronic* and progressive. In severe cases, irreversible coma and death may occur. Acute episodes may recur with variable frequency.

Pathogenesis The specific cause of hepatic encephalopathy is unknown. The most important factors in the pathogenesis are severe hepatocellular dysfunction and/or intrahepatic and extrahepatic shunting of portal venous blood into the systemic circulation so that the liver is largely bypassed. As a result of these processes, various toxic substances absorbed from the intestine are not detoxified by the liver and lead to metabolic abnormalities in the central nervous system (CNS). *Ammonia* is the substance most often incriminated in the pathogenesis of encephalopathy. Many, but not all, patients with hepatic encephalopathy have elevated blood ammonia levels, and recovery from encephalopathy is often accompanied by declining blood ammonia levels. Other compounds and metabolites that may contribute to the development of encephalopathy include mercaptans (derived from intestinal metabolism of methionine), short-chain fatty acids, and phenol. *False neurochemical transmitters* (e.g., octopamine), resulting in part from alterations in plasma levels of aromatic and branched-chain amino acids, may also play a role. An increase in the permeability of the blood-brain barrier to some of these substances may be an additional factor involved in the pathogenesis of hepatic encephalopathy. Several observations suggest that excessive concentrations of g-aminobutyric acid (GABA), an inhibitory neurotransmitter, in theCNS are important in the reduced levels of consciousness seen in hepatic encephalopathy. Increased CNS GABA may reflect failure of the liver to extract precursor amino acids efficiently or to remove GABA produced in the intestine. In support of this, there is also evidence to suggest that endogenous benzodiazepines, which act through the GABA receptor, may contribute to the development of hepatic encephalopathy. This evidence includes isolation of 1,4-benzodiazepines from brain tissue of patients with fulminant hepatic failure as well as the partial response observed in some patients and experimental animals after administration of flumazenil, a benzodiazepine antagonist. However, the inconsistent effect of flumazenil in patients with encephalopathy, as well as potential methodologic pitfalls in the measurement of endogenous benzodiazepines, preclude definitive attribution of a role to these substances in the pathogenesis of hepatic encephalopathy. The finding of direct enhancement of GABA receptor activation by ammonia suggests that several of the factors described above may be operating via a final common pathway to produce the neuronal depression of hepatic encephalopathy. Finally, the observation of hyperintensity in the basal ganglia by magnetic resonance imaging in cirrhotic patients suggests that excessive *manganese* deposition may also contribute to the pathogenesis of hepatic encephalopathy. Further studies are needed to determine whether chelation therapy exerts long-term benefit.

In the patient with otherwise stable cirrhosis, hepatic encephalopathy often follows a clearly identifiable precipitating event (Table 299-3). Perhaps the most common predisposing factor is *gastrointestinal bleeding*, which leads to an increase in the production of ammonia and other nitrogenous substances, which are then absorbed. Similarly, *increased dietary protein* may precipitate encephalopathy as a result of increased production of nitrogenous substances by colonic bacteria. *Electrolyte disturbances*, particularly hypokalemic alkalosis secondary to overzealous use of diuretics, vigorous paracentesis, or vomiting, may precipitate hepatic encephalopathy. Systemic alkalosis causes an increase in the amount of nonionic ammonia (NH_3) relative to ammonium ions NH_{4+}). Only nonionic (uncharged) ammonia readily crosses the blood-brain barrier and accumulates in theCNS. Hypokalemia also directly stimulates renal ammonia production. Injudicious use of CNS-depressing drugs (e.g., barbiturates, benzodiazepines) and acute infection may trigger or aggravate hepatic encephalopathy, although the mechanisms involved are not clear. Other potential precipitating factors include superimposed acute viral hepatitis, alcoholic hepatitis, extrahepatic bile duct obstruction, constipation, surgery, and other coincidental medical complications.

Hepatic encephalopathy has protean manifestations, and any neurologic abnormality, including focal deficits, may be encountered. In patients with acute encephalopathy, neurologic deficits are completely reversible upon correction of underlying precipitating factors and/or improvement in liver function, but in patients with chronic encephalopathy, the deficits may be irreversible and progressive. Cerebral edema is frequently present and contributes to the clinical picture and overall mortality in patients with both acute and chronic encephalopathy.

The diagnosis of hepatic encephalopathy should be considered when four major factors are present: (1) acute or chronic hepatocellular disease and/or extensive portal-systemic collateral shunts (the latter may be either spontaneous, e.g., secondary to portal hypertension, or surgically created, e.g., portacaval anastomosis); (2) disturbances of awareness and mentation, which may progress from forgetfulness and confusion to stupor and finally coma; (3) shifting combinations of neurologic signs, including asterixis, rigidity, hyperreflexia, extensor plantar signs, and rarely, seizures; and (4) a characteristic (but nonspecific) symmetric, high-voltage, triphasic slow-wave (2 to 5 per second) pattern on the electroencephalogram. Asterixis ("liver flap," "flapping tremor") is a nonrhythmic asymmetric lapse in voluntary sustained position of the extremities, head, and trunk. It is best demonstrated by having the patient extend the arms and dorsiflex the hands. Because elicitation of asterixis depends on sustained voluntary muscle contraction, it is not present in the comatose patient. Asterixis is nonspecific and also occurs in patients with other forms of metabolic brain disease. Disturbances of sleep with reversal of sleep/wake cycles are among the earliest signs of encephalopathy. Alterations in personality, mood disturbances, confusion, deterioration in self-care and handwriting, and daytime somnolence are additional clinical features of encephalopathy. *Fetor hepaticus*, a unique musty odor of the breath and urine believed to be due to mercaptans, may be noted in patients with varying stages of hepatic encephalopathy.

Grading or classifying the stages of hepatic encephalopathy is often helpful in following the course of the illness and assessing response to therapy. One useful classification is shown in Table 299-4.

The diagnosis of hepatic encephalopathy is usually one of exclusion. There are no diagnostic liver function test abnormalities, although an elevated serum ammonia level in the appropriate clinical setting is highly suggestive of the diagnosis. Examination of the cerebrospinal fluid is unremarkable, and computed tomography of the brain shows no characteristic abnormalities until late in stage IV when cerebral edema may supervene. A number of conditions, particularly disorders related to acute and chronic alcoholism, can mimic the clinical features of hepatic encephalopathy. These include acute alcohol intoxication, sedative overdose, delirium tremens, Wernicke's encephalopathy, and Korsakoff's psychosis (Chap. 373). Subdural hematoma, meningitis, and hypoglycemia or other metabolic encephalopathies must also be considered, especially in patients with alcoholic cirrhosis. In young patients with liver disease and neurologic abnormalities, Wilson's disease should be excluded.

TREATMENT

(SeeFig. 299-4) Early recognition and prompt treatment of hepatic encephalopathy are essential. Patients with acute, severe hepatic encephalopathy (stage IV) require the usual supportive measures for the comatose patient. Specific treatment of hepatic encephalopathy is aimed at (1) elimination or treatment of precipitating factors and (2) lowering of blood ammonia (and other toxin) levels by decreasing the absorption of protein and nitrogenous products from the intestine. In the setting of acute gastrointestinal bleeding, blood in the bowel should be promptly evacuated with laxatives (and enemas if necessary) in order to reduce the nitrogen load. Protein should be excluded from the diet, and constipation should be avoided. Ammonia absorption can be decreased by the administration of lactulose, a nonabsorbable disaccharide that acts as an osmotic laxative. Metabolism of lactulose by colonic bacteria may also result in an acid pH that favors conversion of ammonia to the poorly absorbed ammonium ion. In addition, lactulose may actually diminish ammonia production through its direct effects on bacterial metabolism. Acutely, lactulose syrup can be administered in a dose of 30 to 60 mL every hour until diarrhea occurs; thereafter the dose is adjusted (usually 15 to 30 mL three times daily) so that the patient has two to four soft stools daily. Intestinal ammonia production by bacteria can also be decreased by oral administration of a "nonabsorbable" antibiotic such as neomycin (0.5 to 1.0 g every 6 h). However, despite poor absorption, neomycin may reach sufficient concentrations in the bloodstream to cause renal toxicity. Equal benefits may be achieved with broad-spectrum antibiotics such as metronidazole. The use of agents such as levodopa, bromocriptine, keto analogues of essential amino acids, and intravenous amino acid formulations rich in branched-chain amino acids in the treatment of acute hepatic encephalopathy remains of unproven benefit. Flumazenil, a short-acting benzodiazepine antagonist, may have a role in management of hepatic encephalopathy precipitated by use of benzodiazepines, if there is a need for urgent therapy. Hemoperfusion to remove toxic substances and therapy directed primarily toward coincident cerebral edema in acute encephalopathy are also of unproven value. The efficacy of extracorporeal liver assist devices employing hepatocytes of porcine or human origin to bridge patients to recovery or transplantation is as yet unproven but is currently being studied.

Chronic encephalopathy may be effectively controlled by administration of lactulose. Management of patients with chronic encephalopathy should include dietary protein restriction (usually to 60 g/d) in combination with low doses of lactulose or neomycin. Nephrotoxicity or ototoxicity may be limiting in prolonged usage of neomycin. There are suggestions that vegetable protein may be preferable to animal protein.

OTHER SEQUELAE OF CIRRHOSIS

Coagulopathy Patients with cirrhosis often demonstrate a variety of abnormalities in both cellular and humoral clotting function. Thrombocytopenia may result from hypersplenism. In the alcoholic patient, there may be direct bone marrow suppression by ethanol. Diminished protein synthesis may lead to reduced production of fibrinogen (factor I), prothrombin (factor II), and factors V, VII, IX, and X. Reduction in levels of all factors except factor V may be worsened by the coincident malabsorption of the fat-soluble cofactor vitamin K due to cholestasis (Chap. 286). Of these, factor VII appears to be pivotal. In cirrhosis, it is the first of the factors to become depleted and, because of its short half-life, replacement with plasma often fails to correct an elevated prothrombin time. Preliminary studies suggest that selective replacement of factor VII can correct the prothrombin time in patients with cirrhosis.

Hepatocellular Carcinoma SeeChap. 91.

HYPOXEMIA AND HEPATOPULMONARY SYNDROME

Definition and Pathogenesis Mild hypoxemia occurs in approximately one-third of patients with chronic liver disease. The hepatopulmonary syndrome is typically manifest by hypoxemia, platypnea, and orthodeoxia. Hypoxemia usually results from right-to-left intrapulmonary shunts through dilatations in intrapulmonary vessels that can be detected by contrast-enhanced echocardiography or a macroaggregated albumin lung perfusion scan. The mechanisms of shunt formation are unclear, but one animal model suggests that endothelin-1 levels and pulmonary nitric oxide, raised in cirrhosis, correlate with degree of shunting.

TREATMENT

No specific treatment is consistently effective, though large arteriovenous shunts may be embolized. It is now increasingly recognized that liver transplantation may eventually lead to amelioration of the hepatopulmonary syndrome in cases that have not yet been complicated by advanced pulmonary hypertension.

(Bibliography omitted in Palm version)

Back to Table of Contents

300. INFILTRATIVE, GENETIC, AND METABOLIC DISEASES AFFECTING THE LIVER - *Daniel K. Podolsky*

Many disseminated, systemic, or metabolic diseases involve the liver in a diffuse manner by the infiltration of abnormal cells or the accumulation of chemical substances or metabolites. Chemical accumulation may be extracellular or intracellular and may involve hepatocytes, Kupffer cells, or other elements of the reticuloendothelial system. Although infiltrative diseases may vary widely in cause and extrahepatic manifestations, the findings in the liver may be quite similar. Generalized enlargement and firmness of the liver, gradual and nonspecific deterioration of liver function, and, less often, signs of portal hypertension or ascites are typical features of this group of diseases. Differential diagnosis by clinical means may be difficult on occasion, but in patients in whom ancillary clinical findings do not establish the diagnosis, the diffusely infiltrated liver provides an excellent source of tissue for diagnostic purposes.

HEPATIC STEATOSIS (FATTY LIVER) AND NONALCOHOLIC STEATOHEPATITIS

Slight to moderate enlargement of the liver due to a diffuse accumulation of neutral fat (triglycerides) in hepatocytes is an important clinical and pathologic finding. Imaging techniques such as computed tomography (CT), ultrasound, and magnetic resonance imaging (MRI) may each yield alterations suggesting increased fat in the liver. Several mechanisms can contribute to lipid accumulation in the liver. Fatty liver can be separated into two categories based on whether the fat droplets in the hepatocytes are macrovesicular or microvesicular (Table 300-1). In addition, fatty infiltration may be accompanied by necroinflammatory activity, a condition designated *nonalcoholic steatohepatitis (NASH)*.

MACROVESICULAR FATTY LIVER

This is the most common type of fatty liver and is seen most frequently in alcoholism or alcoholic liver disease, diabetes mellitus, obesity, and prolonged parenteral nutrition. Hematoxylin and eosin-stained liver sections show hepatocytes with large, empty vacuoles with the nucleus "pushed" to the periphery of the cell. In general, fat in the liver is not damaging per se, and the fat will disappear with improvement or elimination of the predisposing condition.

Etiology The major causes of fatty liver with macrovesicular fat depend on the age, geographic location, and metabolic-nutritional status of the patient population. *Chronic alcoholism* is the most common cause of hepatic steatosis in this country and in other countries with a high alcohol intake. The severity of fatty involvement is roughly proportional to the duration and degree of alcoholic excess. In addition, in western countriesNASH is associated with obesity. Many of these patients (up to one-third) have type 2 diabetes and/or hyperlipidemia. Inflammatory activity when present may reflect the combined effects of oxidative stress, subsequent lipid peroxidation and abnormal cytokine expression, especially increased tumor necrosis factor (TNF).

Protein malnutrition, especially in infancy and early childhood, accounts for most cases of severe fatty liver in the tropical zones of Africa, South America, and Asia. The hepatic changes may be associated with other clinical and pathologic features of kwashiorkor.

Jejunoileal bypass for surgical treatment of morbid obesity was sometimes associated with severe fatty liver and hepatic failure that could be fatal. In patients with Cushing's syndrome and in those receiving large doses of glucocorticoids, fatty infiltration of the liver may occur. In many *chronic illnesses*, especially those complicated by impaired nutrition or malabsorption, increased fat is found in liver cells. For example, patients with severe ulcerative colitis, chronic pancreatitis, or protracted heart failure frequently have moderate hepatic steatosis at the time of death. Patients maintained on prolonged *total parenteral nutrition* also may develop a fatty liver. In some cases, fatty infiltration and steatohepatitis may occur in the absence of an identifiable cause.

Acute fatty liver is caused by a number of hepatotoxins and is frequently accompanied by signs and symptoms of liver failure. Carbon tetrachloride intoxication, DDT poisoning, and ingestion of substances containing yellow phosphorus result in severe hepatic steatosis. Acute and prolonged alcohol ingestion may also be considered in this category and may be associated with a rapidly enlarging and fat-laden liver.

Clinical Features The signs and symptoms of hepatic steatosis are related to the degree of fat infiltration, the time course of its accumulation, and the underlying cause. The obese or diabetic patient with a chronic fatty liver is usually asymptomatic and has only mild tenderness over the enlarged liver. The liver function tests are normal or show mild elevations of alkaline phosphatase or aminotransferases. In contrast, the rapid accumulation of fat seen in the setting of hyperalimentation may lead to marked tenderness, presumably resulting from stretching of Glisson's capsule. Similarly, alcoholic patients with acute fatty liver following a bout of heavy drinking may have right upper quadrant pain and tenderness, often with laboratory evidence of cholestasis. The clinical presentation of fatty liver from hepatotoxins is similar to that of fulminant hepatic failure arising from any cause, with evidence of hepatic encephalopathy, marked elevations of prothrombin time and aminotransferases, and variable degrees of jaundice. Although steatohepatitis is generally thought to have a benign clinical course with improvement following elimination of the associated precipitant, in some individuals it may result in significant fibrosis and even cirrhosis. Recent studies indicate that substantial fibrosis or cirrhosis may be present in 15 to 50% of patients withNASH. In the only long-term follow-up study, 30% of patients with fibrosis had cirrhosis after 10 years. It is possible that some cases of "cryptogenic" cirrhosis are due to longstanding NASH and that the fat leaves the liver as endstage liver disease develops.

Diagnosis The findings of a firm, nontender, and generally enlarged liver with minimal hepatic dysfunction in a patient with chronic alcoholism, malnutrition, poorly controlled diabetes mellitus, or obesity should suggest hepatic steatosis. This can usually be detected byCT,MRI, or ultrasound. Modest elevations of aminotransferases are often found in association with hepatic steatohepatitis. A disproportional elevation in AST leading to an AST/ALT ratio greater than 2 is generally associated with alcoholic hepatitis. When diagnostic uncertainty exists, needle biopsy of the liver will demonstrate the increased fat content, the presence of any fibrosis, and possibly the underlying primary disorder.

TREATMENT

Adequate nutritional intake, removal of alcohol or offending toxins, and correction of any associated metabolic disorders usually result in recovery. There is no clinical rationale for the use of lipotropic agents such as choline. When indicated, attention should be directed to abstinence from alcohol, careful control of diabetes, weight loss, or correction of intestinal absorptive defects. In the alcoholic fatty liver, there is gradual disappearance of fat from the liver after 4 to 8 weeks of adequate diet and abstinence from alcohol. Similarly, fatty infiltration usually resolves within 2 weeks after discontinuation of parenteral hyperalimentation. Pilot studies in patients withNASHhave suggested benefits from vitamin E and phlebotomy. Troglitazone has shown some benefit in those patients with concomitant insulin resistance.

MICROVESICULAR FATTY LIVER

This is the less common form of fatty liver. On microscopic examination, the fat is present in many small vacuoles. Although the droplets consist of triglycerides in both the macrovesicular and microvesicular forms, the reason for this difference in morphologic appearance is not clear.

Acute fatty liver of pregnancy (AFLP) is a syndrome that occurs late in pregnancy and is often associated with jaundice and hepatic failure. The liver is typically small. AFLP is more common when the mother is carrying a male fetus and may be associated with a deficiency of long-chain-3-hydroxy acyl COH dehydrogenase. Preeclampsia or the HELLP syndrome, which may complicate eclampsia, presents in a similar fashion and progresses to severe liver dysfunction, though typically with a normal size liver. Aminotransferase elevations are typically modest in all of these conditions (generally <500). If diagnosed in time, the disease usually resolves with termination of the pregnancy. Recurrence in subsequent pregnancies is rare.

Microvesicular fat accumulation also may be seen as a toxic reaction to *valproic acid* and with excessive doses of *tetracycline*. It is a typical finding in *Jamaican vomiting sickness*, which is caused by hypoglycin A present in unripened ackee fruit. Lactic acidosis and severe liver injury with microvesicular fat has been described as a complication of nucleoside analogue therapy.

REYE'S SYNDROME (FATTY LIVER WITH ENCEPHALOPATHY)

This acute illness is encountered exclusively in children below 15 years of age. It is characterized clinically by vomiting and signs of progressive central nervous system damage, signs of hepatic injury, and hypoglycemia. Morphologically, there is extensive fatty vacuolization of the liver and renal tubules. There is mitochondrial dysfunction with decreased activity of hepatic mitochondrial enzymes. The cause is unknown, although viral agents and drugs, especially salicylates, have been implicated. Increased aspirin use and much higher serum salicylate levels in children with this illness than in the general population have been described during outbreaks of Reye's syndrome. Recognition of this relationship and reduced aspirin use in this setting may account for the decreasing incidence of Reye's syndrome. However, this illness may occur in the absence of exposure to salicylates. In fatal cases, the liver is enlarged and yellow with striking diffuse fatty microvacuolization of cells. Peripheral zonal hepatic necrosis also has been present in some cases. Fatty changes of the renal tubular cells, cerebral edema, and neuronal degeneration of the brain are the major extrahepatic changes.

Electron-microscopic studies show structural alterations of mitochondria in liver, brain, and muscle.

The onset usually follows an upper respiratory tract infection, especially influenza or chickenpox. Within 1 to 3 days, persistent vomiting occurs, together with stupor, which usually progresses rapidly to generalized convulsions and coma. The liver is enlarged, but *jaundice is characteristically absent or minimal*. Elevations in serum aminotransferases and prothrombin time, hypoglycemia, metabolic acidosis, and elevated serum ammonia levels are the major laboratory findings. The mortality rate in Reye's syndrome is approximately 50%. Therapy consists of infusions of 20% glucose and fresh frozen plasma, as well as intravenous mannitol to reduce the cerebral edema. Chronic liver disease has not been reported in survivors.

STORAGE DISEASES

Lipid storage diseases include the hereditary disorders of Gaucher's and Niemann-Pick disease. Other rare diseases associated with increased fat in the liver include abetalipoproteinemia, Tangier disease, Fabry's disease, and types I and V hyperlipoproteinemia (see Chap. 344 for details). Hepatic enlargement caused by distention of liver cells with glycogen is present in some poorly controlled diabetics and frequently in juvenile diabetes. More often, however, hepatomegaly is due to fatty infiltration (see above). Ketoacidosis and vigorous insulin therapy may further enhance hepatic enlargement.

HEPATIC MINERAL ACCUMULATION

WILSON'S DISEASE

This is an uncommon inherited disorder of copper metabolism. Wilson's disease presents clinically in adolescence or young adulthood by which time there is excess copper accumulation in the liver and other tissues. Deficiency of the plasma copper protein ceruloplasmin is a characteristic feature. The accumulation appears to result from impaired copper excretion due to a mutation in a gene that encodes a P-type ATPase copper transporter. Clinically, patients may present in teenage or early adult years with chronic hepatitis, cirrhosis, or their complications. A small number of patients will present with fulminant hepatitis. Liver disease is often accompanied by softening and degeneration of the basal ganglia (hepatolenticular degeneration) due to copper deposition, which results in extrapyramidal neurologic and psychiatric symptoms. Brownish pigmentation of Descemet's membrane in the cornea (Kayser-Fleischer rings) is frequently present. Hemolytic anemia is also common, especially with fulminant disease. Liver biopsy may reveal findings ranging from fulminant hepatitis to chronic hepatitis and macronodular cirrhosis, in addition to excess copper levels. Typically, liver cells are ballooned and show increased glycogen with glycogen vacuolization in the nuclei. All patients under age 40 with unexplained chronic hepatitis or cirrhosis should be evaluated for possible Wilson's disease. Prompt diagnosis is important; treatment, which must be continued throughout life, can prevent progression of end-organ damage.**For further discussion, see Chap. 348.*

HEMOCHROMATOSIS

Hemochromatosis may be the most common genetic disorder of humans; it involves accumulation of abnormal amounts of iron due to inappropriate absorption from the intestine. Between 85 and 95% of patients with genetic hemochromatosis are homozygous for a point mutation (cystine to tyrosine at codon 282:C282). The liver, as a primary site of iron storage, is affected most directly. There is diffuse deposition of excess iron in hepatocytes, in contrast to the characteristic accumulation of iron in the reticuloendothelial compartment typical of secondary iron overload and hemosiderosis. Excess hepatic iron commonly results in hepatomegaly. Although liver function is initially well preserved, if the disease is untreated, progressive impairment is followed by the development of cirrhosis. Prompt diagnosis can permit the institution of effective lifelong therapy to reduce the iron load and halt progression of the disease._*For further discussion, see Chap. 345._

OTHER INFILTRATIVE AND METABOLIC DISEASES

a_1-ANTITRYPSIN DEFICIENCY

Patients with homozygous deficiency of seruma_1-antitrypsin (a_1AT) are prone to develop emphysema in adult life. The disease is suggested by the absence of alpha$_1$globulin on serum electrophoresis (a_1AT makes up 90% of this fraction normally) and confirmed by direct measurement ofa_1AT. The exact phenotype can then be determined by starch electrophoresis. Although there are approximately 75 recognized alleles, only PiZ and PiS are associated with clinical disease. The molecular bases of these altered products have been related to single nucleic acid substitutions -- e.g., PiZ is caused by a G (guanine) to A (adenine) transposition, which results in a substitution of a glutamic acid for lysine at residue 292 in thea_1AT protein. Hepatocytes of some patients with this deficiency contain globules positive with the periodic acid Schiff reaction. Approximately 10% of children with homozygous deficiency (PiZZ phenotype) ofa_1AT will develop significant liver disease, including neonatal hepatitis and progressive cirrhosis. It has been suggested that 15 to 20 percent of all chronic liver disease in infancy may be attributed toa_1AT deficiency. In adults, the most common manifestation ofa_1AT deficiency is asymptomatic cirrhosis, which may progress from a micronodular to a macronodular state and may be complicated by the development of hepatocellular carcinoma. The occurrence of liver disease in these patients is not dependent on the development of lung disease._*For further discussion, see Chap. 258._

HURLER'S SYNDROME

This is an uncommon hereditary disease that is characterized by the widespread tissue deposition of mucopolysaccharide (chondroitin sulfate B and heparan sulfate) in many tissues. The liver is frequently enlarged and firm. Microscopically, Kupffer cells and other macrophages are enlarged and filled with metachromatic granular material. Cirrhosis may be a late complication._*For further discussion, see Chap. 349._

PORPHYRIAS SeeChap. 346.

RETICULOENDOTHELIAL DISORDERS (See also Chaps. 61 and 113)

Moderate to massive hepatomegaly and splenomegaly occur frequently in the various types of *leukemia* and *lymphoma*. Jaundice, when present, is usually slight and results from hemolysis, although cholestasis may occasionally be associated with lymphoma as a paraneoplastic syndrome. Deep and protracted jaundice is distinctly rare and is caused by obstruction of the intrahepatic or extrahepatic bile ducts by tumor. Liver biopsy specimens reveal portal and sinusoidal infiltrates in most cases of leukemia, but the cellular pattern may be mixed and nonspecific. Liver biopsy is diagnostic in only 5% of patients with *Hodgkin's disease*. This percentage is increased in those with advanced disease or splenomegaly. Directed biopsy at laparoscopy or laparotomy is more likely to be positive than "blind" needle biopsy. Nonspecific histologic changes in the liver have been described in patients with lymphoma and may contribute to the abnormal liver function tests.

Myeloid metaplasia and other myeloproliferative disorders associated with extramedullary hematopoiesis produce hepatomegaly which may reach huge proportions, especially following splenectomy. Serum alkaline phosphatase elevations are often found. Ascites and portal hypertension, resulting from diffuse involvement of portal venules and lymphatics, are rare complications.

GRANULOMATOUS INFILTRATIONS

Perhaps as a result of the large population of mononuclear phagocytes, a number of systemic granulomatous diseases involve the liver, including sarcoidosis, miliary tuberculosis, histoplasmosis, brucellosis, schistosomiasis, berylliosis, and drug reactions (Table 300-2). In addition, isolated granulomas of no diagnostic importance may be found occasionally in patients with various forms of cirrhosis and hepatitis. The liver infiltrated by granulomas may be slightly enlarged and firm, but hepatic dysfunction is usually limited. Increases in serum alkaline phosphatase are common and may range from mild to marked. Occasionally, mild serum elevations in aminotransferases are also present. In a few patients with sarcoidosis or brucellosis, portal hypertension may develop, and extensive postnecrotic scarring or postnecrotic cirrhosis may follow healing of the granulomatous lesions, as in schistosomiasis.

Needle biopsy of the liver often provides the first definite evidence of a systemic or disseminated granulomatous disease. In patients with sarcoidosis who have neither clinical nor laboratory evidence of hepatic involvement, needle biopsy shows sarcoid granulomas in about 80% of cases. In cases of suspected miliary tuberculosis, a portion of the biopsy should be cultured and stained for mycobacteria. The organism can be detected in the majority of cases, particularly when caseating granulomas are present. Serial sections of the biopsy specimen should be examined if granulomas are not apparent. Individual granulomas are rarely specific in their microscopic appearance, and final diagnosis usually requires other clinical, laboratory, or histologic data.

In approximately 20% of patients, it is not possible to identify a cause for the granulomatous infiltration. When these infiltrates are accompanied by fever of unknown origin, the diagnosis of *granulomatous hepatitis* should be considered. This is an uncommon disorder of unknown cause and is diagnosed by exclusion. While granulomatous hepatitis invariably responds to moderate doses of glucocorticoids, relapses are frequent, and such therapy should never be undertaken unless tuberculous disease or other causes of granulomatous infiltration have been excluded. This may include an initial empiric trial of antituberculous therapy.

AMYLOIDOSIS (See alsoChap. 319)

Systemic amyloidosis, whether primary and idiopathic, familial, or secondary to chronic inflammatory or neoplastic diseases, often involves the liver. Grossly, the liver infiltrated with amyloid is enlarged and pale and rubbery in consistency. Microscopically, the birefringent amyloid deposits appear as homogeneous waxy material within the space of Disse, often being concentrated in the periportal areas and associated with atrophy of adjacent liver cell plates. Selective involvement of the walls of blood vessels, especially of the hepatic arterioles, may be a striking feature of primary amyloidosis. With this possible exception, however, the hepatic lesions are the same in all forms of amyloidosis and are present in 60 to 90% of cases.

An enlarged and firm liver is found in about 60% of patients, and ascites occurs in advanced stages of the disease in about 20%. Jaundice, portal hypertension, and other signs of chronic liver disease are usually absent. Liver function changes, although frequent, correlate poorly with the extent of liver infiltration. Hypoalbuminemia and elevated serum alkaline phosphatase are common. Hypoalbuminemia, however, may be related to the presence of nephrosis; the prothrombin time is usually normal. The diagnosis is established by biopsy of rectum, skin, liver, or other involved organs and demonstration of the characteristic Congo red-staining deposits by polarizing microscopy.

AIDS-RELATED LIVER DISEASE

In AIDS, evidence of liver disease is quite common but is usually mild with minimal morbidity. In these patients, hepatic granulomatous disease is often present and may be caused by opportunistic infections, with *Mycobacterium avium-intracellulare* being the most frequent pathogen. Cytomegalovirus hepatitis and hepatic mycoses are less common. These patients are frequently being treated for *Pneumocystis carinii* infections with sulfonamides, which also may cause hepatic granulomatous disease. AIDS cholangiopathy has become a well recognized entity. It exhibits features similar to those found in primary sclerosing cholangitis and is typically associated with cryptosporidia, microsporidia, and/or cytomegalovirus infection in the biliary tract. Papillary stenosis is frequently present. In addition, AIDS patients are vulnerable to hepatic injury resulting from drugs used to treat HIV, most notably nucleoside analogues.

ACKNOWLEDGEMENT

This chapter represents a revised version of a chapter by Dr. Kurt J. Isselbacher and Dr. Daniel K. Podolsky that has appeared in previous editions of this textbook.

(Bibliography omitted in Palm version)

Back to Table of Contents

301. LIVER TRANSPLANTATION - *Jules L. Dienstag*

Liver transplantation -- the replacement of the native, diseased liver by a normal organ (allograft) recovered from a brain-dead donor -- has matured from an experimental procedure reserved for desperately ill patients to an accepted, lifesaving operation applied much earlier in the natural history of end-stage liver disease. The preferred and technically most advanced approach is *orthotopic transplantation*, in which the native organ is removed and the donor organ is inserted in the same anatomic location. Pioneered in the 1960s by Starzl at the University of Colorado and, later, at the University of Pittsburgh and by Calne in Cambridge, England, liver transplantation is now performed routinely by dozens of centers throughout North America and western Europe. Success and survival have improved from approximately 30% in the 1970s to>80% today. These improved prospects for prolonged survival, dating back to the early 1980s, resulted from refinements in operative technique (including the introduction of venovenous bypass to allow venous return from the extremities and visceral circulation during clamping of the inferior vena cava), improvements in organ procurement and preservation, advances in immunosuppressive therapy, and, perhaps most influentially, more enlightened patient selection and timing. Despite the perioperative morbidity and mortality, the technical and management challenges of the procedure, and its costs, liver transplantation has become the approach of choice for selected patients whose chronic or acute liver disease is progressive, life-threatening, and unresponsive to medical therapy. Based on the current level of success, the number of liver transplants has continued to grow each year; in 1999,>4000 patients received liver allografts in the United States. Still, the demand for new livers continues to outpace availability; in 1999, >6000 patients in the United States were on a waiting list for a donor liver.

INDICATIONS

Potential candidates for liver transplantation are children and adults who, in the absence of contraindications (see below), suffer from severe, irreversible liver disease for which alternative medical or surgical treatments have been exhausted or are unavailable. *Timing of the operation is of critical importance*. Indeed, improved timing and better patient selection are felt to have contributed more to the increased success of liver transplantation in the 1980s and beyond than all the impressive technical and immunologic advances combined. Although the disease should be advanced, and although opportunities for spontaneous or medically induced stabilization or recovery should be allowed, the procedure should be done sufficiently early to give the surgical procedure a fair chance for success. Ideally, transplantation should be considered in patients with end-stage liver disease who are experiencing or have experienced a life-threatening complication of hepatic decompensation, whose quality of life has deteriorated to unacceptable levels, or whose liver disease will result predictably in irreversible damage to the central nervous system (CNS). If this is done sufficiently early, the patient will not have developed any contraindications or extrahepatic systemic deterioration. Although patients with well-compensated cirrhosis can survive for many years, many patients with quasi-stable chronic liver disease have much more advanced disease than may be apparent. As discussed below, the better the status of the patient prior to transplantation, the higher will be the anticipated success rate of transplantation. The decision about *when* to transplant is complex and requires the combined judgment of an experienced team of hepatologists, transplant surgeons, anesthesiologists, and specialists in support services, not to mention the well-informed consent of the patient and the patient's family.

Transplantation in Children Indications for transplantation in children are listed inTable 301-1. The most common is *biliary atresia*. *Inherited or genetic disorders of metabolism* associated with liver failure constitute another major indication for transplantation in children and adolescents. In Crigler-Najjar disease type I and in certain hereditary disorders of the urea cycle and of amino acid or lactate-pyruvate metabolism, transplantation may be the only way to prevent impending deterioration ofCNSfunction, despite the fact that the native liver is structurally normal. Combined heart and liver transplantation has yielded dramatic improvement in cardiac function and in cholesterol levels in children with homozygous familial hypercholesterolemia; combined liver and kidney transplantation has been successful in patients with hereditary oxalosis. In hemophiliacs with transfusion-associated hepatitis and liver failure, liver transplantation has been associated with recovery of normal factor VIII synthesis.

Transplantation in Adults Liver transplantation is indicated for end-stage *cirrhosis* of all causes (Table 301-1). In sclerosing cholangitis and *Caroli's disease* (multiple cystic dilatations of the intrahepatic biliary tree), recurrent infections and sepsis associated with inflammatory and fibrotic obstruction of the biliary tree may be an indication for transplantation. Because prior biliary surgery complicates, and is a relative contraindication for, liver transplantation, surgical diversion of the biliary tree has been all but abandoned for patients with sclerosing cholangitis. In patients who undergo transplantation for *hepatic vein thrombosis (Budd-Chiari syndrome)*, postoperative anticoagulation is essential; underlying myeloproliferative disorders may have to be treated but are not a contraindication to liver transplantation. If a donor organ can be located quickly, before life-threatening complications -- including cerebral edema -- set in, patients with *fulminant hepatitis* are candidates for liver transplantation. More controversial as candidates for liver transplantation are patients with *alcoholic cirrhosis*, *chronic viral hepatitis*, and *primary hepatocellular malignancies*. Although all three of these categories are considered to be high risk, liver transplantation can be offered to carefully selected patients. Patients with alcoholic cirrhosis can be considered as candidates for transplantation if they meet strict criteria for abstinence and reform. Patients with chronic hepatitis C have done as well as any other subset of patients after transplantation, despite the fact that recurrent infection in the donor organ is the rule. In patients with chronic hepatitis B, in the absence of measures to prevent recurrent hepatitis B, survival after transplantation is reduced by approximately 10 to 20%; however, prophylactic use of hepatitis B immune globulin (HBIG) during and after transplantation increases the success of transplantation to a level comparable to that seen in patients with nonviral causes of liver decompensation. Specific antiviral drugs, such as lamivudine, that can be used for both prophylaxis against and treatment of recurrent hepatitis B will facilitate further the management of patients undergoing liver transplantation for end-stage hepatitis B. Issues of disease recurrence are discussed in more detail below. Patients with nonmetastatic primary hepatobiliary tumors -- primary hepatocellular carcinoma, cholangiocarcinoma, hepatoblastoma, angiosarcoma, epithelioid hemangioendothelioma, and multiple or massive hepatic adenomata -- have undergone liver transplantation; however, for hepatobiliary malignancies, overall survival is significantly lower than that for other categories of liver disease. To minimize the very high likelihood of recurrent tumor after transplantation, some centers are evaluating experimental adjuvant chemotherapy protocols. Some transplantation centers have reported excellent long-term, recurrence-free survival in patients with unresectable hepatocellular carcinoma for single tumors<5 cm in diameter or for three or fewer lesions all<3 cm. Consequently, most centers restrict liver transplanation to patients whose hepatic malignancies are confined to these limits. Because the likelihood of recurrent cholangiocarcinoma is almost universal, this tumor is no longer considered an indication for transplantation.

CONTRAINDICATIONS

Absolute contraindications for transplantation include life-threatening systemic diseases, uncontrolled extrahepatic bacterial or fungal infections, preexisting advanced cardiovascular or pulmonary disease, multiple uncorrectable life-threatening congenital anomalies, metastatic malignancy, active drug or alcohol abuse, and HIV infection (Table 301-2). Because carefully selected patients in their sixties and even seventies have undergone transplantation successfully, advanced age per se is no longer considered an absolute contraindication; however, in older patients, a more thorough preoperative evaluation should be undertaken to exclude ischemic cardiac disease. Advanced age (>70 years), however, may be considered a *relative contraindication* -- that is, a factor to be taken into account with other relative contraindications. Other relative contraindications include highly replicative hepatitis B, portal vein thrombosis, preexisting renal disease not associated with liver disease, intrahepatic or biliary sepsis, severe hypoxemia resulting from right-to-left intrapulmonary shunts, previous extensive hepatobiliary surgery, and any uncontrolled serious psychiatric disorder. Any one of these relative contraindications is insufficient in and of itself to preclude transplantation. For example, the problem of portal vein thrombosis can be overcome by constructing a graft from the donor liver portal vein to the recipient's superior mesenteric vein.

TECHNICAL CONSIDERATIONS

Donor Selection Donor livers for transplantation are procured primarily from victims of head trauma. Organs from brain-dead donors up to age 60 are acceptable if the following criteria are met: hemodynamic stability; adequate oxygenation; absence of bacterial or fungal infection; serologic exclusion of hepatitis B and C viruses and HIV; absence of abdominal trauma; and absence of hepatic dysfunction. Cardiovascular and respiratory functions are maintained artificially until the liver can be removed. Compatibility in ABO blood group and organ size between donor and recipient are important considerations in donor selection; however, ABO-incompatible or reduced-donor-organ transplants can be performed in emergency or marked donor-scarcity situations. Tissue typing for HLA matching is not required, and preformed cytotoxic HLA antibodies do not preclude liver transplantation. Following perfusion with cold electrolyte solution, the donor liver is removed and packed in ice. The use of University of Wisconsin (UW) solution, rich in lactobionate and raffinose, has permitted the extension of cold ischemic time up to 20 h; however, 12 h may be a more reasonable limit. Improved techniques for harvesting multiple organs from the same donor have increased the availability of donor livers, but the availability of donor livers is far outstripped by the demand. Currently, in the United States, all donor livers are distributed through a nationwide organ-sharing network (United Network of Organ Sharing) designed to allocate available organs based on regional considerations and recipient acuity. Recipients who require the highest level of care (intensive care) have the highest priority, as outlined in Table 301-3.

Surgical Technique Removal of the recipient's native liver is technically difficult, particularly in the presence of portal hypertension with its associated collateral circulation and extensive varices, and even more so in the presence of scarring from previous abdominal operations. The combination of portal hypertension and coagulopathy (elevated prothrombin time and thrombocytopenia) translates into large blood product transfusion requirements. After the portal vein and infrahepatic and suprahepatic inferior vena cavae are dissected, a pump-driven venovenous bypass system is applied to reroute blood from the portal vein and inferior vena cava, preventing congestion of visceral organs. After the hepatic artery and common bile duct are dissected, the native liver is removed and the donor organ inserted. During the anhepatic phase, coagulopathy, hypoglycemia, hypocalcemia, and hypothermia are encountered and must be managed by the anesthesiology team. Caval, portal vein, hepatic artery, and bile duct anastomoses are performed in succession, the last by end-to-end suturing of the donor and recipient common bile ducts or by choledochojejunostomy to a Roux en Y loop if the recipient common bile duct cannot be used for reconstruction (e.g., in sclerosing cholangitis). A typical transplant operation lasts 8 h, with a range of 6 to 18 h. Because of excessive bleeding, large volumes of blood, blood products, and volume expanders may be required during surgery.

Emerging alternatives to orthotopic liver transplantation include split-liver grafts, in which one donor organ is divided and inserted into two recipients; and living-related-donor procedures, in which the left lobe of the liver is harvested from a living-related donor for transplantation into the recipient. Heterotopic liver transplantation, in which the donor liver is inserted without removal of the native liver, has met with very limited success and acceptance, except in a very small number of centers. To support desperately ill patients until a suitable donor organ can be identified, several transplantation centers are studying extracorporeal perfusion with bioartificial liver cartridges constructed from hepatocytes bound to hollow fiber systems and used as temporary hepatic-assist devices, but their efficacy remains to be established. Areas of research with the potential to overcome the shortage of donor organs include hepatocyte transplantation and xenotransplantation with genetically modified organs of nonhuman origin (e.g., swine).

POSTOPERATIVE COURSE AND MANAGEMENT

Immunosuppressive Therapy The introduction in 1980 of cyclosporine as an immunosuppressive agent contributed substantially to the improvement in survival after liver transplantation. Cyclosporine inhibits early activation of T cells and is specific for T cell functions that result from the interaction of the T cell with its receptor and that involve the calcium-dependent signal transduction pathway. As a result, the activity of cyclosporine leads to inhibition of lymphokine gene activation, blocking interleukins 2, 3, and 4, tumor necrosis factor a, as well as other lymphokines. Cyclosporine also inhibits B cell functions. This process occurs without affecting rapidly dividing cells in the bone marrow, which may account for the reduced frequency of posttransplantation systemic infections. The most common and important side effect of cyclosporine therapy is nephrotoxicity. Cyclosporine causes dose-dependent renal tubular injury and direct renal artery vasospasm. Following renal function, therefore, is important in monitoring cyclosporine therapy, perhaps even a more reliable indicator than blood levels of the drug. Nephrotoxicity is reversible and can be managed by dose reduction. Other adverse effects of cyclosporine therapy include hypertension, hyperkalemia, tremor, hirsutism, glucose intolerance, and gum hyperplasia.

Tacrolimus (originally labeled FK 506) is a macrolide lactone antibiotic isolated from a Japanese soil fungus, *Streptomyces tsukubaensis*. It has the same mechanism of action as cyclosporine but is 10 to 100 times more potent. Initially applied as "rescue" therapy for patients in whom rejection occurred despite the use of cyclosporine, tacrolimus has been shown in two large, multicenter, randomized trials to be associated with a reduced frequency of acute rejection, refractory rejection, and chronic rejection. Although patient and graft survival are the same with these two drugs, the advantage of tacrolimus in minimizing episodes of rejection, reducing the need for additional glucocorticoid doses, and reducing the likelihood of bacterial and cytomegalovirus infection has simplified the management of patients undergoing liver transplantation. In addition, the oral absorption of tacrolimus is more predictable than that of cyclosporine, especially during the early postoperative period when T-tube drainage interferes with the enterohepatic circulation of cyclosporine. As a result, in most transplantation centers, tacrolimus has now supplanted cyclosporine for primary immunosuppression, and many centers rely on oral, rather than intravenous, administration from the outset. For transplantation centers that prefer cyclosporine, a new, better-absorbed, microemulsion preparation is now available.

Although tacrolimus is more potent than cyclosporine, it is also more toxic and more likely to be discontinued for adverse events. The toxicity of tacrolimus is similar to that of cyclosporine; nephrotoxicity and neurotoxicity are the most commonly encountered adverse effects, and neurotoxicity (tremor, seizures, hallucinations, psychoses, coma) is more likely and more severe in tacrolimus-treated patients. Both drugs can cause diabetes mellitus, but tacrolimus does not cause hirsutism or gingival hyperplasia. Because of overlapping toxicity between cyclosporine and tacrolimus, especially nephrotoxicity, and because tacrolimus reduces cyclosporine clearance, these two drugs should not be used together. Because 99% of tacrolimus is metabolized by the liver, hepatic dysfunction reduces its clearance; in primary graft nonfunction (when, for technical reasons or because of ischemic damage prior to its insertion, the allograft is defective and does not function normally from the outset) tacrolimus doses have to be reduced substantially, especially in children. Both cyclosporine and tacrolimus are metabolized by the cytochrome P450 IIIA system, and, therefore, drugs that induce cytochrome P450 (e.g., phenytoin, phenobarbital, carbamazepine, rifampin) reduce available levels of cyclosporine and tacrolimus; drugs that inhibit cytochrome P450 (e.g., erythromycin, fluconazole, ketoconazole, clotrimazole, itraconazole, verapamil, diltiazem, nicardipine, cimetidine, danazol, metoclopramide, bromocriptine) increase cyclosporine and tacrolimus blood levels. Like azathioprine, cyclosporine and tacrolimus appear to be associated with a risk of lymphoproliferative malignancies (see below), which may occur earlier after cyclosporine or tacrolimus than after azathioprine therapy. Because of these side effects, combinations of cyclosporine or tacrolimus with prednisone and azathioprine -- all at reduced doses -- are preferable regimens for immunosuppressive therapy.

In patients with pretransplant renal dysfunction or renal deterioration that occurs intraoperatively or immediately postoperatively, tacrolimus or cyclosporine therapy may not be practical; under these circumstances, induction or maintenance of immunosuppression with monoclonal antibodies to T cells, OKT3, may be appropriate. Therapy with OKT3 has been especially effective in reversing acute rejection in the posttransplant period and is the standard treatment for acute rejection that fails to respond to methylprednisolone boluses. Intravenous infusions of OKT3 may be complicated by transient fever, chills, and diarrhea. When this drug is used to induce immunosuppression initially or to provide "rescue" in those who reject despite "conventional" therapy, the incidence of bacterial, fungal, and especially cytomegalovirus infections is increased during and after such therapy. In some centers, ganciclovir antiviral therapy is initiated prophylactically as a routine along with OKT3. Another immunosuppressive drug that is likely to be used in the future for patients undergoing liver transplantation is mycophenolic acid, a nonnucleoside purine metabolism inhibitor derived as a fermentation product from several *Penicillium* species. Mycophenolate has been shown to be better than azathioprine, when used with other standard immunosuppressive drugs, in preventing rejection after renal transplantation and has been approved for use in renal transplantation. Rapamycin, an inhibitor of later events in T cell activation, is yet another drug undergoing experimental evaluation as an immunosuppressive agent.

The most important principle of immunosuppression is that the ideal approach strikes a balance between immunosuppression and immunologic competence. Given sufficient immunosuppression, acute liver allograft rejection is always reversible; however, if the cumulative dose of immunosuppressive therapy is too large, the patient will succumb to opportunistic infection. Therefore, immunosuppressive drugs must be used judiciously, with strict attention to the infectious consequences of such therapy.

Postoperative Complications Complications of liver transplantation can be divided into hepatic and nonhepatic categories (Tables 301-4 and301-5). In addition, both immediately postoperative and late complications are encountered. Patients who undergo liver transplantation as a rule have been chronically ill for protracted periods and may be malnourished and wasted. The impact of such chronic illness and the multisystem failure that accompanies liver failure continues to require attention in the postoperative period. Because of the massive fluid losses and fluid shifts that occur during the operation, patients may remain fluid overloaded during the immediate postoperative period, straining cardiovascular reserve; this effect can be amplified in the face of transient renal dysfunction and pulmonary capillary vascular permeability. Continuous monitoring of cardiovascular and pulmonary function, measures to maintain the integrity of the intravascular compartment and to treat extravascular volume overload, and scrupulous attention to potential sources of and sites of infection are of paramount importance. Cardiovascular instability may also result from the electrolyte imbalance that may accompany reperfusion of the donor liver. Pulmonary function may be compromised further by paralysis of the right hemidiaphragm associated with phrenic nerve injury. The hyperdynamic state with increased cardiac output that is characteristic of patients with liver failure reverses rapidly after successful liver transplantation.

Other immediate management issues include renal dysfunction; prerenal azotemia, acute kidney injury associated with hypoperfusion (acute tubular necrosis), and renal toxicity caused by antibiotics, tacrolimus, or cyclosporine are frequently encountered in the postoperative period, sometimes necessitating dialysis. Occasionally, postoperative intraperitoneal bleeding may be sufficient to increase intraabdominal pressure, which, in turn, may reduce renal blood flow; this effect is rapidly reversible when abdominal distention is relieved by exploratory laparotomy to identify and ligate the bleeding site and to remove intraperitoneal clot. Anemia also may result from acute upper gastrointestinal bleeding or from transient hemolytic anemia, which may be autoimmune, especially when blood group O livers are transplanted into blood group A or B recipients. This autoimmune hemolytic anemia is mediated by donor intrahepatic lymphocytes that recognize red blood cell A or B antigens on recipient erythrocytes. Transient in nature, this process resolves once the donor liver is repopulated by recipient bone marrow-derived lymphocytes; the hemolysis can be treated by transfusing blood group O red blood cells and/or by administering higher doses of glucocorticoids. Transient thrombocytopenia is also commonly encountered. Aplastic anemia, a late occurrence, is rare but has been reported in almost 30% of patients who underwent liver transplantation for acute, severe hepatitis of unknown cause.

Bacterial, fungal, or viral infections are common and may be life-threatening postoperatively. Early after transplant surgery, common postoperative infections predominate -- pneumonia, wound infections, infected intraabdominal collections, urinary tract infections, and intravenous line infections -- rather than opportunistic infections; these infections may involve the biliary tree and liver as well. Beyond the first postoperative month, the toll of immunosuppression becomes evident, and opportunistic infections -- cytomegalovirus, herpes viruses, fungal infections (*Aspergillus*, *Candida*, cryptococcal disease), mycobacterial infections, parasitic infections (*Pneumocystis*, *Toxoplasma*), bacterial infections (*Nocardia*, *Legionella*, and *Listeria*) -- predominate. Rarely, early infections represent those transmitted with the donor liver, either infections present in the donor or infections acquired during procurement processing. De novo viral hepatitis infections acquired from the donor organ or from transfused blood products occur after typical incubation periods for these agents (well beyond the first month). Obviously, infections in an immunosuppressed host demand early recognition and prompt management; prophylactic antibiotic therapy is administered routinely in the immediate postoperative period. Use of sulfamethoxazole with trimethoprim reduces the incidence of postoperative *Pneumocystis carinii* pneumonia.

Neuropsychiatric complications include seizures (commonly associated with cyclosporine and tacrolimus toxicity), encephalopathy, depression, and difficult psychosocial adjustment. Rarely, diseases are transmitted by the allograft from the donor to the recipient. In addition to viral and bacterial infections, malignancies of donor origin have occurred. Lymphoproliferative malignancies, especially B cell lymphoma, are a recognized complication associated with immunosuppressive drugs such as azathioprine, tacrolimus, and cyclosporine (see above). Epstein-Barr virus has been shown to play a contributory role in some of these tumors, which may regress when immunosuppressive therapy is reduced.

Hepatic Complications Hepatic dysfunction after liver transplantation is similar to the hepatic complications encountered after major abdominal and cardiothoracic surgery;

however, in addition, there may be complications such as primary graft failure, vascular compromise, failure or obstruction of the biliary anastomoses, and rejection. As in nontransplant surgery, postoperative jaundice may result from prehepatic, intrahepatic, and posthepatic sources. *Prehepatic* sources represent the massive hemoglobin pigment load from transfusions, hemolysis, hematomas, ecchymoses, and other collections of blood. *Early intrahepatic* liver injury includes effects of hepatotoxic drugs and anesthesia; hypoperfusion injury associated with hypotension, sepsis, and shock; and benign postoperative cholestasis. *Late intrahepatic* sources of liver injury include posttransfusion hepatitis and exacerbation of primary disease. *Posthepatic* sources of hepatic dysfunction include biliary obstruction and reduced renal clearance of conjugated bilirubin. Hepatic complications unique to liver transplantation include primary graft failure associated with ischemic injury to the organ during harvesting; vascular compromise associated with thrombosis or stenosis of the portal vein or hepatic artery anastomoses; vascular anastomotic leak; stenosis, obstruction, or leakage of the anastomosed common bile duct; recurrence of primary hepatic disorder (see below); and rejection.

Transplant Rejection Despite the use of immunosuppressive drugs, rejection of the transplanted liver still occurs in a majority of patients, beginning 1 to 2 weeks after surgery. Clinical signs suggesting rejection are fever, right upper quadrant pain, and reduced bile pigment and volume. Leukocytosis may occur, but the most reliable indicators are increases in serum bilirubin and aminotransferase levels. Because these tests lack specificity, distinguishing among rejection and biliary obstruction, primary graft nonfunction, vascular compromise, viral hepatitis, cytomegalovirus infection, drug hepatotoxicity, and recurrent primary disease may be difficult. Radiographic visualization of the biliary tree and/or percutaneous liver biopsy often help to establish the correct diagnosis. Morphologic features of acute rejection include portal infiltration, bile duct injury, and/or endothelial inflammation ("endothelialitis"); some of these findings are reminiscent of graft-versus-host disease and primary biliary cirrhosis. As soon as transplant rejection is suspected, treatment consists of intravenous methylprednisolone in repeated boluses; if this fails to abort rejection, many centers use antibodies to lymphocytes, such as OKT3, or polyclonal antilymphocyte globulin.

Chronic rejection is a relatively rare outcome that may follow repeated bouts of acute rejection or that occurs unrelated to preceding rejection episodes. Morphologically, chronic rejection is characterized by progressive cholestasis, focal parenchymal necrosis, mononuclear infiltration, vascular lesions (intimal fibrosis, subintimal foam cells, fibrinoid necrosis), and fibrosis. This process may be reflected as ductopenia -- the vanishing bile duct syndrome. Some of the histologic hallmarks of chronic rejection may be so similar to those of chronic viral hepatitis that differentiation between the two may be difficult. Reversibility of chronic rejection is limited; in patients with therapy-resistant chronic rejection, retransplantation has yielded encouraging results.

OUTCOME

Survival The survival rate for patients undergoing liver transplantation has improved steadily since 1983. One-year survival rates have increased from approximately 70% in the early 1980s to 80 to 90% in the late 1990s. Currently, the 5-year survival rate exceeds 60%. An important observation is the relation between clinical status before transplantation and outcome. For patients who undergo liver transplantation when their level of compensation is high (e.g., still working or only partially disabled), a 1-year survival rate of 85% is common. For those whose level of decompensation mandates continuous in-hospital care prior to transplantation, the 1-year survival rate is about 70%, while for those who are so decompensated that they require life support in an intensive care unit, the 1-year survival rate is approximately 50%. Indeed, the trend toward transplantation earlier in the natural history of end-stage liver disease is a major factor in the increased success of liver transplantation during the 1980s and 1990s. Another important distinction in survival has been drawn between high-risk and low-risk patient categories. For patients who do not fit any "high-risk" designations, 1-year and 5-year survival rates of 85 and 80%, respectively, have been recorded. In contrast, among patients in high-risk categories -- cancer, fulminant hepatitis, hepatitis B, age >65, concurrent renal failure, respirator dependence, portal vein thrombosis, and history of a portacaval shunt or multiple right upper quadrant operations -- survival statistics fall into the range of 60% at 1 year and 35% at 5 years. Survival after retransplantation for primary graft nonfunction is approximately 50%. Causes of failure of liver transplantation vary with time. Failures within the first 3 months result primarily from technical complications, postoperative infections, and hemorrhage. Transplant failures after the first 3 months are more likely to result from infection, rejection, or recurrent disease (such as malignancy or viral hepatitis).

Recurrence of Primary Disease The recurrence of autoimmune hepatitis or primary sclerosing cholangitis has not been reported. There have been reports of recurrent primary biliary cirrhosis after liver transplantation; however, the histologic features of primary biliary cirrhosis and acute rejection are virtually indistinguishable and occur as frequently in patients with primary biliary cirrhosis as in patients undergoing transplantation for other reasons. Hereditary disorders such as Wilson's disease anda$_1$-antitrypsin deficiency have not recurred after liver transplantation; however, recurrence of disordered iron metabolism has been observed in some patients with hemochromatosis. Hepatic vein thrombosis (Budd-Chiari syndrome) may recur; this can be minimized by treating underlying lymphoproliferative disorders and by anticoagulation. Cholangiocarcinoma recurs almost invariably; therefore, few centers now transplant such patients. In patients with hepatocellular carcinoma, tumor recurrence in the liver is common after approximately 1 year, although better success has been reported in patients with an unresectable isolated lesion <5 cm or with three or fewer lesions all<3 cm. Trials are underway to assess the benefit of adjuvant chemotherapy.

Hepatitis A can recur after transplantation for fulminant hepatitis A, but such acute reinfection has no serious clinical sequelae. In fulminant hepatitis B, recurrence is not the rule; however, in the absence of any prophylactic measures, hepatitis B usually recurs after transplantation for end-stage chronic hepatitis B. With sufficient immunosuppressive therapy to prevent allograft rejection, levels of hepatitis B viremia increase markedly, regardless of pretransplantation values. A majority of patients undergoing transplantation for chronic hepatitis B become carriers of hepatitis B virus (HBV) with high levels of virus replication but without liver injury; however, some patients experience a rapid recapitulation of severe injury -- severe chronic hepatitis or even fulminant hepatitis -- after transplantation. *Fibrosing cholestatic hepatitis* is a histologic feature linked to rapidly progressive liver injury in approximately 10% of patients who undergo liver transplantation for hepatitis B. These patients experience marked hyperbilirubinemia, substantial prolongation of the prothrombin time (both out of proportion to relatively modest elevations of aminotransferase activity), and rapidly progressive liver failure. This lesion has been suggested to represent a "choking off" of the hepatocyte by an overwhelming density of HBV proteins. Complications such as sepsis and pancreatitis have also been observed more frequently in patients undergoing liver transplantation for hepatitis B. Although the risk of recurrent hepatitis B is approximately 20% higher in patients with pretransplantation markers of HBV replication (hepatitis B e antigen and HBV DNA), recurrent hepatitis B occurs in at least 60% of patients whose replicative markers were undetectable prior to transplantation, probably because of the enhancing impact of immunosuppressive drugs on HBV replication. Most transplantation centers will not undertake liver transplantation in patients with hepatitis B unless immunoprophylaxis withHBIG is used. Neither preoperative hepatitis B vaccination, preoperative or postoperative interferon therapy, nor short-term (£2 months) HBIG prophylaxis has been shown to be effective, but a retrospective analysis of data from several hundred European patients followed for 3 years after transplantation has shown that long-term (³6 months) prophylaxis with HBIG is associated with a lowering of the risk of HBV reinfection from approximately 75% to 35% and a reduction in mortality from approximately 50 percent to 20%.

As a result of long-termHBIG use following liver transplantation for chronic hepatitis B, similar improvements in outcome have been observed in the United States, with 1-year survival rates between 75 and 90%. Currently, with HBIG prophylaxis, the outcome of liver transplantation for chronic hepatitis B is indistinguishable from that for chronic liver disease unassociated with chronic hepatitis B; essentially, medical concerns regarding liver transplantation for chronic hepatitis B have been eliminated. Passive immunoprophylaxis with HBIG is begun during the anhepatic stage of surgery, repeated daily for the first 6 postoperative days, then continued with infusions that are given either at regular intervals of 4 to 6 weeks or, alternatively, when anti-HBs levels fall below a threshold of 100 mIU/mL. In all likelihood, indefinite HBIG infusions will be required, and, occasionally "breakthrough"HBVinfection occurs. This approach is very expensive, approximately $20,000 per year, and involves the intravenous administration of a globulin preparation designed for intramuscular injection. Although this approach is now practiced universally, it has not been approved officially by the U.S. Food and Drug Administration; clinical trials of HBIG preparations produced specifically for intravenous administration are in progress.

An alternative and promising but still experimental approach to the prophylaxis of patients with chronic hepatitis B undergoing liver transplantation is the use of nucleoside analogues such as lamivudine (Chap. 297). Limited evidence available to date suggests that lamivudine can be used to prevent recurrence ofHBVinfection when administered *prior* to transplantation, to treat hepatitis B that recurs *after* transplantation, including in patients who break throughHBIGprophylaxis, and to reverse the course of otherwise fatal fibrosing cholestatic hepatitis. Clinical trials have shown that lamivudine monotherapy reduces the level of HBV replication substantially, sometimes even resulting in clearance of HBsAg; reduces ALT levels; and improves histologic features of necrosis and inflammation. Long-term use of lamivudine is safe and effective, but, after several months, a proportion of patients become resistant to lamivudine, resulting from "YMDD" mutations in the HBV polymerase motif (Chap. 297). In approximately half of such resistant patients, hepatic deterioration may ensue. Perhaps the best results with currently available antiviral approaches can be achieved by combining HBIG and lamivudine. In addition, new nucleoside analogues and related drugs are being assessed as antiviral agents against HBV infection. Some of these are effective against lamivudine-associated YMDD variants of HBV; these novel agents also are likely to be used in patients undergoing liver transplantation. Clinical trials are underway to define the optimal application of these antiviral agents in the management of patients undergoing liver transplantation for chronic hepatitis B.

Patients who undergo liver transplantation for chronic hepatitis B plus D have a better survival rate than patients undergoing transplantation for hepatitis B alone. Accounting for up to 40% of all liver transplantation procedures, the most common indication for liver transplantation is end-stage liver disease resulting from chronic hepatitis C. Recurrence of hepatitis C virus (HCV) after liver transplantation can be documented in almost every patient, if sufficiently sensitive virus markers are used. Although acute and chronic liver injury occur after transplantation in patients with chronic hepatitis C, clinical consequences of recurrent hepatitis C are limited during the first 5 years after transplantation. Nonetheless, despite the relative clinical benignity of recurrent hepatitis C in the early years after liver transplantation, and despite the negligible impact on patient survival during these early years, histologic studies have documented the presence of moderate to severe chronic hepatitis in more than half of all patients and bridging fibrosis or cirrhosis in approximately 10%. Ultimately, such histologic evidence of chronic hepatitis and cirrhosis will be expressed clinically as well, and the expectation is that the 10-year outcome will not be as favorable as the 5-year statistics suggest. In a proportion of patients, even during the early posttransplantation period, recurrent hepatitis C may be sufficiently severe biochemically and histologically to merit antiviral therapy. Treatment with interferon monotherapy, which can *suppress* HCV-associated liver injury in approximately half of patients but rarely leads to *sustained* benefit, has been disappointing. The addition of the nucleoside analogue ribavirin to interferon has resulted in improved responses to antiviral therapy, and many centers have adopted some form of combination therapy for their patients with recurrent chronic hepatitis C; however, the efficacy of such combination therapy remains the subject of clinical trials. Of interest is the preliminary observation that the immunosuppressive agent mycophenolate may have a suppressive effect on HCV. A small number succumb to early HCV-associated liver injury, and a syndrome reminiscent of fibrosing cholestatic hepatitis (see above) has been observed rarely. Because patients with more episodes of rejection receive more immunosuppressive therapy, and because immunosuppressive therapy enhances HCV replication, patients with severe or multiple episodes of rejection are more likely to experience early recurrence of hepatitis C after transplantation. Both HCV genotype 1b and high viral load have been linked to recurrent HCV-induced liver disease and to earlier disease recurrence after transplantation; however, the association between genotype and recurrence of HCV-associated liver injury has not been supported by more recent reports.

Patients who undergo liver transplantation for end-stage alcoholic cirrhosis are at risk of resorting to drinking again after transplantation, a potential source of recurrent alcoholic liver injury. Currently, alcoholic liver disease is one of the more common indications for liver transplantation, accounting for 20 to 25% of all liver transplantation procedures, and most transplantation centers screen candidates carefully for predictors of continued abstinence. Recidivism is more likely in patients whose sobriety prior to transplantation was shorter than 6 months. For abstinent patients with alcoholic cirrhosis, liver transplantation can be undertaken successfully, with outcomes comparable to those for other categories of patients with chronic liver disease, when coordinated by a team approach that includes substance abuse counseling.

Posttransplantation Quality of Life Full rehabilitation is achieved in the majority of patients who survive the early postoperative months and escape chronic rejection or unmanageable infection. Psychosocial maladjustment interferes with medical compliance in a small number of patients, but most manage to adhere to immunosuppressive regimens, which must be continued indefinitely. In one study, 85% of patients who survived their transplants returned to gainful activities. In fact, some women have conceived and carried pregnancies to term after transplantation without demonstrable injury to their infants.

(Bibliography omitted in Palm version)

Back to Table of Contents

302. DISEASES OF THE GALLBLADDER AND BILE DUCTS - *Norton J. Greenberger*, *Gustav Paumgartner*

PHYSIOLOGY OF BILE PRODUCTION AND FLOW

Bile Secretion and Composition Bile formed in the hepatic lobules is secreted into a complex network of canaliculi, small bile ductules, and larger bile ducts that run with lymphatics and branches of the portal vein and hepatic artery in portal tracts situated between hepatic lobules. These interlobular bile ducts coalesce to form larger septal bile ducts that join to form the right and left hepatic ducts, which in turn unite to form the common hepatic duct. The common hepatic duct is joined by the cystic duct of the gallbladder to form the common bile duct (CBD), which enters the duodenum (often after joining the main pancreatic duct) through the ampulla of Vater.

Hepatic bile is an isotonic fluid with an electrolyte composition resembling blood plasma. The electrolyte composition of gallbladder bile differs from that of hepatic bile because most of the inorganic anions, chloride and bicarbonate, have been removed by reabsorption across the gallbladder epithelium.

Major components of bile by weight include water (82%), bile acids (12%), lecithin and other phospholipids (4%), and unesterified cholesterol (0.7%). Other constituents include conjugated bilirubin, proteins (IgA, metabolites of hormones, and other proteins metabolized in the liver), electrolytes, mucus, and, often, drugs and their metabolites.

The total daily basal secretion of hepatic bile is approximately 500 to 600 mL. Many substances taken up or synthesized by the hepatocyte are secreted into the bile canalculi. The canalicular membrane forms microvilli and is associated with microfilaments of actin, microtubules, and other contractile elements. Prior to their secretion into the bile, many substances that are taken up into the hepatocyte are conjugated, while others such as phospholipids, a portion of primary bile acids, and some cholesterol are synthesized de novo in the hepatocyte. Three mechanisms are important in regulating bile flow: (1) active transport of bile acids from hepatocytes into the bile canaliculi, (2) active transport of other organic anions, and (3) cholangiocellular secretion. The last is a secretin-mediated and cyclic AMP-dependent mechanism that ultimately results in the secretion of a sodium- and bicarbonate-rich fluid into the bile ducts.

Active vectorial secretion of biliary constituents from the portal blood into the bile canaliculi is driven by a distinct set of polarized transport systems at the basolateral (sinusoidal) and the canalicular plasma membrane domains of the hepatocyte. Two sinusoidal bile salt uptake systems have been cloned in humans, the Na^+/taurocholate cotransporter and the organic anion transporting protein, which also transports a large variety of non-bile salt organic anions. Four ATP-dependent canalicular transport systems ("export pumps") have been identified: a bile salt export pump (BSEP), which was formerly called "sister of P-glycoprotein"; a conjugate export pump (MRP2), also called the canalicular multispecific organic anion transporter, which mediates the canalicular excretion of various amphiphilic conjugates formed by phase II conjugation (e.g., bilirubin diglucuronide); a multidrug export pump (MDR1) for hydrophobic cationic compounds; and a phospholipid export pump (MDR3). The canalicular membrane also contains ATP-independent transport systems such as the Cl^-/HCO_3^-anion exchanger isoform 2 for canalicular bicarbonate secretion. For some of these transporters, genetic defects have been identified that are associated with various forms of cholestasis or defects of biliary excretion. BSEP is defective in progressive familial intrahepatic cholestasis (PFIC) type 2. Mutations of MRP2 cause the Dubin-Johnson syndrome, an inherited form of conjugated hyperbilirubinemia. A defective MDR3 results in PFIC-3. The cystic fibrosis transmembrane regulator located on bile duct epithelial cells is defective in cystic fibrosis, which may be associated with impaired cholangiocellular bile formation and chronic cholestatic liver disease.

The Bile Acids The primary bile acids, cholic acid and chenodeoxycholic acid (CDCA), are synthesized from cholesterol in the liver, conjugated with glycine or taurine, and excreted into the bile. Secondary bile acids, including deoxycholate and lithocholate, are formed in the colon as bacterial metabolites of the primary bile acids. However, lithocholic acid is much less efficiently absorbed from the colon than deoxycholic acid. Another secondary bile acid, found in low concentration is ursodeoxycholic acid (UDCA), a stereoisomer of CDCA. In normal bile, the ratio of glycine to taurine conjugates is about 3:1.

Bile acids are detergents that in aqueous solutions and above a critical concentration of about 2 m*M* form molecular aggregates called *micelles*. Cholesterol alone is poorly soluble in aqueous environments, and its solubility in bile depends on both the total lipid concentration and the relative molar percentages of bile acids and lecithin. Normal ratios of these constituents favor the formation of solubilizing *mixed micelles*, while abnormal ratios promote the precipitation of cholesterol crystals in bile.

In addition to facilitating the biliary excretion of cholesterol, bile acids are necessary for the normal intestinal absorption of dietary fats via a micellar transport mechanism (Chap. 286). Bile acids also serve as a major physiologic driving force for hepatic bile flow and aid in water and electrolyte transport in the small bowel and colon.

Enterohepatic Circulation Bile acids are efficiently conserved under normal conditions. Unconjugated, and to a lesser degree also conjugated, bile acids are absorbed by *passive diffusion* along the entire gut. Quantitatively much more important for bile salt recirculation, however, is the *active transport* mechanism for conjugated bile acids in the distal ileum (Chap. 286). The reabsorbed bile acids enter the portal bloodstream and are taken up rapidly by hepatocytes, reconjugated, and resecreted into bile (enterohepatic circulation).

The normal bile acid pool size is approximately 2 to 4 g. During digestion of a meal, the bile acid pool undergoes at least one or more enterohepatic cycles, depending on the size and composition of the meal. Normally, the bile acid pool circulates approximately 5 to 10 times daily. Intestinal absorption of the pool is about 95% efficient, so fecal loss of bile acids is in the range of 0.3 to 0.6 g/d. This fecal loss is compensated by an equal daily synthesis of bile acids by the liver, and thus the size of the bile acid pool is maintained. Bile acids returning to the liver suppress de novo hepatic synthesis of primary bile acids from cholesterol by inhibiting the rate-limiting enzyme cholesterol 7a-hydroxylase. While the loss of bile salts in stool is usually matched by increased hepatic synthesis, the maximum rate of synthesis is approximately 5 g/d, which may be insufficient to replete the bile acid pool size when there is pronounced impairment of intestinal bile salt reabsorption.

Gallbladder and Sphincteric Functions In the fasting state, the sphincter of Oddi offers a high-pressure zone of resistance to bile flow from the common bile duct into the duodenum. This tonic contraction serves to (1) prevent reflux of duodenal contents into the pancreatic and bile ducts and (2) promote bile filling of the gallbladder. The major factor controlling the evacuation of the gallbladder is the peptide hormone cholecystokinin (CCK), which is released from the duodenal mucosa in response to the ingestion of fats and amino acids.CCKproduces (1) powerful contraction of the gallbladder, (2) decreased resistance of the sphincter of Oddi, (3) increased hepatic secretion of bile, and thus (4) enhanced flow of biliary contents into the duodenum.

Hepatic bile is "concentrated" within the gallbladder by energy-dependent transmucosal absorption of water and electrolytes. Almost the entire bile acid pool may be sequestered in the gallbladder following an overnight fast for delivery into the duodenum with the first meal of the day. The normal capacity of the gallbladder is 30 to 50 mL of bile.

DISEASES OF THE GALLBLADDER

CONGENITAL ANOMALIES

Anomalies of the biliary tract may be found in 10 to 20% of the population, including abnormalities in number, size, and shape (e.g., agenesis of the gallbladder, duplications, rudimentary or oversized "giant" gallbladders, and diverticula). Phrygian cap is a clinically innocuous entity in which a partial or complete septum (or fold) separates the fundus from the body. Anomalies of position or suspension are not uncommon and include left-sided gallbladder, intrahepatic gallbladder, retrodisplacement of the gallbladder, and "floating" gallbladder. The latter condition predisposes to acute torsion, volvulus, or herniation of the gallbladder.

GALLSTONES

Pathogenesis Gallstones are quite prevalent in most western countries. In the United States, autopsy series have shown gallstones in at least 20% of women and in 8% of men over the age of 40. It is estimated that 16 to 20 million persons in the United States have gallstones and that approximately 1 million new cases of cholelithiasis develop each year.

Gallstones are formed by concretion or accretion of normal or abnormal bile constituents. They are divided into three major types; cholesterol and mixed stones account for 80% of the total, with pigment stones comprising the remaining 20%. Mixed and cholesterol gallstones usually contain more than 50% cholesterol monohydrate plus an admixture of calcium salts, bile pigments, proteins, and fatty acids. Pigment stones are composed primarily of calcium bilirubinate; they contain less than 20% cholesterol.

Cholesterol and Mixed Stones and Biliary Sludge Cholesterol is essentially water insoluble and requires aqueous dispersion into either micelles or vesicles, both of which require the presence of a second lipid to "liquefy" the cholesterol. Cholesterol and phospholipids are secreted into bile as unilamellar bilayered vesicles, which are converted into mixed micelles consisting of bile acids, phospholipids, and cholesterol by the action of bile acids. If there is an excess of cholesterol in relation to phospholipids and bile acids, unstable cholesterol-rich vesicles remain, which aggregate into large multilamellar vesicles from which cholesterol crystals precipitate (Fig. 302-1).

There are several important mechanisms in the formation of lithogenic (stone-forming) bile. The most important is increased biliary secretion of cholesterol. This may occur in association with obesity, high-caloric and cholesterol-rich diets, or drugs (e.g., clofibrate) and may result from increased activity of HMG-CoA reductase, the rate-limiting enzyme of hepatic cholesterol synthesis, and increased hepatic uptake of cholesterol from blood. In patients with gallstones, dietary cholesterol *increases* biliary cholesterol secretion. This does not occur in non-gallstone patients on high-cholesterol diets. In addition to environmental factors such as high-caloric and cholesterol-rich diets, genetic factors play an important role in cholesterol hypersecretion and gallstone formation. A high prevalence of gallstones is found among first-degree relatives of gallstone carriers and in certain ethnic populations such as American Indians as well as Chilean Indians and Chilean Hispanics. A common genetic trait has been identified for some of these populations by mitochondrial DNA analysis. A genetic defect in the control of cholesterol secretion also exists in certain strains of inbred mice who develop gallstones under a lithogenic diet. In some patients, impaired hepatic conversion of cholesterol to bile acids may also occur, resulting in an increase of the lithogenic cholesterol/bile acid ratio. Lithogenic bile may also result from conditions affecting the enterohepatic circulation of bile acids (e.g., prolonged parenteral alimentation or ileal disease or resection). In addition, most patients with gallstones may have reduced activity of hepatic cholesterol 7a-hydroxylase, the rate-limiting enzyme for primary bile acid synthesis.

Thus an excess of biliary cholesterol in relation to bile acids and phospholipids is primarily due to hypersecretion of cholesterol, but hyposecretion of bile acids may contribute. Two additional disturbances of bile acid metabolism that are likely to contribute to supersaturation of bile with cholesterol are (1) reduction of the bile acid pool and (2) enhanced conversion of cholic acid to deoxycholic acid, with replacement of the cholic acid pool by an expanded deoxycholic acid pool. The first disorder may be caused by more rapid loss of primary bile acid from the small intestine into the colon. The second disturbance may result from enhanced dehydroxylation of cholic acid and increased absorption of newly formed deoxycholic acid. An increased deoxycholate secretion is associated with hypersecretion of cholesterol into bile. While supersaturation of bile with cholesterol is an important prerequisite for gallstone formation, it is not sufficient by itself to produce cholesterol precipitation in vivo. Most people with supersaturated bile do not develop stones because the time required for cholesterol crystals to nucleate and grow is longer than the time bile spends in the gallbladder.

A second important mechanism is *nucleation* of cholesterol monohydrate crystals, which is greatly accelerated in human lithogenic bile; it is this feature rather than the degree of cholesterol supersaturation that distinguishes lithogenic from normal gallbladder bile. Accelerated nucleation of cholesterol monohydrate in bile may be due to either an *excess of pronucleating factors* or a *deficiency of antinucleating factors*. Mucin and certain non-mucin glycoproteins appear to be pronucleating factors, while apolipoproteins AI and AII and other glycoproteins appear to be antinucleating factors. Cholesterol monohydrate crystal nucleation and crystal growth probably occur within the mucin gel layer. Vesicle fusion leads to liquid crystals, which, in turn, nucleate into solid cholesterol monohydrate crystals. Continued growth of the crystals occurs by direct nucleation of cholesterol molecules from supersaturated unilamellar or multilamellar biliary vesicles.

A third important mechanism in cholesterol gallstone formation is *gallbladder hypomotility*. If the gallbladder emptied all supersaturated or crystal-containing bile completely, stones would not be able to grow. A high percentage of patients with gallstones exhibits abnormalities of gallbladder emptying. Ultrasonographic studies show that gallstone patients have an increased gallbladder volume during fasting and also after a test meal (residual volume) and that fractional emptying after gallbladder stimulation is decreased. Gallbladder emptying is a major determinant of gallstone recurrence in patients who underwent biliary lithotripsy. Within 3 years, only 13% of patients with good but 53% of patients with poor gallbladder emptying form recurrent stones.

Biliary sludge is a thick mucous material that upon microscopic examination reveals lecithin-cholesterol crystals, cholesterol monohydrate crystals, calcium bilirubinate, and mucin thread or mucous gels. Biliary sludge typically forms a crescent-like layer in the most dependent portion of the gallbladder and is recognized by characteristic echoes on ultrasonography (see below). The presence of biliary sludge implies two abnormalities: (1) the normal balance between gallbladder mucin secretion and elimination has become deranged and (2) nucleation of biliary solutes has occurred. That biliary sludge may be a precursor form of gallstone disease is evident from several observations. In one study, 96 patients with gallbladder sludge were followed prospectively by serial ultrasound studies. In 18%, biliary sludge disappeared and did not recur for at least 2 years. In 60%, biliary sludge disappeared and reappeared; in 14%, gallstones (8% asymptomatic, 6% symptomatic) developed, and in 6%, severe biliary pain with or without acute pancreatitis occurred. In 12 patients, cholecystectomies were performed, 6 for gallstone-associated biliary pain and 3 in symptomatic patients with sludge but without gallstones who had prior attacks of pancreatitis; the latter did not recur after cholecystectomy. It should be emphasized that biliary sludge can develop with disorders that cause gallbladder hypomotility, i.e., surgery, burns, total parenteral nutrition, pregnancy, and oral contraceptives -- all of which are associated with gallstone formation.

Two other conditions are associated with cholesterol stone or biliary sludge formation: pregnancy and very low calorie diet. There appear to be two key changes during pregnancy that contribute to a "cholelithogenic state." First, the composition of the bile acid pool and the cholesterol-carrying capacity of bile change, with a resultant marked increase in cholesterol saturation during the third trimester. Second, ultrasonographic studies have demonstrated that gallbladder contraction in response to a standard meal is sluggish, resulting in impaired gallbladder emptying. That these changes are related to pregnancy per se is supported by several studies that show reversal of these abnormalities after delivery. During pregnancy, gallbladder sludge develops in 20 to 30% of women and gallstones in 5 to 12%. While biliary sludge is a common finding during pregnancy, it is usually asymptomatic and often resolves spontaneously after delivery. Gallstones, which are less common than sludge and frequently associated with biliary colic, may also disappear after delivery because of spontaneous dissolution related to bile becoming unsaturated with cholesterol post partum.

From 10 to 20% of people having rapid weight reduction through very low calorie dieting develop gallstones. In a study involving 600 patients who completed a 16-week, 520-kcal/d diet,UDCA in a dosage of 600 mg/d proved highly effective in preventing gallstone formation; gallstones developed in only 3% of UDCA recipients compared to 28% of placebo-treated patients.

To summarize, cholesterol gallstone disease occurs because of several defects, which include (1) bile supersaturation with cholesterol, (2) nucleation of cholesterol monohydrate with subsequent crystal retention and stone growth, and (3) abnormal gallbladder motor function with delayed emptying and stasis. Other important factors known to predispose to cholesterol stone formation are summarized inTable 302-1.

Pigment Stones Gallstones composed largely of calcium bilirubinate are much more common in the Far East than in western countries. The presence of increased amounts of unconjugated, insoluble bilirubin in bile results in the precipitation of bilirubin, which may aggregate to form pigment stones or may form the nidus for growth of mixed cholesterol gallstones. In western countries, chronic hemolytic states (with increased conjugated bilirubin in bile) or alcoholic liver disease are associated with an increased incidence of pigment stones. Deconjugation of an excess of soluble bilirubin mono- and diglucuronide may be mediated by endogenous b-glucuronidase but may also occur by spontaneous alkaline hydrolysis. Sometimes, the enzyme is also produced when bile is chronically infected by bacteria. Pigment stone formation is especially prominent in Asians and is often associated with infections in the biliary tree (Table 302-1).

Diagnosis Procedures of potential use in the diagnosis of cholelithiasis and other diseases of the gallbladder are detailed inTable 302-2. The plain abdominal film may detect gallstones containing sufficient calcium to be radiopaque (10 to 15% of cholesterol and mixed stones and approximately 50% of pigment stones). Plain radiography may also be of use in the diagnosis of emphysematous cholecystitis, porcelain gallbladder, limey bile, and gallstone ileus.

Ultrasonography of the gallbladder is very accurate in the identification of cholelithiasis and has several advantages over oral cholecystography (Fig. 302-2*A*). The gallbladder is easily visualized with the technique, and in fact, failure to image the gallbladder successfully in a fasting patient correlates well with the presence of underlying gallbladder disease. Stones as small as 2 mm in diameter may be confidently identified provided that firm criteria are used [e.g., acoustic "shadowing" of opacities that are within the gallbladder lumen and that change with the patient's position (by gravity)]. In major medical centers, the false-negative and false-positive rates for ultrasound in gallstone patients are about 2 to 4%. Biliary sludge is material of low echogenic activity that typically forms a layer in the most dependent position of the gallbladder. This layer shifts with postural changes but fails to produce acoustic shadowing; these two characteristics distinguish sludges from gallstones. Ultrasound can also be used to assess the emptying function of the gallbladder.

Oral cholecystography (OCG) is a useful procedure for the diagnosis of gallstones but has been largely replaced by ultrasound. However, OCG is still useful for the selection of patients for nonsurgical therapy of gallstone disease such as lithotripsy or bile acid dissolution therapy. In both these settings, OCG is used to assess the patency of the cystic duct and gallbladder emptying function. Further, OCG can also delineate the size and number of gallstones and determine whether they are calcified. Factors that may produce nonvisualization of the OCG are summarized in Table 302-2.

Radiopharmaceuticals such as$_{99m}$Tc-labeled *N*-substituted iminodiacetic acids (HIDA, DIDA, DISIDA, etc.) are rapidly extracted from the blood and are excreted into the biliary tree in high concentration even in the presence of mild to moderate serum bilirubin elevations. Failure to image the gallbladder in the presence of biliary ductal visualization may indicate cystic duct obstruction, acute or chronic cholecystitis, or surgical absence of the organ. Such scans have their greatest application in the diagnosis of acute cholecystitis.

Symptoms of Gallstone Disease Gallstones usually produce symptoms by causing inflammation or obstruction following their migration into the cystic duct orCBD. The most specific and characteristic symptom of gallstone disease is biliary colic. Obstruction of the cystic duct or CBD by a stone produces increased intraluminal pressure and distention of the viscus that cannot be relieved by repetitive biliary contractions. The resultant visceral pain is characteristically a severe, steady ache or pressure in the epigastrium or right upper quadrant (RUQ) of the abdomen with frequent radiation to the interscapular area, right scapula, or shoulder.

Biliary colic begins quite suddenly and may persist with severe intensity for 30 min to 5 h, subsiding gradually or rapidly. An episode of biliary pain is sometimes followed by a residual mild ache or soreness in theRUQ, which may persist for 24 h or so. Nausea and vomiting frequently accompany episodes of biliary colic. An elevated level of serum bilirubin and/or alkaline phosphatase suggests a common duct stone. Fever or chills (rigors) with biliary colic usually imply a complication, i.e., cholecystitis, pancreatitis, or cholangitis. Complaints of vague epigastric fullness, dyspepsia, eructation, or flatulence, especially following a fatty meal, should not be confused with biliary colic. Such symptoms are frequently elicited from patients with gallstone disease but are not specific for biliary calculi. Biliary colic may be precipitated by eating a fatty meal, by consumption of a large meal following a period of prolonged fasting, or by eating a normal meal.

Natural History Gallstone disease discovered in an asymptomatic patient or in a patient whose symptoms are not referable to cholelithiasis is a common clinical problem. The natural history of "silent" or asymptomatic gallstones has occasioned much debate. A study of predominantly male silent gallstone patients suggests that the cumulative risk for the development of symptoms or complications requiring surgery is relatively low -- 10% at 5 years, 15% at 10 years, and 18% at 15 years. Patients remaining asymptomatic for 15 years were found to be unlikely to develop symptoms during further follow-up, and most patients who did develop complications from their gallstones experienced *prior* warning symptoms. Similar conclusions apply to diabetic patients with silent gallstones. Decision analysis has suggested that (1) the cumulative risk of death due to gallstone disease while on expectant management is small, and (2) prophylactic cholecystectomy is not warranted.

Complications requiring cholecystectomy are much more common in gallstone patients who have developed symptoms of biliary colic. Patients found to have gallstones at a young age are more likely to develop symptoms from cholelithiasis than are patients older than 60 years at the time of initial diagnosis. Patients with diabetes mellitus and gallstones may be somewhat more susceptible to septic complications, but the magnitude of risk of septic biliary complications in diabetic patients is incompletely defined. In addition, asymptomatic gallstone patients with nonvisualization of the gallbladder onOCGappear to have an increased tendency to develop symptoms and complications.

TREATMENT

Surgical Therapy In asymptomatic gallstone patients, the risk of developing symptoms or complications requiring surgery is quite small (in the range of 1 to 2% per year). Thus a recommendation for cholecystectomy in a patient with gallstones should probably be based on assessment of three factors: (1) the presence of symptoms that are frequent enough or severe enough to interfere with the patient's general routine; (2) the presence of a prior complication of gallstone disease, i.e., history of acute cholecystitis, pancreatitis, gallstone fistula, etc.; or (3) the presence of an underlying condition predisposing the patient to increased risk of gallstone complications (e.g., calcified or porcelain gallbladder and/or a previous attack of acute cholecystitis regardless of current symptomatic status). Patients with very large gallstones (over 2 cm in diameter) and patients having gallstones in a congenitally anomalous gallbladder might also be considered for prophylactic cholecystectomy. Although age under 50 years is a worrisome factor in asymptomatic gallstone patients, few authorities would now recommend routine cholecystectomy in all young patients with silent stones. Laparoscopic cholecystectomy is a minimal-access approach for the removal of the gallbladder together with its stones. Its advantages include a markedly shortened hospital stay as well as decreased cost, and it is the procedure of choice for most patients referred for elective cholecystectomy.

From several studies involving over 4000 patients undergoing laparoscopic cholecystectomy, the following key points emerge: (1) complications develop in about 4% of patients, (2) conversion to laparotomy occurs in 5%, (3) the death rate is remarkably low (i.e.,<0.1%), and (4) bile duct injuries are unusual (i.e., 0.2 to 0.5%). These data indicate why laparoscopic cholecystectomy has become the "gold standard" for treating symptomatic cholelithiasis.

Medical Therapy -- Gallstone Dissolution UDCAdecreases cholesterol saturation of bile and also appears to produce a lamellar liquid crystalline phase in bile that allows a dispersion of cholesterol from stones by physiochemical means. UDCA may also retard cholesterol crystal nucleation. In carefully selected patients with a functioning gallbladder and with radiolucent stones <10 mm in diameter, complete dissolution can be achieved in about 50% of patients within 6 months to 2 years with UDCA at a dose of 8 to 10 mg/kg per day. The highest success rate (i.e., >70%) occurs in patients with small (<5 mm) floating radiolucent gallstones. Probably no more than 10% of patients with *symptomatic* cholelithiasis are candidates for such treatment. However, in addition to the vexing problem of recurrent stones (30 to 50% over 3 to 5 years of follow-up), there is also the factor of taking an expensive drug for an indefinite period of time. The advantages and success of laparoscopic cholecystectomy have largely reduced the role of gallstone dissolution to patients who wish to avoid or are not candidates for elective cholecystectomy.

Gallbladder stones may be fragmented by extracorporeal shock waves. While such shock wave lithotripsy combined with medical litholytic therapy is safe and effective in carefully selected patients with gallbladder calculi (radiolucent, solitary stone<2 cm in well-contracting gallbladder), the procedure is employed infrequently because of the emergence of laparoscopic cholystectomy as the procedure of choice for symptomatic cholelithiasis, the recurrence of gallstones in 30% of patients within 5 years after lithotripsy combined with medical litholytic therapy, and the cost of takingUDCA for a variable period after the procedure.

ACUTE AND CHRONIC CHOLECYSTITIS

Acute Cholecystitis Acute inflammation of the gallbladder wall usually follows obstruction of the cystic duct by a stone. Inflammatory response can be evoked by three factors: (1) *mechanical inflammation* produced by increased intraluminal pressure and distention with resulting ischemia of the gallbladder mucosa and wall, (2) *chemical inflammation* caused by the release of lysolecithin (due to the action of phospholipase on lecithin in bile) and other local tissue factors, and (3) *bacterial inflammation*, which may play a role in 50 to 85% of patients with acute cholecystitis. The organisms most frequently isolated by culture of gallbladder bile in these patients include *Escherichia coli*, *Klebsiella* spp., group D *Streptococcus*, *Staphylococcus* spp., and *Clostridium* spp.

Acute cholecystitis often begins as an attack of biliary colic that progressively worsens. Approximately 60 to 70% of patients report having experienced prior attacks that resolved spontaneously. As the episode progresses, however, the pain of acute cholecystitis becomes more generalized in the right upper abdomen. As with biliary colic, the pain of cholecystitis may radiate to the interscapular area, right scapula, or shoulder. Peritoneal signs of inflammation such as increased pain with jarring or on deep respiration may be apparent. The patient is anorectic and often nauseated. Vomiting is relatively common and may produce symptoms and signs of vascular and extracellular volume depletion. Jaundice is unusual early in the course of acute cholecystitis but may occur when edematous inflammatory changes involve the bile ducts and surrounding lymph nodes.

A low-grade fever is characteristically present, but shaking chills or rigors are not uncommon. TheRUQ of the abdomen is almost invariably tender to palpation. An enlarged, tense gallbladder is palpable in one-quarter to one-half of patients. Deep inspiration or cough during subcostal palpation of the RUQ usually produces increased pain and inspiratory arrest (Murphy's sign). A light blow delivered to the right subcostal area may elicit a marked increase in pain. Localized rebound tenderness in the RUQ is common, as are abdominal distention and hypoactive bowel sounds from paralytic ileus, but generalized peritoneal signs and abdominal rigidity are usually lacking, absent perforation.

The diagnosis of acute cholecystitis is usually made on the basis of a characteristic history and physical examination. The triad of sudden onset ofRUQtenderness, fever, and leukocytosis is highly suggestive. Typically, leukocytosis in the range of 10,000 to 15,000 cells per microliter with a left shift on differential count is found. The serum bilirubin is mildly elevated [<85.5 umol/L (5 mg/dL)] in 45% of patients, while 25% have modest elevations in serum aminotransferases (usually less than a fivefold elevation). The radionuclide (e.g., HIDA) biliary scan may be confirmatory if bile duct imaging is seen without visualization of the gallbladder. Ultrasound will demonstrate calculi in 90 to 95% of cases.

Approximately 75% of patients treated medically have remission of acute symptoms within 2 to 7 days following hospitalization. In 25%, however, a complication of acute cholecystitis will occur despite conservative treatment (see below). In this setting, prompt surgical intervention is required. Of the 75% of patients with acute cholecystitis who undergo remission of symptoms, approximately one-quarter will experience a recurrence of cholecystitis within 1 year, and 60% will have at least one recurrent bout within 6 years. In view of the natural history of the disease, acute cholecystitis is best treated by early surgery whenever possible.

Acalculous Cholecystitis In 5 to 10% of patients with acute cholecystitis, calculi obstructing the cystic duct are not found at surgery. In over 50% of such cases, an underlying explanation for acalculous inflammation is not found. An increased risk for the development of acalculous cholecystitis is especially associated with serious trauma or burns, with the postpartum period following prolonged labor, and with orthopedic and other nonbiliary major surgical operations in the postoperative period. Other precipitating factors include vasculitis, obstructing adenocarcinoma of the gallbladder, diabetes mellitus, torsion of the gallbladder, "unusual" bacterial infections of the gallbladder (e.g., *Leptospira*, *Streptococcus*, *Salmonella*, or *Vibrio cholerae*), and parasitic infestation of the gallbladder. Acalculous cholecystitis may also be seen with a variety of other systemic disease processes (sarcoidosis, cardiovascular disease, tuberculosis, syphilis, actinomycosis, etc.) and may possibly complicate periods of prolonged parenteral hyperalimentation.

Although the clinical manifestations of acalculous cholecystitis are indistinguishable from those of calculous cholecystitis, the setting of acute gallbladder inflammation complicating severe underlying illness is characteristic of acalculous disease. Ultrasound, computed tomography (CT) scanning, or radionuclide examinations demonstrating a large, tense, static gallbladder without stones and with evidence of poor emptying over a prolonged period may be diagnostically useful in some cases. The complication rate for acalculous cholecystitis exceeds that for calculous cholecystitis. Successful management of acute acalculous cholecystitis appears to depend primarily on early diagnosis and surgical intervention, with meticulous attention to postoperative care.

Acalculous Cholecystopathy Disordered motility of the gallbladder can produce recurrent biliary pain in patients without gallstones. Infusion of an octapeptide ofCCK can be used to measure the gallbladder ejection fraction during cholescintigraphy. In a representative study, CCK cholescintigraphy using$_{99m}$Tc-diisopropyl iminodiacetic acid

(DIDA) identified 21 patients with an abnormal gallbladder ejection fraction (<40% at 45 min); 10 of 11 patients who underwent surgery became asymptomatic; all 10 showed abnormalities, i.e., chronic cholecystitis, gallbladder muscle hypertrophy, and/or a markedly narrowed cystic duct. From this and other similar studies, the following criteria can be used to identify patients with acalculous cholecystopathy: (1) recurrent episodes of typicalRUQ pain characteristic of biliary tract pain, (2) abnormal CCK cholescintigraphy demonstrating a gallbladder ejection fraction of less than 40%, and (3) infusion of CCK reproduces the patient's pain. An additional clue would be the identification of a large gallbladder on ultrasound examination. Finally, it should be noted that sphincter of Oddi dysfunction can also give rise to recurrent RUQ pain and CCK-scintigraphic abnormalities.

Emphysematous Cholecystitis So-called emphysematous cholecystitis is thought to begin with acute cholecystitis (calculous or acalculous) followed by ischemia or gangrene of the gallbladder wall and infection by gas-producing organisms. Bacteria most frequently cultured in this setting include anaerobes, such as *C. welchii* or *C. perfringens*, and aerobes, such as *E. coli.* This condition occurs most frequently in elderly men and in patients with diabetes mellitus. The clinical manifestations are essentially indistinguishable from those of nongaseous cholecystitis. The diagnosis is usually made on plain abdominal film by the finding of gas within the gallbladder lumen, dissecting within the gallbladder wall to form a gaseous ring, or in the pericholecystic tissues. The morbidity and mortality rates with emphysematous cholecystitis are considerable. Prompt surgical intervention coupled with appropriate antibiotics is mandatory.

Chronic Cholecystitis Chronic inflammation of the gallbladder wall is almost always associated with the presence of gallstones and is thought to result from repeated bouts of subacute or acute cholecystitis or from persistent mechanical irritation of the gallbladder wall. The presence of bacteria in the bile occurs in more than one-quarter of patients with chronic cholecystitis. Although the presence of infected bile in a patient with *chronic* cholecystitis undergoing elective cholecystectomy probably adds little to the operative risk, intraoperative Gram's staining and routine culturing of bile have been advocated to identify those patients whose gallbladder is colonized with *Clostridium* spp. Appropriate antibiotics intra- and postoperatively are recommended in such patients because colonization with these organisms may be associated with devastating septic complications following surgery. Chronic cholecystitis may be asymptomatic for years, may progress to symptomatic gallbladder disease or to acute cholecystitis, or may present with complications (see below).

Complications of Cholecystitis

Empyema and Hydrops Empyema of the gallbladder usually results from progression of acute cholecystitis with persistent cystic duct obstruction to superinfection of the stagnant bile with a pus-forming bacterial organism. The clinical picture resembles that of cholangitis with high fever, severeRUQpain, marked leukocytosis, and often, prostration. Empyema of the gallbladder carries a high risk of gram-negative sepsis and/or perforation. Emergency surgical intervention with proper antibiotic coverage is required as soon as the diagnosis is suspected.

Hydrops or mucocele of the gallbladder may also result from prolonged obstruction of the cystic duct, usually by a large solitary calculus. In this instance, the obstructed gallbladder lumen is progressively distended, over a period of time, by mucus (mucocele) or by a clear transudate (hydrops) produced by mucosal epithelial cells. A visible, easily palpable, nontender mass sometimes extending from theRUQ into the right iliac fossa may be found on physical examination. The patient with hydrops of the gallbladder frequently remains asymptomatic, although chronic RUQ pain may also occur. Cholecystectomy is indicated, since empyema, perforation, or gangrene may complicate the condition.

Gangrene and Perforation Gangrene of the gallbladder results from ischemia of the wall and patchy or complete tissue necrosis. Underlying conditions often include marked distention of the gallbladder, vasculitis, diabetes mellitus, empyema, or torsion resulting in arterial occlusion. Gangrene usually predisposes to perforation of the gallbladder, but perforation may also occur in chronic cholecystitis without premonitory warning symptoms. *Localized perforations* are usually contained by the omentum or by adhesions produced by recurrent inflammation of the gallbladder. Bacterial superinfection of the walled-off gallbladder contents results in abscess formation. Most patients are best treated with cholecystectomy, but some seriously ill patients may be managed with cholecystostomy and drainage of the abscess. *Free perforation* is less common but is associated with a mortality rate of approximately 30%. Such patients may experience a sudden transient relief ofRUQ pain as the distended gallbladder decompresses; this is followed by signs of generalized peritonitis.

Fistula Formation and Gallstone Ileus Fistulization into an adjacent organ adherent to the gallbladder wall may result from inflammation and adhesion formation. Fistulas into the duodenum are most common, followed in frequency by those involving the hepatic flexure of the colon, stomach or jejunum, abdominal wall, and renal pelvis. Clinically "silent" biliary-enteric fistulas occurring as a complication of chronic cholecystitis have been found in up to 5% of patients undergoing cholecystectomy. Asymptomatic cholecystoenteric fistulas may sometimes be diagnosed by finding gas in the biliary tree on plain abdominal films. Barium contrast studies or endoscopy of the upper gastrointestinal tract or colon may demonstrate the fistula. Treatment in the symptomatic patient usually consists of cholecystectomy,CBDexploration, and closure of the fistulous tract.

Gallstone ileus refers to mechanical intestinal obstruction resulting from the passage of a large gallstone into the bowel lumen. The stone customarily enters the duodenum through a cholecystoenteric fistula at that level. The site of obstruction by the impacted gallstone is usually at the ileocecal valve, provided that the more proximal small bowel is of normal caliber. The majority of patients do not give a history of either prior biliary tract symptoms or complaints suggestive of acute cholecystitis or fistulization. Large stones over 2.5 cm in diameter are thought to predispose to fistula formation by gradual erosion through the gallbladder fundus. Diagnostic confirmation may occasionally be found on the plain abdominal film (e.g., small-intestinal obstruction with gas in the biliary tree and a calcified, ectopic gallstone) or following an upper gastrointestinal series (cholecystoduodenal fistula with small-bowel obstruction at the ileocecal valve). Laparotomy with stone extraction (or propulsion into the colon) remains the procedure of choice to relieve obstruction. Evacuation of large stones within the gallbladder should also be performed. In general, the gallbladder and its attachment to the intestines should be left alone.

Limey (Milk of Calcium) Bile and Porcelain Gallbladder Calcium salts may be secreted into the lumen of the gallbladder in sufficient concentration to produce calcium precipitation and diffuse, hazy opacification of bile or a layering effect on plain abdominal roentgenography. This so-called limey bile, or milk of calcium bile, is usually clinically innocuous, but cholecystectomy is recommended because limey bile most often occurs in a hydropic gallbladder. In the entity called *porcelain gallbladder*, calcium salt deposition within the wall of a chronically inflamed gallbladder may be detected on the plain abdominal film. Cholecystectomy is advised in all patients with porcelain gallbladder because in a high percentage of cases this finding appears to be associated with the development of carcinoma of the gallbladder.

TREATMENT

Medical Therapy Although surgical intervention remains the mainstay of therapy for acute cholecystitis and its complications, a period of in-hospital stabilization may be required before cholecystectomy. Oral intake is eliminated, nasogastric suction may be indicated, and extracellular volume depletion and electrolyte abnormalities are repaired. Meperidine or nonsteroidal antiinflammatory drugs (NSAIDs) are usually employed for analgesia because they may produce less spasm of the sphincter of Oddi than drugs such as morphine. Intravenous antibiotic therapy is usually indicated in patients with severe acute cholecystitis even though bacterial superinfection of bile may not have occurred in the early stages of the inflammatory process. Postoperative complications of wound infection, abscess formation, or sepsis are reduced in antibiotic-treated patients. Effective antibiotics include ureidopenicillins, ampicillin, metronidazole, and cephalosporins. Combination with an aminoglycoside or other antibiotics may be considered in diabetic or debilitated patients and in those with signs of gram-negative sepsis (Chap. 134).

Surgical Therapy The optimal timing of surgical intervention in patients with acute cholecystitis depends on stabilization of the patient. The clear trend is toward earlier surgery, and this is due in part to requirements for shorter hospital stays. Urgent (emergency) cholecystectomy or cholecystostomy is probably appropriate in most patients in whom a complication of acute cholecystitis such as empyema, emphysematous cholecystitis, or perforation is suspected or confirmed. In uncomplicated cases of acute cholecystitis, up to 30% of patients fail to resolve their symptoms on appropriate medical therapy, and progression of the attack or a supervening complication leads to the performance of early operation (within 24 to 72 h). The technical complications of surgery are not increased in patients undergoing early as opposed to delayed cholecystectomy. Delayed surgical intervention is probably best reserved for (1) patients in whom the overall medical condition imposes an unacceptable risk for early surgery and (2) patients in whom the diagnosis of acute cholecystitis is in doubt. Early cholecystectomy is the treatment of choice for most patients with acute cholecystitis. Mortality figures for emergency cholecystectomy in most centers approach 3%, while the mortality risk for elective or early cholecystectomy approximates 0.5% in patients under age 60. Of course, the operative risks increase with age-related diseases of other organ systems and with the presence of long- or short-term complications of gallbladder disease. Seriously ill or debilitated patients with cholecystitis may be managed with cholecystostomy and tube drainage of the gallbladder. Elective cholecystectomy may then be done at a later date.

Postcholecystectomy Complications Early complications following cholecystectomy include atelectasis and other pulmonary disorders, abscess formation (often subphrenic), external or internal hemorrhage, biliary-enteric fistula, and bile leaks. Jaundice may indicate absorption of bile from an intraabdominal collection following a biliary leak or mechanical obstruction of theCBD by retained calculi, intraductal blood clots, or extrinsic compression. Routine performance of intraoperative cholangiography during cholecystectomy has helped to reduce the incidence of these early complications.

Overall, cholecystectomy is a very successful operation that provides total or near-total relief of preoperative symptoms in 75 to 90% of patients. The most common cause of persistent postcholecystectomy symptoms is an overlooked extrabiliary disorder (e.g., reflux esophagitis, peptic ulceration, pancreatitis, or -- most often -- irritable bowel syndrome). In a small percentage of patients, however, a disorder of the extrahepatic bile ducts may result in persistent symptomatology. These so-called postcholecystectomy syndromes may be due to (1) biliary strictures, (2) retained biliary calculi, (3) cystic duct stump syndrome, (4) stenosis or dyskinesia of the sphincter of Oddi, or (5) bile salt-induced diarrhea or gastritis.

Cystic Duct Stump Syndrome In the absence of cholangiographically demonstrable retained stones, symptoms resembling biliary colic or cholecystitis in the postcholecystectomy patient have frequently been attributed to disease in a long (>1 cm) cystic duct remnant (cystic duct stump syndrome). Careful analysis, however, reveals that postcholecystectomy complaints are attributable to other causes in almost all patients in whom the symptom complex was originally thought to result from the existence of a long cystic duct stump. Accordingly, considerable care should be taken to investigate the possible role of other factors in the production of postcholecystectomy symptoms before attributing them to cystic duct stump syndrome.

Papillary dysfunction, papillary stenosis, spasm of the sphincter of Oddi, and biliary dyskinesia Symptoms of biliary colic accompanied by signs of recurrent, intermittent biliary obstruction may be produced by papillary stenosis, papillary dysfunction, spasm of the sphincter of Oddi, and biliary dyskinesia. Papillary stenosis is thought to result from acute or chronic inflammation of the papilla of Vater or from glandular hyperplasia of the papillary segment. Five criteria have been used to define papillary stenosis: (1) upper abdominal pain, usuallyRUQ or epigastric; (2) abnormal liver tests; (3) dilatation of the common bile duct upon endoscopic retrograde cholangiopancreatography (ERCP) examination; (4) delayed (>45 min) drainage of contrast material from the duct; and (5) increased basal pressure of the sphincter of Oddi, a finding that may be of only minor significance. An alternative to ERCP is magnetic resonance cholangiography if ERCP and/or biliary manometry are either unavailable or not feasible. In patients with papillary stenosis, quantitative hepatobiliary scintigraphy has revealed delayed transit from the common bile duct to the bowel, ductal dilatation, and abnormal time-activity dynamics. This technique can also be used before and after sphincterotomy to document improvement in biliary emptying. Treatment consists of endoscopic or surgical sphincteroplasty to ensure wide patency of the distal portions of both the bile and pancreatic ducts. The greater the number of the preceding criteria present, the greater the likelihood that a patient does have a degree of papillary stenosis sufficient to justify correction. The factors usually considered as indications for sphincterotomy include (1) prolonged duration of symptoms, (2) lack of response to symptomatic treatment, (3) presence of severe disability, and (4) the patient's choice of sphincterotomy over surgery (given a clear understanding on his or her part of the risks involved in both procedures).

Criteria for diagnosing dyskinesia of the sphincter of Oddi are even more controversial than those for papillary stenosis. Proposed mechanisms include spasm of the sphincter, denervation sensitivity resulting in hypertonicity, and abnormalities of the sequencing or frequency rates of sphincteric contraction waves. When thorough evaluation has failed to demonstrate another cause for the pain, and when cholangiographic and manometric criteria suggest a diagnosis of biliary dyskinesia, medical treatment with nitrites or anticholinergics to attempt pharmacologic relaxation of the sphincter has been proposed. Endoscopic biliary sphincterotomy (EBS) or surgical sphincteroplasty may be indicated in patients who fail to respond to a 2- to 3-month trial of medical therapy, especially if basal sphincter of Oddi pressures are elevated. EBS has become a well-established procedure for removing bile duct stones and for other biliary and pancreatic problems. Approximately 150,000 such procedures are performed annually in the United States. Key findings in a recent study of EBS include: (1) Dysfunction of the sphincter of Oddi was the most frequent patient-related risk factor for complications; (2) pancreatitis was more frequent in young patients; (3) difficulty in cannulating the bile duct and the use of "precut" sphincterotomy were important technique-related risk factors for complications; and (4) experience in the volume of procedures proved to be important; endoscopists who perform more than one EBS per week had lower complication rates than endoscopists who performed a smaller number of procedures.

Bile Salt-Induced Diarrhea and Gastritis Postcholecystectomy patients may develop symptoms and signs of gastritis, which has been attributed to duodenogastric reflux of bile. However, firm data linking an increased incidence of bile gastritis with surgical removal of the gallbladder are lacking. Cholecystectomy induces persistent changes in gut transit, and these changes effect a noticeable modification of bowel habits. Cholecystectomy shortens gut transit time by accelerating passage of the fecal bolus through the colon with marked acceleration in the right colon, thus causing an increase in colonic bile acid output and a shift in bile acid composition toward the more diarrheagenic secondary bile acids. Diarrhea that is severe enough, i.e., three or more watery movements per day, can be classified as postcholecystectomy diarrhea, and this occurs in 8 to 12% of patients undergoing elective cholecystectomy. Treatment with a bile acid sequestering agent, such as cholestyramine, is often effective in ameliorating troublesome diarrhea.

THE HYPERPLASTIC CHOLECYSTOSES

The term *hyperplastic cholecystoses* is used to denote a group of disorders of the gallbladder characterized by excessive proliferation of normal tissue components.

Adenomyomatosis is characterized by a benign proliferation of gallbladder surface epithelium with glandlike formations, extramural sinuses, transverse strictures, and/or fundal nodule ("adenoma" or "adenomyoma") formation. Outpouchings of mucosa termed *Rokitansky-Aschoff sinuses* may be seen on oral cholecystography in conjunction with hyperconcentration of contrast medium. Characteristic dimpled filling defects also may be seen.

Cholesterolosis is characterized by abnormal deposition of lipid, especially cholesterol esters, in the lamina propria of the gallbladder wall. In its diffuse form ("strawberry gallbladder"), the gallbladder mucosa is brick red and speckled with bright yellow flecks of lipid. The localized form shows solitary or multiple "cholesterol polyps" studding the gallbladder wall. Cholesterol stones of the gallbladder are found in nearly half the cases. Cholecystectomy is indicated in both adenomyomatosis and cholesterolosis when symptomatic or when cholelithiasis is present.

DISEASES OF THE BILE DUCTS

CONGENITAL ANOMALIES

Biliary Atresia and Hypoplasia Atretic and hypoplastic lesions of the extrahepatic and major intrahepatic bile ducts are the most common biliary anomalies of clinical relevance encountered in infancy. The clinical picture is one of severe obstructive jaundice during the first month of life, with pale stools. The diagnosis is confirmed by surgical exploration with operative cholangiography. Approximately 10% of cases of biliary atresia are treatable with roux-en-Y choledochojejunostomy, with the Kasai procedure (hepatic portoenterostomy) being attempted in the remainder in an effort to restore some bile flow. Most patients, even those having successful biliary-enteric anastomoses, eventually develop chronic cholangitis, extensive hepatic fibrosis, and portal hypertension.

Choledochal Cysts Cystic dilatation may involve the free portion of theCBD, i.e., choledochal cyst, or may present as diverticulum formation in the intraduodenal segment. In the latter situation, chronic reflux of pancreatic juice into the biliary tree can produce inflammation and stenosis of the extrahepatic bile ducts leading to cholangitis or biliary obstruction. Because the process may be gradual, approximately 50% of patients present with onset of symptoms after age 10. The diagnosis may be made by ultrasound, abdominalCT, or cholangiography. Only one-third of patients show the classic triad of abdominal pain, jaundice, and an abdominal mass. Ultrasonographic detection of a cyst separate from the gallbladder should suggest the diagnosis of choledochal cyst, which can be confirmed by demonstrating the entrance of extrahepatic bile ducts into the cyst. Surgical treatment involves excision of the "cyst" and biliary-enteric anastomosis. Patients with choledochal cysts are at increased risk for the subsequent development of cholangiocarcinoma.

Congenital Biliary Ectasia Cystic dilatation of the intrahepatic bile ducts may involve either the major intrahepatic radicles (Caroli's disease), the inter- and intralobular ducts (congenital hepatic fibrosis), or both. In Caroli's disease, clinical manifestations include recurrent cholangitis, abscess formation in and around the affected ducts, and, sometimes, gallstone formation within portions of ectatic intrahepatic biliary radicles. TheCT scan and cholangiographic patterns are usually diagnostic, and treatment with ongoing antibiotic therapy is usually undertaken in an effort to limit the frequency and severity of recurrent bouts of cholangitis. Progression to secondary biliary cirrhosis with portal hypertension, amyloidosis, extrahepatic biliary obstruction, cholangiocarcinoma, or recurrent episodes of sepsis with hepatic abscess formation is common.

CHOLEDOCHOLITHIASIS

Pathophysiology and Clinical Manifestations Passage of gallstones into theCBDoccurs in approximately 10 to 15% of patients with cholelithiasis. The incidence of common duct stones increases with increasing age of the patient, so that up to 25% of elderly patients may have calculi in the common duct at the time of cholecystectomy. Undetected duct stones are left behind in approximately 1 to 5% of cholecystectomy patients. The overwhelming majority of bile duct stones are cholesterol or mixed stones formed in the gallbladder, which then migrate into the extrahepatic biliary tree through the cystic duct. Primary calculi arising de novo in the ducts are usually pigment stones developing in patients with (1) chronic hemolytic diseases; (2) hepatobiliary parasitism or chronic, recurrent cholangitis; (3) congenital anomalies of the bile ducts (especially Caroli's disease); or (4) dilated, sclerosed, or strictured ducts. Common duct stones may remain asymptomatic for years, may pass spontaneously into the duodenum, or (most often) may present with biliary colic or a complication.

Complications

Cholangitis Cholangitis may be acute or chronic, and symptoms result from inflammation, which usually requires at least partial obstruction to the flow of bile. Bacteria are present on bile culture in approximately 75% of patients with acute cholangitis early in the symptomatic course. The characteristic presentation of acute cholangitis involves biliary colic, jaundice, and spiking fevers with chills (Charcot's triad). Blood cultures are frequently positive, and leukocytosis is typical. *Nonsuppurative* acute cholangitis is most common and may respond relatively rapidly to supportive measures and to treatment with antibiotics. In *suppurative* acute cholangitis, however, the presence of pus under pressure in a completely obstructed ductal system leads to symptoms of severe toxicity -- mental confusion, bacteremia, and septic shock. Response to antibiotics alone in this setting is relatively poor, multiple hepatic abscesses are often present, and the mortality rate approaches 100% unless prompt endoscopic or surgical relief of the obstruction and drainage of infected bile are carried out. Endoscopic management of bacterial cholangitis is as effective as surgical intervention.ERCPwith endoscopic sphincterotomy is safe and the preferred initial procedure for both establishing a definitive diagnosis and providing effective therapy.

Obstructive Jaundice Gradual obstruction of theCBD over a period of weeks or months usually leads to initial manifestations of jaundice or pruritus without associated symptoms of biliary colic or cholangitis. Painless jaundice may occur in patients with choledocholithiasis, but this manifestation is much more characteristic of biliary obstruction secondary to malignancy of the head of the pancreas, bile ducts, or ampulla of Vater.

In patients whose obstruction is secondary to choledocholithiasis, associated chronic calculous cholecystitis is very common, and the gallbladder in this setting may be relatively indistensible. The absence of a palpable gallbladder in most patients with biliary obstruction from duct stones is the basis for *Courvoisier's law*, i.e., that the presence of a palpably enlarged gallbladder suggests that the biliary obstruction is secondary to an underlying malignancy rather than to calculous disease. Biliary obstruction causes progressive dilatation of the intrahepatic bile ducts as intrabiliary pressures rise. Hepatic bile flow is suppressed, and regurgitation of conjugated bilirubin into the bloodstream leads to jaundice accompanied by dark urine (bilirubinuria) and light-colored (acholic) stools.

CBDstones should be suspected in any patient with cholecystitis whose serum bilirubin level exceeds 85.5 umol/L (5 mg/dL). The maximum bilirubin level is seldom over 256.5 umol/L (15.0 mg/dL) in patients with choledocholithiasis unless concomitant hepatic disease or another factor leading to marked hyperbilirubinemia exists. Serum bilirubin levels of 342.0 umol/L (20 mg/dL) or more should suggest the possibility of neoplastic obstruction. The serum alkaline phosphatase level is almost always elevated in biliary obstruction. A rise in alkaline phosphatase often precedes clinical jaundice and may be the only abnormality in routine liver function tests. There may be a two- to tenfold elevation of serum aminotransferases, especially in association with acute obstruction. Following relief of the obstructing process, serum aminotransferase elevations usually return rapidly to normal, while the serum bilirubin level may take 1 to 2 weeks to return to normal. The alkaline phosphatase level usually falls slowly, lagging behind the decrease in serum bilirubin.

Pancreatitis The most common associated entity discovered in patients with nonalcoholic acute pancreatitis is biliary tract disease. Biochemical evidence of pancreatic inflammation complicates acute cholecystitis in 15% of cases and choledocholithiasis in over 30%, and the common factor appears to be the passage of gallstones through the common duct. Coexisting pancreatitis should be suspected in patients with symptoms of cholecystitis who develop (1) back pain or pain to the left of the abdominal midline, (2) prolonged vomiting with paralytic ileus, or (3) a pleural effusion, especially on the left side. Surgical treatment of gallstone disease is usually associated with resolution of the pancreatitis.

Secondary Biliary Cirrhosis Secondary biliary cirrhosis may complicate prolonged or intermittent duct obstruction with or without recurrent cholangitis. Although this complication may be seen in patients with choledocholithiasis, it is more common in cases of prolonged obstruction from stricture or neoplasm. Once established, secondary biliary cirrhosis may be progressive even after correction of the obstructing process, and increasingly severe hepatic cirrhosis may lead to portal hypertension or to hepatic failure and death. Prolonged biliary obstruction may also be associated with clinically relevant deficiencies of the fat-soluble vitamins A, D, and K.

Diagnosis and Treatment The diagnosis of choledocholithiasis is usually made by cholangiography (Table 302-3), either preoperatively byERCP or intraoperatively at the time of cholecystectomy. As many as 15% of patients undergoing cholecystectomy will prove to haveCBDstones. With the advent of laparoscopic cholecystectomy, the management of CBD stones in the presence of gallstones is gradually being clarified. Preoperative ERCP with endoscopic papillotomy and stone extraction is the preferred approach. It not only provides stone clearance but also defines the anatomy of the biliary tree in relationship to the cystic duct. ERCP is indicated in gallstone patients who have any of the following risk factors: (1) a history of jaundice or pancreatitis, (2) abnormal tests of liver function, and (3) ultrasonographic evidence of a dilated CBD or stones in the duct. Alternatively, if intraoperative cholangiography reveals retained stones, postoperative ERCP can be carried out. The need for preoperative ERCP is expected to decrease further as laparoscopic techniques improve.

The widespread use of laparoscopic cholecystectomy andERCP has decreased the incidence of complicated biliary tract disease and the need for choledocholithotomy and T-tube drainage of the bile ducts.EBSfollowed by spontaneous passage or stone extraction is the treatment of choice in the management of patients with common duct stones, especially in elderly or poor-risk patients.

TRAUMA, STRICTURES, AND HEMOBILIA

Benign strictures of the extrahepatic bile ducts result from surgical trauma in approximately 95% of cases and occur in about 1 in 500 cholecystectomies. Strictures may present with bile leak or abscess formation in the immediate postoperative period or with biliary obstruction or cholangitis as long as 2 years or more following the inciting trauma. The diagnosis is established by percutaneous or endoscopic cholangiography. Endoscopic brushing of biliary strictures is an effective way to establish the nature of the lesion and is more accurate than bile cytology alone. When positive exfoliative cytology is obtained, the diagnosis of a neoplastic stricture is established. This procedure is especially important in patients with primary sclerosing cholangitis who are predisposed to the development of cholangiocarcinomas. Successful operative correction by a skillful surgeon with duct-to-bowel anastomosis is usually possible, although mortality rates from surgical complications, recurrent cholangitis, or secondary biliary cirrhosis are high.

Hemobilia may follow traumatic or operative injury to the liver or bile ducts, intraductal rupture of a hepatic abscess or aneurysm of the hepatic artery, biliary or hepatic tumor hemorrhage, or mechanical complications of choledocholithiasis or hepatobiliary parasitism. Diagnostic procedures such as liver biopsy, percutaneous transhepatic cholangiography (PTHC), and transhepatic biliary drainage catheter placement may also be complicated by hemobilia. Patients often present with a classic triad of biliary colic, obstructive jaundice, and melena or occult blood in the stools. The diagnosis is sometimes made by cholangiographic evidence of blood clot in the biliary tree, but selective angiographic verification may be required. Although minor episodes of hemobilia may resolve without operative intervention, surgical ligation of the bleeding vessel is frequently required.

EXTRINSIC COMPRESSION OF THE BILE DUCTS

Partial or complete biliary obstruction may sometimes be produced by extrinsic compression of the ducts. The most common cause of this form of obstructive jaundice is carcinoma of the head of the pancreas. Biliary obstruction may also occur as a complication of either acute or chronic pancreatitis or involvement of lymph nodes in the porta hepatis by lymphoma or metastatic carcinoma. The latter should be distinguished from cholestasis resulting from massive replacement of the liver by tumor.

HEPATOBILIARY PARASITISM

Infestation of the biliary tract by adult helminths or their ova may produce a chronic, recurrent pyogenic cholangitis with or without multiple hepatic abscesses, ductal stones, or biliary obstruction. This condition is relatively rare but does occur in inhabitants of southern China and elsewhere in Southeast Asia. The organisms most commonly involved are trematodes or flukes, including *Clonorchis sinensis*, *Opisthorchis viverrini* or *O. felineus*, and *Fasciola hepatica*. The biliary tract also may be involved by intraductal migration of adult *Ascaris lumbricoides* from the duodenum or by intrabiliary rupture of hydatid cysts of the liver produced by *Echinococcus* spp. The diagnosis is made by cholangiography and the presence of characteristic ova on stool examination. When obstruction is present, the treatment of choice is laparotomy under antibiotic coverage, with common duct exploration and a biliary drainage procedure. It should be emphasized that in the Far East, one also sees cholangiohepatitis associated with pigment lithiasis, which may, in fact, be more common than cholangitis due to parasites.

SCLEROSING CHOLANGITIS

Primary or idiopathic sclerosing cholangitis is characterized by a progressive, inflammatory, sclerosing, and obliterative process affecting the extrahepatic and/or the intrahepatic bile ducts. The disorder occurs in about 70% in association with inflammatory bowel disease, especially ulcerative colitis. It may also be associated (albeit rarely) with multifocal fibrosclerosis syndromes such as retroperitoneal, mediastinal, and/or periureteral fibrosis; Riedel's struma; or pseudotumor of the orbit. In patients with AIDS, cholangiopancreatography may demonstrate a broad range of biliary tract changes as well as pancreatic duct obstruction and occasionally pancreatitis (Chap. 309). Further, biliary tract lesions in AIDS include infection and cholangiopancreatographic changes similar to primary sclerosing cholangitis. Changes noted include: (1) diffuse involvement of intrahepatic bile ducts alone, (2) involvement of both intra- and extrahepatic bile ducts, (3) ampullary stenosis, (4) stricture of the intrapancreatic portion of the common bile duct, and (5) pancreatic duct involvement. Associated infectious organisms include *Cryptosporidium*, *Mycobacterium avium-intracellulare*, cytomegalovirus, *Microsporidia*, and *Isospora*. In addition, acalculous cholecystitis occurs in up to 10% of patients.ERCPsphincterotomy, while not without risk, provides significant pain reduction in patients with AIDS-associated papillary stenosis. Secondary sclerosing cholangitis may occur as a long-term complication of choledocholithiasis, cholangiocarcinoma, operative or traumatic biliary injury, or contiguous inflammatory processes.

Patients with primary sclerosing cholangitis often present with signs and symptoms of chronic or intermittent biliary obstruction: jaundice, pruritus,RUQabdominal pain, or acute cholangitis. Late in the course, complete biliary obstruction, secondary biliary cirrhosis, hepatic failure, or portal hypertension with bleeding varices may occur. The diagnosis is usually established by finding multifocal, diffusely distributed strictures with intervening segments of normal or dilated ducts, producing a beaded appearance on cholangiography (Fig. 302-2*D*). The cholangiographic technique of choice in suspected cases isERCP, since intrahepatic ductal involvement may makePTHCdifficult. When a diagnosis of sclerosing cholangitis has been established, a search for associated diseases, especially for chronic inflammatory bowel disease, should be carried out.

A recent study describes the natural history and outcome for 305 patients of Swedish descent with primary sclerosing cholangitis; 134 (44%) of the patients were asymptomatic at the time of diagnosis and, not surprisingly, had a significantly higher survival rate with a median follow-up time of 63 months. The independent predictors of a bad prognosis were age, serum bilirubin concentration, and liver histologic changes. Cholangiocarcinoma was found in 24 patients (8%). Inflammatory bowel disease was closely associated with primary sclerosing cholangitis and had a prevalence of 81% in this study population.

TREATMENT

Therapy with cholestyramine may help control symptoms of pruritus, and antibiotics are useful when cholangitis complicates the clinical picture. Vitamin D and calcium supplementation may help prevent the loss of bone mass frequently seen in patients with chronic cholestasis. Glucocorticoids, methotrexate, and cyclosporine have not been shown to be efficacious.UDCAimproves serum liver tests, but an effect on survival has not been documented. In cases where complete or high-grade biliary obstruction (dominant strictures) has occurred, balloon dilatation, stenting, or (rarely) surgical intervention may be appropriate. Efforts at biliary-enteric anastomosis or stent placement may, however, be complicated by recurrent cholangitis and further progression of the stenosing process. The role of colectomy in patients with sclerosing cholangitis complicating chronic ulcerative colitis is uncertain. The prognosis is unfavorable, with a median survival of 9 to 12 years following the diagnosis, regardless of therapy. Four variables (age, serum bilirubin level, histologic stage, and splenomegaly) predict survival in patients with primary sclerosing cholangitis and serve as the basis for a risk score. Primary sclerosing cholangitis is one of the most common indications for liver transplantation.

In two large studies involving 627 and 3147 patients, the prevalence of gallbladder polyps was 6.7 and 6.9%, respectively, with a marked male predominance. Few significant changes occurred over a 5-year period in asymptomatic patients in whom gallbladder polyps<10 mm in diameter were found. If polyps>10 mm are present and show rapid growth, cholecystectomy should be considered.

ACKNOWLEDGEMENT

This chapter was written by Dr. Norton J. Greenberger and Dr. Kurt J. Isselbacher in the previous edition.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 3 -DISORDERS OF THE PANCREAS

303. APPROACH TO THE PATIENT WITH PANCREATIC DISEASE - *Phillip P. Toskes*, *Norton J. Greenberger*

GENERAL CONSIDERATIONS

Inflammatory disease of the pancreas may be acute or chronic. Although good data exist concerning the frequency of acute pancreatitis (about 5000 new cases per year in the United States, with a mortality rate of about 10%), the number of patients who suffer with recurrent acute pancreatitis or chronic pancreatitis is largely undefined. Only one prospective study on the incidence of chronic pancreatitis is available; it showed an incidence of 8.2 new cases per 100,000 per year and a prevalence of 26.4 cases per 100,000. These numbers probably underestimate considerably the true incidence and prevalence, because non-alcohol-induced pancreatitis was largely ignored. At autopsy, the prevalence of chronic pancreatitis ranges from 0.04 to 5%. The relative inaccessibility of the pancreas to direct examination and the nonspecificity of the abdominal pain associated with pancreatitis make the diagnosis of pancreatitis difficult and usually dependent on elevation of blood amylase levels. Many patients with chronic pancreatitis do not have elevated blood amylase levels. Some patients with chronic pancreatitis develop signs and symptoms of pancreatic exocrine insufficiency, and thus objective evidence for pancreatic disease can be demonstrated. However, there is a very large reservoir of pancreatic exocrine function. More than 90% of the pancreas must be damaged before maldigestion of fat and protein is manifested. Even the secretin stimulation test, which is the most sensitive method of assessing pancreatic exocrine function, is probably abnormal only when more than 60% of exocrine function has been lost. Noninvasive, indirect tests of pancreatic exocrine function (bentiromide, serum trypsinogen) are much more likely to give abnormal results in patients with obvious pancreatic disease, i.e., pancreatic calcification, steatorrhea, or diabetes mellitus, than in patients with occult disease. Thus, the number of patients who have subclinical exocrine dysfunction (less than 90% loss of function) is unknown.

The clinical manifestations of acute and chronic pancreatitis and pancreatic insufficiency are protean. Thus, patients may present with hypertriglyceridemia, vitamin B_{12}malabsorption, hypercalcemia, hypocalcemia, hyperglycemia, ascites, pleural effusions, and chronic abdominal pain with normal blood amylase levels. Indeed, if the clinician considers pancreatitis as a possible diagnosis only when presented with a patient having classic symptoms (i.e., severe, constant epigastric pain that radiates through to the back, along with an elevated blood amylase level), only a minority of patients with pancreatitis will be diagnosed correctly.

As emphasized inChap. 304, the etiologies as well as the clinical manifestations of pancreatitis are quite varied. Although it is well appreciated that pancreatitis is frequently secondary to alcohol abuse and biliary tract disease, it can also be caused by drugs, trauma, and viral infections and is associated with metabolic and connective tissue disorders. In approximately 30% of patients with acute pancreatitis and 25 to 40% of patients with chronic pancreatitis, the etiology is obscure.

TESTS USEFUL IN THE DIAGNOSIS OF PANCREATIC DISEASE

Several tests have proved of value in the evaluation of pancreatic exocrine function. Examples of specific tests and their usefulness in the diagnosis of acute and chronic pancreatitis are summarized inTable 303-1. At most institutions, pancreatic function tests are performed if the diagnosis of pancreatic disease remains a possibility after noninvasive tests [ultrasound, computed tomography (CT)] and invasive tests [endoscopic retrograde cholangiopancreatography (ERCP)] have given normal or inconclusive results. In this regard, tests employing *direct* stimulation of the pancreas are the most sensitive.

PANCREATIC ENZYMES IN BODY FLUIDS

The serum amylase level is widely used as a screening test for acute pancreatitis in the patient with acute abdominal pain or back pain. A value greater than 65 U/L should raise the question of acute pancreatitis. Levels greater than 130 U/L make the diagnosis more likely, and values greater than three times normal virtually clinch the diagnosis if gut perforation or infarction is excluded. In acute pancreatitis, the serum amylase is usually elevated within 24 h of onset and remains so for 1 to 3 days. Levels return to normal within 3 to 5 days unless there is extensive pancreatic necrosis, incomplete ductal obstruction, or pseudocyst formation. Approximately 85% of patients with acute pancreatitis have an elevated serum amylase level. This index may be normal, however, if (1) there is a delay (of 2 to 5 days) before blood samples are obtained, (2) the underlying disorder is chronic pancreatitis rather than acute pancreatitis, or (3) hypertriglyceridemia is present. Patients with hypertriglyceridemia and proven pancreatitis have been found to have spuriously low levels of amylase and lipase activity.

The serum amylase is often elevated in other conditions (Table 303-2), in part because the enzyme is found in many organs in addition to the pancreas (salivary glands, liver, small intestine, kidney, fallopian tube) and can be produced by various tumors (carcinomas of the lung, esophagus, breast, and ovary). Isoenzymes of amylase fall into two general categories: those arising from the pancreas (P isoamylases) and those arising from nonpancreatic sources (S isoamylases). The measurement of serum isoamylases is of clinical importance. In normal serum, about 35 to 45% of the amylase is of pancreatic origin. For example, in patients with acute pancreatitis, the total serum amylase level returns to normal more rapidly than the level of pancreatic isoamylase. Thus, in patients seen after the first day, the pancreatic isoamylase level is a more sensitive indicator of pancreatitis than the total serum amylase level. In the past, elevations in serum amylase seen in certain conditions, such as the postoperative state, acute alcohol intoxication, and diabetic ketoacidosis, were assumed to indicate acute pancreatitis. However, the elevation of serum amylase in such conditions has been shown to be due to an elevation of the S isoamylase. Simple tests to distinguish pancreatic from nonpancreatic amylase are no longer readily available, and such tests are often not reliable when the total amylase is minimally to moderately elevated. An assay of serum trypsinogen (performed by several commercial laboratories) is quite helpful in this regard. Since this enzyme is secreted specifically by the pancreas, a normal serum trypsinogen level in a patient with minimal elevation of serum amylase essentially rules out acute pancreatitis. Urinary amylase measurements, including the amylase/creatinine clearance ratio, are no more sensitive or specific than blood amylase levels.

Elevation of ascitic fluid amylase occurs in acute pancreatitis as well as in (1) pancreatogenous ascites due to disruption of the main pancreatic duct or a leaking pseudocyst and (2) other abdominal disorders that simulate pancreatitis (e.g., intestinal obstruction, intestinal infarction, and perforated peptic ulcer). Elevation of pleural fluid amylase occurs in acute pancreatitis, chronic pancreatitis, carcinoma of the lung, and esophageal perforation.

Lipase may now be the single best enzyme to measure for the diagnosis of acute pancreatitis. Improvements in substrates and technology offer clinicians improved options, especially when a turbidimetric assay is used. The newer lipase assays have colipase as a cofactor and are fully automated.

An assay for trypsinogen (or for trypsin-like immunoreactivity) has a theoretical advantage over amylase and lipase determinations in that the pancreas is the only organ that contains this enzyme. The test appears to be useful in the diagnosis of both acute and chronic pancreatitis. Sensitivity and specificity are comparable to those of amylase and lipase determinations. Since trypsinogen is also excreted by the kidney, elevated serum values are found in renal failure, as is the case with serum amylase and lipase levels. *No single blood test is reliable for the diagnosis of acute pancreatitis in patients with renal failure.* Determining whether a patient with renal failure and abdominal pain has pancreatitis remains a difficult clinical problem. A recent study found that serum amylase levels were elevated in patients with renal dysfunction only when creatinine clearance was less than 50 mL/min. In such patients, the serum amylase level was invariably less than 500 IU/L in the absence of objective evidence of acute pancreatitis. In that study, serum lipase and trypsin levels paralleled serum amylase values.

A recent study evaluated the sensitivity and specificity of five assays used to diagnose acute pancreatitis: two for amylase, one for lipase, one for trypsin-like immunoreactivity (TLI), and one for pancreatic isoamylase. The data obtained (1) show that, if the best cutoff level is used, all these assays have similar specificities and (2) suggest that total serum amylase is as good an indicator of acute pancreatitis as any of the alternatives. However, inherent in many such studies is the problem that the recognition and diagnosis of acute pancreatitis hinge on the finding of an elevated serum amylase level. The question arises as to whether any diagnostic test result can be proved superior to the total serum amylase level if hyperamylasemia is required for the diagnosis. In other studies, when "objective" confirmation of the clinical diagnosis of pancreatitis was required (ultrasonography,CT, laparotomy), the sensitivity of the serum amylase has been found to be as low as 68%. With these limitations in mind, the recommended screening tests for acute pancreatitis are *total serum amylase* and *serum lipase activities*. Serum amylase values greater than three times normal are highly specific.

STUDIES PERTAINING TO PANCREATIC STRUCTURE

Radiologic Tests Plain films of the abdomen provide useful information in 30 to 50% of patients with acute pancreatitis. The most frequent abnormalities include (1) a localized ileus, usually involving the jejunum ("sentinel loop"); (2) a generalized ileus with air-fluid levels; (3) the "colon cutoff sign," which results from isolated distention of the transverse colon; (4) duodenal distention with air-fluid levels; and (5) a mass, which is frequently a pseudocyst. In chronic pancreatitis, an important radiographic finding is pancreatic calcification, which characteristically is localized adjacent to and superimposed on the second lumbar vertebra (see Fig. 304-3*A*).

Upper gastrointestinal x-rays may reveal displacement of the stomach by the retroperitoneal mass (see Fig. 304-2*A*) or widening and effacement of the duodenal C loop, which also suggest the presence of a pancreatic mass, which could be inflammatory, cystic, or neoplastic. However, the use of x-ray films has been largely superseded by ultrasound.

Ultrasonography can provide important information in patients with acute pancreatitis, chronic pancreatitis, pancreatic calcification, pseudocyst, and pancreatic carcinoma. Echographic appearances can indicate the presence of edema, inflammation, and calcification (not obvious on plain films of the abdomen), as well as pseudocysts, mass lesions, and gallstones (see Figs. 304-2*B* and 304-3*B*). In acute pancreatitis, the pancreas is characteristically enlarged. In pancreatic pseudocyst, the usual appearance is that of an echo-free, smooth, round fluid collection. Pancreatic carcinoma distorts the usual landmarks, and mass lesions greater than 3.0 cm are usually detected as localized, echo-free solid lesions. Ultrasound is often the initial investigation for most patients with suspected pancreatic disease. However, obesity, excess small- and large-bowel gas, and recently performed barium contrast examinations can interfere with ultrasound studies.

CTis the best imaging study for initial evaluation of a suspected chronic pancreatic disorder and for the complications of acute and chronic pancreatitis. It is especially useful in the detection of pancreatic tumors, fluid-containing lesions such as pseudocysts and abscesses, and calcium deposits (seeFigs. 304-3*C* and 304-4*A*). Most lesions are characterized by (1) enlargement of the pancreatic outline, (2) distortion of the pancreatic contour, and/or (3) a fluid filling that has a different attenuation coefficient than normal pancreas. However, it is occasionally difficult to distinguish between inflammatory and neoplastic lesions. Oral water-soluble contrast agents may be used to opacify the stomach and duodenum during CT scans; this strategy permits more precise delineation of various organs as well as mass lesions. Dynamic CT (using rapid intravenous administration of contrast) is useful in estimating the degree of pancreatic necrosis and in predicting morbidity and mortality. Spiral (helical) CT provides clear images much more rapidly and essentially negates artifact caused by patient movement (see Fig. 304-2*D*).

Endoscopic ultrasonography (*EUS*) produces high-resolution images of the pancreatic parenchyma and pancreatic duct with a transducer fixed to an endoscope that can be directed onto the surface of the pancreas through the stomach or duodenum. Although criteria for abnormalities on EUS in severe pancreatic disease have been developed, the true sensitivity and specificity of this procedure has yet to be determined. In particular, it is not clear whether EUS can detect early pancreatic disease before abnormalities appear on more conventional radiograph tests such as ultrasonography orCT. The exact role of EUS versusERCP and CT has yet to be defined.

Magnetic resonance cholangiopancreatography (*MRCP*) is now being used to view both the bile duct and the pancreatic duct. Nonbreath-hold and 3D turbo spin-echo techniques are being utilized to produce superb MRCP images. The main pancreatic duct and common bile duct can be seen well, but there is still a question as to whether changes can be detected consistently in the secondary ducts. MRCP may be particularly useful to evaluate the pancreatic duct in high-risk patients such as the elderly because this is a noninvasive procedure.

BothEUS andMRCP may replaceERCP in some patients. As these techniques become more refined, they may well be the diagnostic tests of choice to evaluate the pancreatic duct. ERCP is still needed to perform therapy of bile duct and pancreatic duct lesions.

Selective catheterization of the celiac and superior mesenteric arteries combined with superselective catheterization of others arteries, such as the hepatic, splenic, and gastroduodenal arteries permits visualization of the pancreas and detection of pancreatic neoplasms and pseudocysts. Pancreatic neoplasms can be identified by the sheathing of blood vessels by a mass lesion (seeFig. 304-1*D*). Hormone-producing pancreatic tumors are especially likely to exhibit increased vascularity and tumor staining. Angiographic abnormalities are noted in many patients with pancreatic carcinoma but are uncommon in patients without pancreatic disease. Angiography complements ultrasonography andERCP in the study of patients with a suspected pancreatic lesion and may be carried out if ERCP is either unsuccessful or nondiagnostic.

ERCPmay provide useful information on the status of the pancreatic ductal system and thus aid in the differential diagnosis of pancreatic disease (seeFigs. 304-1*C*, 304-3*D*, and 304-4*B*). Pancreatic carcinoma is characterized by stenosis or obstruction of either the pancreatic duct or the common bile duct; both ductal systems are often abnormal. In chronic pancreatitis, ERCP abnormalities include (1) luminal narrowing, (2) irregularities in the ductal system with stenosis, dilation, sacculation, and ectasia, and (3) blockage of the pancreatic duct by calcium deposits. The presence of ductal stenosis and irregularity can make it difficult to distinguish chronic pancreatitis from carcinoma. It is important to be aware that ERCP changes interpreted as indicating chronic pancreatitis actually may be due to the effects of aging on the pancreatic duct or to the fact that the procedure was performed within several weeks of an attack of acute pancreatitis. Although aging may cause impressive ductal alterations, it does not affect the results of pancreatic function tests (i.e., the secretin test). Elevated serum and/or urine amylase levels after ERCP have been reported in 25 to 75% of patients, but clinical pancreatitis is uncommon. In a series of 300 patients, pancreatitis occurred in only 5 patients after ERCP. If no lesion is found in the biliary and/or pancreatic ducts in a patient with repeated attacks of acute pancreatitis, manometric studies of the sphincter of Oddi may be indicated. Such studies, however, do increase the risk of post-ERCP/manometry acute pancreatitis. Such pancreatitis appears to be more common in patients with a nondilated pancreatic duct.

Pancreatic Biopsy with Radiologic Guidance Percutaneous aspiration biopsy of a pancreatic mass often distinguishes a pancreatic inflammatory mass from a pancreatic neoplasm.

TESTS OF EXOCRINE PANCREATIC FUNCTION

Pancreatic function tests (Table 303-1) can be divided into the following:

1. *Direct stimulation of the pancreas* by intravenous infusion of secretin or secretin plus cholecystokinin (CCK) followed by collection and measurement of duodenal contents

2. *Indirect stimulation of the pancreas* using nutrients or amino acids, fatty acids, and synthetic peptides followed by assays of proteolytic, lipolytic, and amylolytic enzymes

3. Study of *intraluminal digestion products*, such as undigested meat fibers, stool fat, and fecal nitrogen

4. *Measurement of fecal pancreatic enzymes* such as elastase

The secretin test, used to detect diffuse pancreatic disease, is based on the physiologic principle that the pancreatic secretory response is directly related to the functional mass of pancreatic tissue. In the standard assay, secretin is given intravenously in a dose of 1 clinical unit (CU) per kilogram, as either a bolus or a continuous infusion. The results will vary with the secretin preparation used, the dose, the mode of administration, and the completeness with which the duodenal contents are collected. Normal values for the standard secretin test are (1) volume output >2.0 mL/kg per hour, (2) bicarbonate (HCO_{3^-}) concentration >80 meql/L, and (3) HCO_{3^-}output >10 meq/L in 1 h. The most reproducible measurement, giving the highest level of discrimination between normal subjects and patients with chronic pancreatitis, appears to be the maximal bicarbonate concentration.

The *combined secretin-CCKtest* permits measurement of pancreatic amylase, lipase, trypsin, and chymotrypsin. Although there is overlap in the distributions of enzyme output in normal subjects and patients with pancreatitis in response to this test, markedly low enzyme outputs suggest advanced damage and destruction of acinar cells. With frank exocrine pancreatic insufficiency, there is usually an overall reduction in both HCO_{3^-}concentration and output of several enzymes. However, with lesser degrees of pancreatic damage there may be a dissociation between HCO_{3^-}concentration and enzyme output. There also may be a dissociation between the results of the secretin test and those of tests of absorptive function. For example, patients with chronic pancreatitis often have abnormally low outputs of HCO_{3^-}after secretin but have normal fecal fat excretion. Thus the secretin test measures the secretory capacity of ductular epithelium, while fecal fat excretion indirectly reflects intraluminal lipolytic activity. Steatorrhea does not occur until intraluminal levels of lipase are markedly reduced, underscoring the fact that only small amounts of enzymes are necessary for intraluminal digestive activities. An abnormal secretin test result suggests only that chronic pancreatic damage is present; it will not consistently distinguish between chronic pancreatitis and pancreatic carcinoma.

Another test of exocrine pancreatic function is the *bentiromide test*. This test is an indirect measure of pancreatic function and reflects intraluminal chymotrypsin activity. The test has excellent specificity but is not very sensitive. It no longer is available for clinical use in the United States.

The *serum trypsinogen level*, which is determined by radioimmunoassay, also has excellent specificity but is not very sensitive. It is a simple blood test that can detect severe damage to the exocrine pancreas. The normal values are 28 to 58 ng/mL, and any value below 20 ng/mL reflects pancreatic steatorrhea.

Measurement of *intraluminal digestion products*, i.e., undigested muscle fibers, stool fat, and fecal nitrogen, is discussed in Chap. 286. The amount of elastase in stool reflects the pancreatic output of this proteolytic enzyme. Decreased elastase activity in stool has been reported in patients with chronic pancreatitis and cystic fibrosis.**Tests useful in the diagnosis of exocrine pancreatic insufficiency and the differential diagnosis of malabsorption are also discussed in Chaps. 286 and 304.*

Back to Table of Contents

304. ACUTE AND CHRONIC PANCREATITIS - *Norton J. Greenberger*, *Phillip P. Toskes*

BIOCHEMISTRY AND PHYSIOLOGY OF PANCREATIC EXOCRINE SECRETION

GENERAL CONSIDERATIONS

The pancreas secretes 1500 to 3000 mL of isosmotic alkaline (pH >8.0) fluid per day containing about 20 enzymes and zymogens. The pancreatic secretions provide the enzymes needed to effect the major digestive activity of the gastrointestinal tract and provide an optimal pH for the function of these enzymes.

REGULATION OF PANCREATIC SECRETION

The exocrine pancreas is influenced by intimately interacting hormonal and neural systems. *Gastric acid* is the stimulus for the release of secretin, a peptide with 27 amino acids. Sensitive radioimmunoassay studies for secretin suggest that the pH threshold for its release from the duodenum and jejunum is 4.5. Secretin stimulates the secretion of pancreatic juice rich in *water and electrolytes*. Release of cholecystokinin (CCK) from the duodenum and jejunum is largely triggered by long-chain fatty acids, certain essential amino acids (tryptophan, phenylalanine, valine, methionine), and gastric acid itself. CCK evokes an *enzyme-rich secretion from the pancreas*. Gastrin, although it has the same terminal tetrapeptide as CCK, is a weak stimulus for pancreatic enzyme output. The *parasympathetic nervous system* (via the vagus nerve) exerts significant control over pancreatic secretion. Secretion evoked by secretin and CCK depends on permissive roles of vagal afferent and efferent pathways. This is particularly true for enzyme secretion, whereas water and bicarbonate secretion is heavily dependent on the hormonal effects of secretin and CCK. Also, vagal stimulation effects the release of vasoactive intestinal peptide (VIP), a secretin agonist. Bile salts also stimulate pancreatic secretion, thereby integrating the functions of the biliary tract, pancreas, and small intestine.

Somatostatin acts on multiple sites to induce inhibition of pancreatic secretion. The appropriate roles of other peptides, such as peptide YY, pancreastatin, gastrin-releasing peptide, pituitary adenylate cyclase-activating polypeptide, calcitonin gene-related peptide, and galanin are still being defined. Nitric oxide is an important neurotransmitter in the regulation of pancreatic exocrine secretion, although its mechanism of action has not been fully elucidated.

WATER AND ELECTROLYTE SECRETION

Although sodium, potassium, chloride, calcium, zinc, phosphate, and sulfate are found in pancreatic secretions, *bicarbonate is the ion of primary physiologic importance*. In the acini and in the ducts, secretin causes the cells to add water and bicarbonate to the fluid. In the ducts, an exchange occurs between bicarbonate and chloride. There is a good correlation between the maximal bicarbonate output after stimulation with secretin and the pancreatic mass. The bicarbonate output of 120 to 300 mmol/d helps neutralize gastric acid and creates the appropriate pH for the activity of the pancreatic enzymes.

ENZYME SECRETION

The pancreas secretes amylolytic, lipolytic, and proteolytic enzymes. *Amylolytic enzymes*, such as amylase, hydrolyze starch to oligosaccharides and to the disaccharide maltose. The *lipolytic enzymes* include lipase, phospholipase A, and cholesterol esterase. Bile salts *inhibit* lipase in isolation; but colipase, another constituent of pancreatic secretion, binds to lipase and prevents this inhibition. Bile salts *activate* phospholipase A and cholesterol esterase. *Proteolytic enzymes* include *endopeptidases* (trypsin, chymotrypsin), which act on internal peptide bonds of proteins and polypeptides; *exopeptidases* (carboxypeptidases, aminopeptidases), which act on the free carboxyl- and amino-terminal ends of peptides, respectively; and elastase. The proteolytic enzymes are secreted as inactive precursors (zymogens). Ribonucleases (deoxyribonucleases, ribonuclease) are also secreted. Although pancreatic enzymes usually are secreted in parallel, nonparallel secretion can occur as a result of exocytosis from heterogeneous sources in the pancreas. *Enterokinase*, an enzyme found in the duodenal mucosa, cleaves the lysine-isoleucine bond of trypsinogen to form trypsin. Trypsin then activates the other proteolytic zymogens in a cascade phenomenon. All pancreatic enzymes have pH optima in the alkaline range.

AUTOPROTECTION OF THE PANCREAS

Autodigestion of the pancreas is prevented by the packaging of proteases in precursor form and by the synthesis of protease inhibitors. These protease inhibitors are found in the acinar cell, the pancreatic secretions, and the alpha$_1$- and alpha$_2$-globulin fractions of plasma.

EXOCRINE-ENDOCRINE RELATIONSHIPS

Insulin appears to be needed locally for secretin andCCK to promote exocrine secretion; thus, it acts in a permissive role for these two hormones.

ENTEROPANCREATIC AXIS AND FEEDBACK INHIBITION

Pancreatic enzyme secretion is controlled, at least in part, by a negative feedback mechanism induced by the presence of active serine proteases in the duodenum. To illustrate, perfusion of the duodenal lumen with phenylalanine causes a prompt increase in plasmaCCKlevels as well as increased secretion of chymotrypsin. However, simultaneous perfusion with trypsin blunts both responses. Conversely, perfusion of the duodenal lumen with protease inhibitors actually leads to enzyme hypersecretion. The available evidence supports the concept that the duodenum contains a peptide called CCK releasing factor (CCK-RF) that is involved in stimulating CCK release. Two peptides, luminal CCK-RF and diazepam-binding inhibitor, have been found that may be the CCK-RF. Serine proteases inhibit pancreatic secretion by acting on CCK-RF. It appears that serine proteases inhibit pancreatic secretion by acting on a CCK-releasing peptide in the lumen of the small intestine.

ACUTE PANCREATITIS

GENERAL CONSIDERATIONS

Pancreatic inflammatory disease may be classified as (1) acute pancreatitis and (2) chronic pancreatitis. The pathologic spectrum of acute pancreatitis varies from *edematous pancreatitis*, which is usually a mild and self-limited disorder, to *necrotizing pancreatitis*, in which the degree of pancreatic necrosis correlates with the severity of the attack and its systemic manifestations. The term *hemorrhagic pancreatitis* is less meaningful in a clinical sense because variable amounts of interstitial hemorrhage can be found in pancreatitis as well as in other disorders such as pancreatic trauma, pancreatic carcinoma, and severe congestive heart failure.

The incidence of pancreatitis varies in different countries and depends on cause, e.g., alcohol, gallstones, metabolic factors, and drugs (Table 304-1). In the United States, for example, acute pancreatitis is related to alcohol ingestion more commonly than to gallstones; in England, the opposite obtains. There are 185,000 new cases of acute pancreatitis per year in the United States.

ETIOLOGY AND PATHOGENESIS

There are many causes of acute pancreatitis (Table 304-1), but the mechanisms by which these conditions trigger pancreatic inflammation have not been identified. Alcoholic patients with pancreatitis may represent a special subset, since most alcoholics do not develop pancreatitis. The list of identifiable causes is growing, and it is likely that pancreatitis related to viral infections, drugs, and as yet undefined factors is more common than heretofore recognized.

Approximately 2 to 5% of cases of acute pancreatitis are drug-related (Table 304-1). Drugs cause pancreatitis either by a hypersensitivity reaction or by the generation of a toxic metabolite, although in some cases it is not clear which of these mechanisms is operative.

Autodigestion is one pathogenetic theory, according to which pancreatitis results when proteolytic enzymes (e.g., trypsinogen, chymotrypsinogen, proelastase, and phospholipase A) are activated in the pancreas rather than in the intestinal lumen. A number of factors (e.g., endotoxins, exotoxins, viral infections, ischemia, anoxia, and direct trauma) are believed to activate these proenzymes. Activated proteolytic enzymes, especially trypsin, not only digest pancreatic and peripancreatic tissues but also can activate other enzymes, such as elastase and phospholipase. The active enzymes then digest cellular membranes and cause proteolysis, edema, interstitial hemorrhage, vascular damage, coagulation necrosis, fat necrosis, and parenchymal cell necrosis. Cellular injury and death result in the liberation of activated enzymes. In addition, activation and release of bradykinin peptides and vasoactive substances (e.g., histamine) are believed to produce vasodilation, increased vascular permeability, and edema. Thus, a cascade of events culminates in the development of acute necrotizing pancreatitis.

The autodigestion theory has largely eclipsed two older theories. First, according to the "common channel" theory, the existence of a common anatomic channel for pancreatic secretions and bile permits reflux of bile into the pancreatic duct, which results in activation of pancreatic enzymes. (Actually, a common channel with free communication between the common bile duct and the main pancreatic duct is infrequently encountered.) The second theory is that obstruction and hypersecretion are pivotal in the development of pancreatitis. Obstruction of the main pancreatic duct, however, produces pancreatic edema but generally not pancreatitis.

A recent hypothesis to explain the intrapancreatic activation of zymogens is that they become activated by *lysosomal hydrolases* in the pancreatic acinar cell itself. In two different types of experimental pancreatitis, it has been demonstrated that digestive enzymes and lysosomal hydrolases become admixed; as a result, the latter can activate the former in the acinar cell. In vitro, lysosomal enzymes such as cathepsin B can activate trypsinogen, and trypsin can activate the other protease precursors. It is still not clear, however, whether the human acinar cell can provide the pH (about 3.0) necessary for activation of trypsinogen by lysosomal hydrolases. It is now believed that ischemia/hypoperfusion can alone result in activation of trypsinogen and pancreatic injury.

CLINICAL FEATURES

Abdominal pain is the major symptom of acute pancreatitis. Pain may vary from a mild and tolerable discomfort to severe, constant, and incapacitating distress. Characteristically, the pain, which is steady and boring in character, is located in the epigastrium and periumbilical region and often radiates to the back as well as to the chest, flanks, and lower abdomen. The pain is frequently more intense when the patient is supine, and patients often obtain relief by sitting with the trunk flexed and knees drawn up. Nausea, vomiting, and abdominal distention due to gastric and intestinal hypomotility and chemical peritonitis are also frequent complaints.

Physical examination frequently reveals a distressed and anxious patient. Low-grade fever, tachycardia, and hypotension are fairly common. Shock is not unusual and may result from (1) hypovolemia secondary to exudation of blood and plasma proteins into the retroperitoneal space (a "retroperitoneal burn"); (2) increased formation and release of kinin peptides, which cause vasodilation and increased vascular permeability; and (3) systemic effects of proteolytic and lipolytic enzymes released into the circulation. Jaundice occurs infrequently; when present, it usually is due to edema of the head of the pancreas with compression of the intrapancreatic portion of the common bile duct. Erythematous skin nodules due to subcutaneous fat necrosis may occur. In 10 to 20% of patients, there are pulmonary findings, including basilar rales, atelectasis, and pleural effusion, the latter most frequently left-sided. Abdominal tenderness and muscle rigidity are present to a variable degree, but, compared with the intense pain, these signs may be unimpressive. Bowel sounds are usually diminished or absent. A pancreatic pseudocyst may be palpable in the upper abdomen. A faint blue discoloration around the umbilicus (Cullen's sign) may occur as the result of hemoperitoneum, and a blue-red-purple or green-brown discoloration of the flanks (Turner's sign) reflects tissue catabolism of hemoglobin. The latter two findings, which are uncommon, indicate the presence of a severe necrotizing pancreatitis.

LABORATORY DATA

The diagnosis of acute pancreatitis is usually established by the detection of an increased level of serum amylase. Values threefold or more above normal virtually clinch the diagnosis if overt salivary gland disease and gut perforation or infarction are excluded. However, there appears to be no definite correlation between the severity of pancreatitis and the degree of serum amylase elevation. After 48 to 72 h, even with continuing evidence of pancreatitis, total serum amylase values tend to return to normal. However, pancreatic isoamylase and lipase levels may remain elevated for 7 to 14 days. It will be recalled that amylase elevations in serum and urine occur in many conditions other than pancreatitis (seeTable 303-2). Importantly, patients with *acidemia* (arterial pH £7.32) may have spurious elevations in serum amylase. In one study, 12 of 33 patients with acidemia had elevated serum amylase, but only 1 had an elevated lipase value; in 9, salivary-type amylase was the predominant serum isoamylase. This finding explains why patients with diabetic ketoacidosis may have marked elevations in serum amylase without any other evidence of acute pancreatitis. Serum lipase activity increases in parallel with amylase activity, and measurement of both enzymes increases the diagnostic yield. An elevated serum lipase or trypsin value is usually diagnostic of acute pancreatitis; these tests are especially helpful in patients with nonpancreatic causes of hyperamylasemia (seeTable 303-2). Markedly increased levels of peritoneal or pleural fluid amylase [>1500 nmol/L (> 5000 U/dL)] are also helpful, if present, in establishing the diagnosis.

Leukocytosis (15,000 to 20,000 leukocytes per microliter) occurs frequently. Patients with more severe disease may show hemoconcentration with hematocrit values exceeding 50% because of loss of plasma into the retroperitoneal space and peritoneal cavity. *Hyperglycemia* is common and is due to multiple factors, including decreased insulin release, increased glucagon release, and an increased output of adrenal glucocorticoids and catecholamines. *Hypocalcemia* occurs in approximately 25% of patients, and its pathogenesis is incompletely understood. Although earlier studies suggested that the response of the parathyroid gland to a decrease in serum calcium is impaired, subsequent observations have failed to confirm this idea. Intraperitoneal saponification of calcium by fatty acids in areas of fat necrosis occurs occasionally, with large amounts (up to 6.0 g) dissolved or suspended in ascitic fluid. Such "soap formation" also may be significant in patients with pancreatitis, mild hypocalcemia, and little or no obvious ascites. *Hyperbilirubinemia* [serum bilirubin>68 umol/L (> 4.0 mg/dL)] occurs in approximately 10% of patients. However, jaundice is transient, and serum bilirubin levels return to normal in 4 to 7 days. Serum alkaline phosphatase and aspartate aminotransferase (AST) levels are also transiently elevated and parallel serum bilirubin values. Markedly elevated serum lactic dehydrogenase (LDH) levels [>8.5 umol/L (> 500 U/dL)] suggest a poor prognosis. Serum albumin is decreased to £30 g/L (£3.0 g/dL) in about 10% of patients; this sign is associated with more severe pancreatitis and a higher mortality rate (Table 304-2). *Hypertriglyceridemia* occurs in 15 to 20% of patients, and serum amylase levels in these individuals are often spuriously normal (Chap. 303). Most patients with hypertriglyceridemia and pancreatitis, when subsequently examined, show evidence of an underlying derangement in lipid metabolism which probably antedated the pancreatitis (see below). Approximately 25% of patients have *hypoxemia* (arterial P_{O_2}£ 60 mmHg), which may herald the onset of adult respiratory distress syndrome. Finally, the electrocardiogram is occasionally abnormal in acute pancreatitis with ST-segment and T-wave abnormalities simulating myocardial ischemia.

Although one or more radiologic abnormalities are found in over 50% of patients, the findings are inconstant and nonspecific. The chief value of conventional x-rays [chest films; kidney, ureter, and bladder (KUB) studies] in acute pancreatitis is to help exclude other diagnoses, especially a perforated viscus. Upper gastrointestinal tract x-rays have been superseded by ultrasonography and computed tomography (CT). A CT scan may confirm the clinical impression of acute pancreatitis even in the face of normal serum amylase levels. Importantly, CT is quite helpful in indicating the severity of acute pancreatitis and the risk of morbidity and mortality (see below). Sonography and radionuclide scanning [*N*-*p*-isopropylacetanilide-iminodiacetic acid (PIPIDA) scan; hepatic 2,6-dimethyliminodiacetic acid (HIDA) scan] are useful in acute pancreatitis to evaluate the gallbladder and biliary tree.**Radiologic studies useful in the diagnosis of acute pancreatitis are discussed in Chap. 303 and listed in Table 303-1.*

DIAGNOSIS

Any severe acute pain in the abdomen or back should suggest acute pancreatitis. The diagnosis is usually entertained when a patient with a possible predisposition to pancreatitis presents with severe and constant abdominal pain, nausea, emesis, fever, tachycardia, and abnormal findings on abdominal examination. Laboratory studies frequently reveal leukocytosis, an abnormal appearance on x-rays of the abdomen and chest, hypocalcemia, and hyperglycemia. The diagnosis is usually confirmed by the finding of an elevated level of serum amylase and/or lipase. Not all the above features have to be present for the diagnosis to be established.

The *differential diagnosis* should include the following disorders: (1) perforated viscus, especially peptic ulcer; (2) acute cholecystitis and biliary colic; (3) acute intestinal obstruction; (4) mesenteric vascular occlusion; (5) renal colic; (6) myocardial infarction; (7) dissecting aortic aneurysm; (8) connective tissue disorders with vasculitis; (9) pneumonia; and (10) diabetic ketoacidosis. A penetrating duodenal ulcer usually can be identified by upper gastrointestinal x-rays and/or endoscopy. A perforated duodenal ulcer is readily diagnosed by the presence of free intraperitoneal air. It may be difficult to differentiate acute cholecystitis from acute pancreatitis, since an elevated serum amylase may be found in both disorders. Pain of biliary tract origin is more right-sided and gradual in onset, and ileus is usually absent; sonography and radionuclide scanning are helpful in establishing the diagnosis of cholelithiasis and cholecystitis. Intestinal obstruction due to mechanical factors can be differentiated from pancreatitis by the history of colicky pain, findings on abdominal examination, and x-rays of the abdomen showing changes characteristic of mechanical obstruction. Acute mesenteric vascular occlusion is usually evident in elderly debilitated patients with brisk leukocytosis, abdominal distention, and bloody diarrhea, in whom paracentesis shows sanguineous fluid and arteriography shows vascular occlusion. Serum as well as peritoneal fluid amylase levels are increased, however, in patients with intestinal infarction. Systemic lupus erythematosus and polyarteritis nodosa may be confused with pancreatitis, especially since pancreatitis may develop as a complication of these diseases. Diabetic ketoacidosis is often accompanied by abdominal pain and elevated total serum amylase levels, thus closely mimicking acute pancreatitis. However, the serum lipase and pancreatic isoamylase levels are not elevated in diabetic ketoacidosis.

COURSE OF THE DISEASE AND COMPLICATIONS

It is important to identify patients with acute pancreatitis who have an increased risk of dying. Ranson and Imrie have used multiple prognostic criteria and have demonstrated that there is an increased mortality rate when three or more risk factors are identifiable either at the time of admission to the hospital or during the initial 48 h of hospitalization (Table 304-2). Recent studies indicate that obesity is a major risk factor for severe pancreatitis, presumably because the increased deposits of peripancreatic fat in such patients may predispose them to more extensive pancreatic and peripancreatic necrosis. The acute physiology and chronic health evaluation scoring system (APACHE II) uses the worst values of 12 physiologic measurements plus age and previous health status and provides a good description of illness severity for a wide range of common diseases; this score also correlates with outcome. Prospective studies have compared APACHE II with multiple prognostic criteria, i.e., Ranson and Imrie scores, in predicting the severity of acute pancreatitis. On admission, APACHE II identified approximately two-thirds of severe attacks, and after 48 h, the prognostic accuracy of APACHE II is comparable with that of Ranson and Imrie's scoring system. The drawbacks of APACHE II are (1) its complexity, (2) the requirement of a computer for scoring, and (3) standardization regarding peak values and cutoff scores. McMahon and colleagues have shown that the presence of a "toxic broth" or dark (hemorrhagic) fluid in abdominal pancreatitis is also an important prognostic indicator in acute pancreatitis. These multiple-factor scoring systems are difficult to use and have not been embraced consistently by clinicians. There is a great need for a reliable, simple biochemical test that consistently predicts outcome in patients with acute pancreatitis. Three candidate markers that show great promise are C-reactive protein, serum granulocyte elastase, and urinary trypsinogen activation peptide (TAP). The key indicators of a severe attack of acute pancreatitis are also listed inTable 304-2. Importantly, the presence of any one of these factors is associated with an increased risk of complications, and the presence of any two, with a 20 to 30% mortality rate. The high mortality rate of such severely ill patients is due in large part to infection and warrants intensive radiologic intervention and monitoring and/or a combination of radiologic and surgical means, as discussed in detail below.

The local and systemic complications of acute pancreatitis are listed in Table 304-3. In the first 2 to 3 weeks after pancreatitis patients frequently develop an inflammatory mass, which may be due to pancreatic necrosis (with or without infection) or may represent an abscess or pseudocyst (see below). Systemic complications include pulmonary, cardiovascular, hematologic, renal, metabolic, and central nervous system abnormalities. Pancreatitis and hypertriglyceridemia constitute an association in which cause and effect remain incompletely understood. However, several reasonable conclusions can be drawn. First, hypertriglyceridemia can precede and apparently cause pancreatitis. Second, the vast majority (>80%) of patients with acute pancreatitis do not have hypertriglyceridemia. Third, almost all patients with pancreatitis and hypertriglyceridemia have preexisting abnormalities in lipoprotein metabolism. Fourth, many of the patients with this association have persistent hypertriglyceridemia after recovery from pancreatitis and are prone to recurrent episodes of pancreatitis. Fifth, any factor (e.g., drugs or alcohol) that causes an abrupt increase in serum triglycerides to levels greater than 11 mmol/L (1000 mg/dL) can precipitate a bout of pancreatitis that can be associated with significant complications and even become fulminant. To avert the risk of triggering pancreatitis, a fasting serum triglyceride measurement should be obtained before estrogen replacement therapy is begun in postmenopausal women. Fasting levels less than 300 mg/dL pose no risk, whereas levels greater than 750 mg/dL are associated with a high probability of developing pancreatitis. Finally, patients with a deficiency of apolipoprotein CII have an increased incidence of pancreatitis; apolipoprotein CII activates lipoprotein lipase, which is important in clearing chylomicrons from the bloodstream.

Purtscher's retinopathy, a relatively unusual complication, is manifested by a sudden and severe loss of vision in a patient with acute pancreatitis. It is characterized by a peculiar funduscopic appearance with cotton-wool spots and hemorrhages confined to an area limited by the optic disk and macula; it is believed to be due to occlusion of the posterior retinal artery with aggregated granulocytes.

The two most common causes of acute pancreatitis are alcoholism and biliary tract disease; other causes are listed in Table 304-1. The risk of acute pancreatitis in patients with at least one gallstone smaller than 5 mm in diameter is fourfold greater than that in patients with larger stones. However, after a conventional workup, a specific cause is not identified in about 30% of patients. It is important to note that ultrasound examinations fail to detect gallstones, especially microlithiasis and/or sludge, in 4 to 7% of patients. In one series of 31 patients diagnosed initially as having idiopathic acute pancreatitis, 23 were found to have occult gallstone disease. Thus, approximately two-thirds of patients with recurrent acute pancreatitis without an obvious cause actually have occult gallstone disease due to microlithiasis. Examination of duodenal aspirates in such cases often reveals cholesterol crystals, which confirm the diagnosis. Other diseases of the biliary tree and pancreatic ducts that can cause acute pancreatitis include choledochocele, ampullary tumors, pancreas divisum, and pancreatic duct stones, stricture, and tumor. Approximately 2% of patients with pancreatic carcinoma present with acute pancreatitis.

Pancreatitis in Patients with AIDS The incidence of acute pancreatitis is increased in patients with AIDS for two reasons: (1) the high incidence of infections involving the pancreas, such as infections with cytomegalovirus, *Cryptosporidium*, and the *Mycobacterium avium* complex; and (2) the frequent use by patients with AIDS of medications such as didanosine, pentamidine, and trimethoprim-sulfamethoxazole (Chap. 309).

TREATMENT

In most patients (approximately 85 to 90%) with acute pancreatitis, the disease is self-limited and subsides spontaneously, usually within 3 to 7 days after treatment is instituted. Conventional measures include (1) analgesics for pain, (2) intravenous fluids and colloids to maintain normal intravascular volume, (3) no oral alimentation, and (4) nasogastric suction to decrease gastrin release from the stomach and prevent gastric contents from entering the duodenum. Recent controlled trials, however, have shown that nasogastric suction offers no clear-cut advantages in the treatment of mild to moderately severe acute pancreatitis. Its use, therefore, must be considered elective rather than mandatory.

It has been demonstrated thatCCK-stimulated pancreatic secretion is almost abolished in four different experimental models of acute pancreatitis. This finding probably explains why drugs to block pancreatic secretion in acute pancreatitis have failed to have any therapeutic benefit. For this and other reasons, anticholinergic drugs are not indicated in acute pancreatitis. In addition to nasogastric suction and anticholinergic drugs, other therapies designed to "rest the pancreas" by inhibiting pancreatic secretion have not changed the course of the disease. Although antibiotics have been used in the treatment of acute pancreatitis, randomized, prospective trials have shown no benefit from their use in acute pancreatitis of mild to moderate severity.

However, current experimental evidence favors the use of prophylactic antibiotics in severe acute pancreatitis. Results of four contemporary randomized clinical trials restricted to patients with prognostically severe acute pancreatitis have demonstrated an improved outcome, i.e., reduced rate of infection and/or mortality, associated with the antibiotic treatment. The carbapenem group of antibiotics, including imipenem, has a very broad spectrum including activity against *Pseudomonas*, *Staphylococcus*, and *Enterococcus*; and these agents penetrate well into pancreatic tissue. Furthermore, because secondary infection of necrotic pancreatic tissue (abscess, pseudocyst or obstructed biliary passages, ascending cholangitis complicating choledocholithiasis) contributes to many of the late deaths from pancreatitis, appropriate antibiotic therapy of established infections is quite important.

Several other drugs have been evaluated by prospective controlled trials and found ineffective in the treatment of acute pancreatitis. The list, by no means complete, includes glucagon, H_2blockers, protease inhibitors such as aprotinin, glucocorticoids, calcitonin, nonsteroidal anti-inflammataory drugs (NSAIDs) and lexiplafant, a platelet-activating factor inhibitor. A recent meta-analysis of somatostatin, ocreotide, and the antiprotease gabexate methylate in therapy of acute pancreatitis suggested (1) a reduced mortality rate but no change in complications with octreotide, and (2) no effect on the mortality rate but reduced pancreatic damage with gabexate.

Intraabdominal *Candida* infection during acute necrotizing pancreatitis is increasing in frequency and is associated with an increased mortality rate. In one representatitve trial, intraabdominal *Candida* infection was found in 13 of 37 cases and was associated with a mortality rate fourfold greater than that associated with intraabdominal bacterial infection alone. Given the impact of *Candida* infection on the mortality rate in acute necrotizing pancreatitis and the apparent benefit of prophylactic chemotherapy, these data suggest earlier use of fungicides.

ACT scan, especially a contrast-enhanced dynamic CT (CECT) scan, provides valuable information on the severity and prognosis of acute pancreatitis (Fig. 304-1 andTable 304-4). In particular, a CECT scan allows estimation of the presence and extent of pancreatic necrosis. Recent studies suggest that the likelihood of prolonged pancreatitis or a serious complication is negligible when the CT severity index is 1 or 2 and low with scores of 3 to 6. However, patients with scores of 7 to 10 had a 92% morbidity rate and a 17% mortality rate. Necrosis is present in 20 to 30% of patients. Those with necrosis have a morbidity rate>20%, whereas those without necrosis have a morbidity rate <10% and a negligible mortality rate. A CECT scan is indicated in patients with three or more of Ranson's signs, in all seriously ill patients, and in patients who show evidence of clinical deterioration. The patient with mild to moderate pancreatitis usually requires treatment with intravenous fluids, fasting, and possibly nasogastric suction for 2 to 4 days. A clear liquid diet is frequently started on the third to sixth day, and a regular diet by the fifth to seventh day. The patient with unremitting *fulminant pancreatitis* usually requires inordinate amounts of fluid and close attention to complications such as cardiovascular collapse, respiratory insufficiency, and pancreatic infection. The latter should be managed by a combination of radiologic and surgical means (see below). While earlier uncontrolled studies suggested that *peritoneal lavage* through a percutaneous dialysis catheter was helpful in severe pancreatitis, subsequent studies indicate that this treatment does not influence the outcome of such attacks. Aggressive surgical pancreatic debridement (necrosectomy) should be undertaken soon after confirmation of the presence of infected necrosis, and multiple operations may be required. Since the mortality rate from sterile acute necrotizing pancreatitis is approximately 10%, laparotomy with adequate drainage and removal of necrotic tissue should be considered if conventional therapy does not halt the patient's deterioration. The use of parenteral nutrition makes it possible to give nutritional support to patients with severe, acute, or protracted pancreatitis who are unable to eat normally. Patients with severe gallstone-induced pancreatitis may improve dramatically if papillotomy is carried out within the first 36 to 72 h of the attack. Studies indicate that only those patients with gallstone pancreatitis who are in the very severe group should be considered for urgent endoscopic retrograde cholangiopancreatography (ERCP). Finally, the treatment for patients with hypertriglyceridemia-associated pancreatitis includes (1) weight loss to ideal weight, (2) a lipid-restricted diet, (3) exercise, (4) avoidance of alcohol and of drugs that can elevate serum triglycerides (i.e., estrogens, vitamin A, thiazides, and beta-blockers), and (5) control of diabetes.

INFECTED PANCREATIC NECROSIS, ABSCESS, AND PSEUDOCYST

Infected pancreatic necrosis should be differentiated from pancreatic abscess. The former is a diffuse infection of an acutely inflamed, necrotic pancreas occurring in the first 1 to 2 weeks after the onset of pancreatitis. In contrast, a pancreatic abscess is an ill-defined, liquid collection of pus that evolves over a longer period, often 4 to 6 weeks. It tends to be less life-threatening and is associated with a lower rate of surgical mortality. Infected pancreatic necrosis should be treated by surgical debridement because the solid component of the infected pancreas is not amenable to effective radiologically guided percutaneous evacuation. Pancreatic abscess can be treated surgically or, in selected cases, by percutaneous drainage. The necrotic pancreas becomes secondarily infected in 40 to 60% of patients, most frequently with gram-negative bacteria of alimentary origin. Whether infection occurs depends on several factors, including the extent of pancreatic and peripancreatic necrosis, the degree of pancreatic ischemia and hypoperfusion, and the presence of organ or multiorgan failure.

The early diagnosis of pancreatic infection can be accomplished byCT-guided needle aspiration. In one study, 60 patients, representing 5% of all admissions for acute pancreatitis, were suspected of harboring a pancreatic infection on the basis of fever, leukocytosis, and an abnormal CT scan (pseudocyst or extrapancreatic fluid collection). Importantly, 60% of these patients had a pancreatic infection, and 55% of these infections developed in the first 2 weeks. These findings suggest that only guided aspiration can reliably distinguish sterile from infected pancreatic necrosis. The following are guidelines for patients meeting the above selection criteria: (1) Pseudocysts should be aspirated promptly, because more than half may be infected; (2) extrapancreatic fluid collections need not be aspirated promptly, because most are sterile; (3) if a necrotic pancreas is found initially to be sterile but fever and leukocytosis persist, several days of observation should be allowed to pass before reaspiration is considered, as clinical improvement frequently occurs; and (4) if fever and leukocytosis recur after an interval of well-being, reaspiration should be considered.

Severe pancreatitis with the presence of three or more risk factors, postoperative pancreatitis, early oral feeding, early laparotomy, and perhaps injudicious use of antibiotics predispose to the development of pancreatic abscess, which occurs in 3 to 4% of patients with acute pancreatitis. Pancreatic abscess also may develop because of a communication between a pseudocyst and the colon, inadequate surgical drainage of a pseudocyst, or needling of a pseudocyst. The characteristic signs of abscess are fever, leukocytosis, ileus, and rapid deterioration in a patient previously recovering from pancreatitis. Sometimes, however, the only manifestations are persistent fever and signs of continuing pancreatic inflammation. Drainage of pancreatic abscesses by percutaneous catheter techniques, usingCTguidance, has been only moderately successful (resolution in 50 to 60% of patients). Accordingly, laparotomy with radical sump drainage and possibly resection of necrotic tissue is usually required, because the mortality rate for undrained pancreatic abscess approaches 100%. Multiple abscesses are common, and reoperation is frequently necessary.

Pseudocysts of the pancreas are collections of tissue, fluid, debris, pancreatic enzymes, and blood which develop over a period of 1 to 4 weeks after the onset of acute pancreatitis; they form in approximately 15% of patients with acute pancreatitis. In contrast to true cysts, pseudocysts do not have an epithelial lining; their walls consist of necrotic tissue, granulation tissue, and fibrous tissue. Disruption of the pancreatic ductal system is common. However, the subsequent course of this disruption varies widely, ranging from spontaneous healing to continuous leakage of pancreatic juice, which results in tense ascites. Pseudocysts are preceded by pancreatitis in 90% of cases and by trauma in 10%. Approximately 85% are located in the body or tail of the pancreas and 15% in the head. Some patients have two or more pseudocysts. Abdominal pain, with or without radiation to the back, is the usual presenting complaint. A palpable, tender mass may be found in the middle or left upper abdomen. The serum amylase level is elevated in 75% of patients at some point during their illness and may fluctuate markedly.

On x-ray examination, 75% of pseudocysts can be seen to displace some portion of the gastrointestinal tract (Fig. 304-2). Sonography, however, is reliable in detecting pseudocysts. Sonography also permits differentiation between an edematous, inflamed pancreas, which can give rise to a palpable mass, and an actual pseudocyst. Furthermore, serial ultrasound studies will indicate whether a pseudocyst has resolved.CTcomplements ultrasonography in the diagnosis of pancreatic pseudocyst (Fig. 304-2), especially when the pseudocyst is infected.

In studies with sonography, pseudocysts were seen to resolve in 25 to 40% of patients. Pseudocysts that are greater than 5 cm in diameter and that persist for longer than 6 weeks should be considered for drainage. Recent natural history studies have suggested that noninterventional, expectant management is the best course in selected patients with minimal symptoms and no evidence of active alcohol use in whom the pseudocyst appears mature by radiography and does not resemble a cystic neoplasm. A significant number of these pseudocysts resolve spontaneously more than 6 weeks after their formation. Also, these studies demonstrate that large pseudocyst size is not an absolute indication for interventional therapy and that many peripancreatic fluid collections detected onCT in cases of acute pancreatitis resolve spontaneously. A pseudocyst that does not resolve spontaneously may lead to serious complications, such as (1) pain caused by expansion of the lesion and pressure on other viscera, (2) rupture, (3) hemorrhage, and (4) abscess. Rupture of a pancreatic pseudocyst is a particularly serious complication. Shock almost always supervenes, and mortality rates range from 14% if the rupture is not associated with hemorrhage to over 60% if hemorrhage has occurred. Rupture and hemorrhage are the prime causes of death from pancreatic pseudocyst. A triad of findings -- an increase in the size of the mass, a localized bruit over the mass, and a sudden decrease in hemoglobin level and hematocrit without obvious external blood loss -- should alert one to the possibility of hemorrhage from a pseudocyst. Thus, in patients who are stable and free of complications and in whom serial ultrasound studies show that the pseudocyst is shrinking, conservative therapy is indicated. Conversely, if the pseudocyst is expanding and is complicated by rupture, hemorrhage, or abscess, the patient should be operated on. With ultrasound or CT guidance, sterile chronic pseudocysts can be treated safely with single or repeated needle aspiration or more prolonged catheter drainage with a success rate of 45 to 75%. The success rate of these techniques for infected pseudocysts is considerably less (40 to 50%). Patients who do not respond to drainage require surgical therapy for internal or external drainage of the cyst.

Pseudoaneurysms develop in up to 10% of patients with acute pancreatitis at sites reflecting the distribution of pseudocysts and fluid collections (Fig. 304-2*D*). The splenic artery is most frequently involved, followed by the inferior and superior pancreatic duodenal arteries. This diagnosis should be suspected in patients with pancreatitis who develop upper gastrointestinal bleeding without an obvious cause or in whom thin-cutCTscanning reveals a contrast-enhanced lesion within or adjacent to a suspected pseudocyst. Arteriography is necessary to confirm the diagnosis.

PANCREATIC ASCITES AND PANCREATIC PLEURAL EFFUSIONS

Pancreatic ascites is usually due to disruption of the main pancreatic duct, often by an internal fistula between the duct and the peritoneal cavity or a leaking pseudocyst (Chap. 43). This diagnosis is suggested in a patient with an elevated serum amylase level in whom the ascites fluid has both increased levels of albumin [>30 g/L (>3.0 g/dL)] and a markedly elevated level of amylase. The fluid in true pancreatic ascites usually has an amylase concentration of >20,000 U/L as a result of the ruptured duct or leaking pseudocyst. Lower amylase elevations may be found in the peritoneal fluid of patients with acute pancreatitis. In addition,ERCPoften demonstrates passage of contrast material from a major pancreatic duct or a pseudocyst into the peritoneal cavity. As many as 15% of patients with pseudocysts have concurrent pancreatic ascites. The differential diagnosis should include intraperitoneal carcinomatosis, tuberculous peritonitis, constrictive pericarditis, and Budd-Chiari syndrome.

If the pancreatic duct disruption is posterior, an internal fistula may develop between the pancreatic duct and the pleural space, producing a pleural effusion, which is usually left-sided and often massive. This complication often requires thoracentesis or chest tube drainage.

Treatment usually requires the use of nasogastric suction and parenteral alimentation to decrease pancreatic secretion. In addition, paracentesis is performed to keep the peritoneal cavity free of fluid and, it is hoped, to effect sealing of the leak. The long-acting somatostatin analogue octreotide, which inhibits pancreatic secretion, is useful in cases of pancreatic ascites and pleural effusion. If ascites continues to recur after 2 to 3 weeks of medical management, the patient should be operated on after pancreatography to define the anatomy of the abnormal duct. A disrupted main pancreatic duct can also be treated effectively by stenting. Patients in whomERCPidentifies two or more sites of extravasation are unlikely to respond to conservative management and/or stenting.

CHRONIC PANCREATITIS AND PANCREATIC EXOCRINE INSUFFICIENCY

GENERAL AND ETIOLOGIC CONSIDERATIONS

Chronic inflammatory disease of the pancreas may present as episodes of acute inflammation in a previously injured pancreas or as chronic damage with persistent pain or malabsorption. The causes of relapsing chronic pancreatitis are similar to those of acute pancreatitis (Table 304-1), except that there is an appreciable incidence of cases of undetermined origin. In addition, the pancreatitis associated with gallstones is predominantly acute or relapsing-acute in nature. A cholecystectomy is almost always performed in patients after the first or second attack of gallstone-associated pancreatitis. Patients with chronic pancreatitis may present with persistent abdominal pain, with or without steatorrhea; some (~15%) present with steatorrhea and no pain.

Patients with chronic pancreatitis in whom there is extensive destruction of the pancreas (less than 10% of exocrine function remaining) have steatorrhea and azotorrhea. Among American adults, alcoholism is the most common cause of clinically apparent pancreatic exocrine insufficiency, while cystic fibrosis is the most frequent cause in children. In up to 25% of American adults with chronic pancreatitis, the cause is not known; that is, they have idiopathic chronic pancreatitis. Mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene have been documented in patients with idiopathic chronic pancreatitis. It has been estimated that in patients with idiopathic pancreatitis the frequency of a single CFTR mutation is 11 times the expected frequency and the frequency of two mutant alleles is 80 times the expected frequency. The results of sweat chloride testing are not diagnostic of cystic fibrosis in these patients. However, these patients have functional evidence of a defect in CFTR-mediated ion transport in nasal epithelium. It is suggested that up to 25% of patients with idiopathic chronic pancreatitis may have abnormalities of the CFTR gene. The therapeutic and prognostic implication of these findings remain to be determined. In other parts of the world, severe protein-calorie malnutrition is a common cause.Table 304-5 lists other causes of pancreatic exocrine insufficiency, but they are relatively uncommon.

PATHOPHYSIOLOGY

The events that initiate an inflammatory process in the pancreas are still not well understood, and the many hypotheses will not be reviewed here. In the case of alcohol-induced pancreatitis, it has been suggested that the primary defect may be the precipitation of protein (inspissated enzymes) in the ducts. The resulting ductal obstruction could lead to duct dilation, diffuse atrophy of the acinar cells, fibrosis, and eventual calcification of some of the protein plugs. However, the fact that some alcoholic patients with recurrent acute pancreatitis show no evidence of chronic pancreatitis does not support this hypothesis. In fact, experimental and clinical observations have shown that alcohol has direct toxic effects on the pancreas. While patients with alcohol-induced pancreatitis generally consume large amounts of alcohol, some consume very little (50 g/d or less). Thus prolonged consumption of "socially acceptable" amounts of alcohol is compatible with the development of pancreatitis. In addition, the finding of extensive pancreatic fibrosis in patients who died during their first attack of clinical acute alcohol-induced pancreatitis supports the concept that such patients already have chronic pancreatitis.

CLINICAL FEATURES

Patients with relapsing chronic pancreatitis may present with symptoms identical to those of acute pancreatitis, but pain may be continuous, intermittent, or absent. The pathogenesis of this pain is poorly understood. Although the classic description is of epigastric pain radiating through the back, the pain pattern is often atypical; the pain may be worst in the right or left upper quadrant of the back or may be diffuse throughout the upper abdomen; it may even be referred to the anterior chest or flank. Characteristically it is persistent, deep-seated, and unresponsive to antacids. It often is worsened by ingestion of alcohol or a heavy meal (especially one rich in fat). Often the pain is severe enough to necessitate the frequent use of narcotics.

Weight loss, abnormal stools, and other signs or symptoms suggestive of malabsorption (seeTable 286-5) are common in chronic pancreatitis. However, clinically apparent deficiencies of fat-soluble vitamins are surprisingly rare. The physical findings in these patients are usually not impressive, so that there is a disparity between the severity of the abdominal pain and the physical signs (other than some abdominal tenderness and mild temperature elevation).

DIAGNOSTIC EVALUATION (See also Chap. 303)

In contrast to relapsing acute pancreatitis, the serum amylase and lipase levels are usually not elevated in chronic pancreatitis. Elevations of serum bilirubin and alkaline phosphatase levels may indicate cholestasis secondary to chronic inflammation around the common bile duct (Fig. 304-3). Many patients demonstrate impaired glucose tolerance, and some have an elevated fasting blood glucose level.

The classic triad of pancreatic calcification, steatorrhea, and diabetes mellitus usually establishes the diagnosis of chronic pancreatitis and exocrine pancreatic insufficiency but is found in less than one-third of chronic pancreatitis patients. Accordingly, it is often necessary to perform an intubation test such as the *secretin stimulation test*, which usually gives abnormal results when 60% or more of pancreatic exocrine function has been lost. Approximately 40% of patients with chronic pancreatitis have *cobalamin* (*vitamin* B_{12}) malabsorption, which can be corrected by the administration of oral pancreatic enzymes. There is usually a marked excretion of fecal fat (Chap. 286), which can be reduced by the administration of oral pancreatic enzymes. The serum trypsinogen (Chap. 303) and the D-xylose urinary excretion test are useful in patients with "pancreatic steatorrhea," since the trypsinogen level will be abnormal, and D-xylose excretion usually is normal. A decreased serum trypsinogen level strongly suggests severe pancreatic exocrine insufficiency.

The radiographic hallmark of chronic pancreatitis is the presence of scattered calcification throughout the pancreas (Fig. 304-3). Diffuse pancreatic calcification indicates that significant damage has occurred and obviates the need for a secretin test. While alcohol is by far the most common cause, pancreatic calcification also may be seen in cases of severe protein-calorie malnutrition, hereditary pancreatitis, posttraumatic pancreatitis, hyperparathyroidism, islet cell tumors, and idiopathic chronic pancreatitis. A large prospective study has shown convincingly that pancreatic calcification decreases or even disappears spontaneously in one-third of patients with severe chronic pancreatitis; this outcome may also follow ductal decompression. Pancreatic calcification is a dynamic process that is incompletely understood.

Sonography,CT, andERCPgreatly aid the diagnosis of pancreatic disease. In addition to excluding pseudocysts and pancreatic cancer, sonography and CT may show calcification or dilated ducts associated with chronic pancreatitis (Fig. 304-4). ERCP is the only major technique that provides a direct view of the pancreatic duct. In patients with alcohol-induced pancreatitis, ERCP may reveal a pseudocyst missed by sonography or CT.

COMPLICATIONS OF CHRONIC PANCREATITIS

The complications of chronic pancreatitis are protean. *Cobalamin* (*vitamin* B_{12}) malabsorption occurs in 40% of patients with alcohol-induced chronic pancreatitis and in virtually all with cystic fibrosis. It is consistently corrected by the administration of pancreatic enzymes (containing proteases). It may be due to excessive binding of cobalamin by cobalamin-binding proteins other than intrinsic factor, which ordinarily are destroyed by pancreatic proteases and therefore do not compete with intrinsic factor for cobalamin binding. Although most patients show *impaired glucose tolerance*, diabetic ketoacidosis and coma are uncommon. Similarly, end-organ damage (retinopathy, neuropathy, nephropathy) is also uncommon, and the appearance of these complications should raise the question of concomitant genetic diabetes mellitus. A nondiabetic retinopathy, peripheral in location and secondary to vitamin A and/or zinc deficiency, is common in these patients. *Effusions* containing high concentrations of amylase may occur into the pleural, pericardial, or peritoneal space. *Gastrointestinal bleeding* may occur from peptic ulceration, gastritis, a pseudocyst eroding into the duodenum, or ruptured varices secondary to splenic vein thrombosis due to inflammation of the tail of the pancreas. *Icterus* may occur, caused either by edema of the head of the pancreas, which compresses the common bile duct, or by chronic cholestasis secondary to a chronic inflammatory reaction around the intrapancreatic portion of the common bile duct (Fig. 304-3). The chronic obstruction may lead to cholangitis and ultimately to biliary cirrhosis. *Subcutaneous fat necrosis* may appear as tender red nodules on the lower extremities. *Bone pain* may be secondary to intramedullary fat necrosis. Inflammation of the large and small joints of the upper and lower extremities may occur. The incidence of pancreatic carcinoma is increased in patients with chronic pancreatitis who have been followed for 2 or more years. Twenty years after the diagnosis of chronic pancreatitis, the cumulative risk of pancreatic carcinoma is 4%. Perhaps the most common and troublesome complication is addiction to narcotics.

TREATMENT

Therapy for patients with chronic pancreatitis is directed toward two major problems -- pain and malabsorption. Patients with intermittent attacks of pain are treated essentially like those with acute pancreatitis (see above). Patients with severe and persistent pain should avoid alcohol completely and avoid large meals rich in fat. Since the pain is often severe enough to require frequent use of narcotics (and hence addiction), a number of surgical procedures have been developed for pain relief.ERCPallows the surgeon to plan the operative approach. If there is a stricture of the pancreatic duct, a *local resection* may ameliorate the pain. Unfortunately, isolated localized strictures are not common. In most patients with alcohol-induced disease, the pancreas is diffusely involved, and surgically correctible localized ductal disease is rare. When there is primary ductal obstruction and dilation, ductal decompression may provide effective pain palliation. Short-term pain relief may be achieved in up to 80% of patients, while long-term pain relief occurs in approximately 50%. In some of these patients, however, pain relief can be achieved only by resecting 50 to 95% of the gland. Although pain relief is achieved in three-quarters of these patients, they tend to develop pancreatic endocrine and exocrine insufficiency and must be treated with pancreatic enzyme replacement therapy. It is important to screen patients carefully, for such radical surgery is contraindicated in those who are severely depressed or suicidal or who continue to drink. Procedures such as splanchnicectomy, celiac ganglionectomy, and nerve blocks usually bring only temporary relief and are not recommended. Endoscopic treatment of chronic pancreatitis may involve sphincterotomy of the minor or major pancreatic sphincter, dilatation of strictures, removal of calculi, or stenting of the ventral or dorsal pancreatic duct. Although many of these techniques are technically impressive, none has been subjected to a randomized trial in patients with chronic pancreatitis. In addition, significant complications -- acute pancreatitis, pancreatic abscess, damage to the pancreatic duct, and death -- have occurred in up to 36% of patients after stent placement.

Three double-blind trials have demonstrated that administration of pancreatic enzymes decreases abdominal pain in selected patients with chronic pancreatitis. In these trials, approximately 75% of the patients evaluated experienced pain relief. The patients most likely to respond are those with mild to moderate exocrine pancreatic dysfunction, as evidenced by an abnormal secretin test, normal fat absorption, and minimal abnormalities onERCPexamination. These clinical observations seem to fit with data from human beings and experimental animals demonstrating a negative feedback regulation for pancreatic exocrine secretion controlled by the amount of proteases within the lumen of the proximal small intestine. It seems reasonable to use the following approach for patients with severe, persistent, or continuous abdominal pain thought to be caused by chronic pancreatitis. After other causes of abdominal pain (peptic ulcer, gallstones, etc.) have been excluded, a pancreatic *sonogram* should be done. If no mass is found, a *secretin test* may be performed, because its results usually are abnormal in cases of chronic pancreatitis with pain. If the results are abnormal (i.e., decreased bicarbonate concentration or volume output), a 3- to 4-week *trial of pancreatic enzyme administration* is appropriate. Eight conventional tablets or capsules are taken at meals and at bedtime. There are a number of studies suggesting that patients may have small-duct chronic pancreatitis and chronic abdominal pain with a normal appearance on radiographic evaluations (ultrasound,CT, ERCP) but abnormal results on hormone stimulation tests (secretin test) and/or abnormal pancreatic histology. Such minimal-change chronic pancreatitis may respond well to pancreatic enzyme therapy (non-enteric-coated) for relief of abdominal pain. If no relief is obtained, and especially if the volume secreted during the secretin test is very low, ERCP should be performed. If a pseudocyst or a localized ductal obstruction is found, surgery should be considered. A patient who has dilated ducts may be a candidate for a surgical ductal decompression procedure. This procedure provides short-term relief in up to 80% of patients, although long-term results are closer to 50%. Some studies have shown octreotide to be effective in decreasing abdominal pain in patients with severe large-duct disease. If no surgically remediable lesion is found and severe pain continues despite abstinence from alcohol, subtotal pancreatic resection may be necessary.

The treatment of malabsorption rests on the use of pancreatic enzyme replacement therapy. Diarrhea and steatorrhea are usually improved by this treatment, although the steatorrhea may not be completely corrected. The major problem is delivering enough active enzyme into the duodenum. Steatorrhea could be abolished if 10% of the normal amount of lipase could be delivered to the duodenum at the proper time. This concentration of lipase cannot be achieved with the current preparations of pancreatic enzymes, even if the latter are given in large doses. The reason for these poor results may be that lipase is inactivated by gastric acid, that food empties from the stomach faster than do the pancreatic enzymes, and that batches of commercially available pancreatic extracts vary in enzyme activity.

For the usual patient, two or three enteric-coated capsules or eight conventional (non-enteric-coated) tablets of a potent enzyme preparation should be administered with meals. Some patients using conventional tablets require adjuvant therapy to improve enzyme replacement treatment. H_2receptor antagonists, sodium bicarbonate, and proton pump inhibitors are effective adjuvants. Antacids containing calcium carbonate or magnesium hydroxide are not effective and may actually result in increased steatorrhea. Several publications have reported colonic strictures in patients with cystic fibrosis receiving extraordinarily high doses of high-potency pancreatic enzyme preparations. Such lesions have not been reported in adults with chronic pancreatitis.

Supportive measures include diet restriction and pain medications. The diet should be moderate in fat (30%), high in protein (24%), and low in carbohydrate (40%). Restriction of long-chain triglyceride intake can help patients who do not respond satisfactorily to pancreatic enzyme therapy. Use of foods containing mainly medium-chain fatty acids, which do not require lipase for digestion, may be beneficial. Nonnarcotic analgesics should be emphasized. Patients taking narcotic drugs for pain relief often become addicted and continue to have pain.

Patients with severe exocrine pancreatic insufficiency secondary to alcohol who continue to drink have a high mortality rate (in one series, 50% of patients who were followed for 5 to 12 years died during this period) and significant morbidity (weight loss, lassitude, vitamin deficiency, and narcotic addiction). Chronic pancreatitis carries significant medical and social costs. A recent study found that pancreatitis led to retirement in 11% of patients with the disease, accounting for 45% of all retirements. In 87% of patients with chronic pancreatitis unable to maintain gainful employment, alcoholism was a contributing factor. Patients with chronic pancreatitis also use substantial medical resources. In 1987 in the United States, this diagnosis accounted for 122,000 recorded outpatient visits and 56,000 hospital admissions. Pain may abate if progressive severe exocrine insufficiency continues. Patients who abstain from alcohol and use vigorous replacement therapy for maldigestion-malabsorption do reasonably well.

HEREDITARY PANCREATITIS

Hereditary pancreatitis is a rare disease that is similar to chronic pancreatitis except for an early age of onset and evidence of hereditary factors (involving an autosomal dominant gene with incomplete penetrance). A genome-wide search using genetic linkage analysis identified the hereditary pancreatitis gene on chromosome 7. An R117H mutation in the cationic trypsinogen gene occurs in most of the families with hereditary pancreatitis that have been studied. Molecular modeling predicts the formation of hydrolysis-resistant trypsin that could lead to pancreatic autodigestion. These patients have recurring attacks of severe abdominal pain which may last from a few days to a few weeks. The serum amylase and lipase levels may be elevated during acute attacks but usually are normal. Patients frequently develop pancreatic calcification, diabetes mellitus, and steatorrhea, and, in addition, they have an increased incidence of pancreatic carcinoma. Such patients often require ductal decompression for pain relief. Abdominal complaints in relatives of patients with hereditary pancreatitis should raise the question of pancreatic disease.

PANCREATIC ENDOCRINE TUMORS

**Pancreatic endocrine tumors are summarized in Table 304-6 and are discussed in Chap. 93.*

OTHER CONDITIONS

ANNULAR PANCREAS

When the ventral pancreatic anlage fails to migrate correctly to make contact with the dorsal anlage, the result may be a ring of pancreatic tissue encircling the duodenum. Such an annular pancreas may cause intestinal obstruction in the neonate or the adult. Symptoms of postprandial fullness, epigastric pain, nausea, and vomiting may be present for years before the diagnosis is entertained. The radiographic findings are symmetric dilation of the proximal duodenum with bulging of the recesses on either side of the annular band, effacement but not destruction of the duodenal mucosa, accentuation of the findings in the right anterior oblique position, and lack of change on repeated examinations. The differential diagnosis should include duodenal webs, tumors of the pancreas or duodenum, postbulbar peptic ulcer, regional enteritis, and adhesions. Patients with annular pancreas have an increased incidence of pancreatitis and peptic ulcer. Because of these and other potential complications, the treatment is surgical even if the condition has been present for years. Retrocolic duodenojejunostomy is the procedure of choice, although some surgeons advocate Billroth II gastrectomy, gastroenterostomy, and vagotomy.

PANCREAS DIVISUM

Pancreas divisum occurs when the embryologic ventral and dorsal pancreatic anlagen fail to fuse, so that pancreatic drainage is accomplished mainly through the accessory papilla. Pancreas divisum is the most common congenital anatomic variant of the human pancreas. Current evidence indicates that this anomaly does not predispose to the development of pancreatitis in the great majority of patients who harbor it. However, the combination of pancreas divisum and a small accessory orifice could result in dorsal duct obstruction. The challenge is to identify this subset of patients with dorsal duct pathology. Cannulation of the dorsal duct byERCP is not as easily done as is cannulation of the ventral duct. Patients with pancreatitis and pancreas divisum demonstrated by ERCP should be treated with conservative measures. In many of these patients, pancreatitis is idiopathic and unrelated to the pancreas divisum. Endoscopic or surgical intervention is indicated only when the above methods fail. If marked dilation of the dorsal duct can be demonstrated, surgical ductal decompression should be performed. The appropriate therapy for patients without dilation of the dorsal duct is not yet defined. It should be stressed that the ERCP appearance of pancreas divisum -- i.e., a small-caliber ventral duct with an arborizing pattern -- may be mistaken as representing an obstructed main pancreatic duct secondary to a mass lesion.

MACROAMYLASEMIA

In macroamylasemia, amylase circulates in the blood in a polymer form too large to be easily excreted by the kidney. Patients with this condition demonstrate an elevated serum amylase value, a low urinary amylase value, and a C_{am}/C_{cr}ratio of less than 1%. The presence of macroamylase can be documented by chromatography of the serum. The prevalence of macroamylasemia is 1.5% of the nonalcoholic general adult hospital population. Usually macroamylasemia is an incidental finding and is not related to disease of the pancreas or other organs.

Macrolipasemia has now been documented in a few patients with cirrhosis or non-Hodgkin's lymphoma. In these patients, the pancreas appeared normal on ultrasound andCTexamination. Lipase was shown to be complexed with immunoglobulin A. Thus, the possibility of *both* macroamylasemia and macrolipasemia should be considered in patients with elevated blood levels of these enzymes.

ACKNOWLEDGEMENT

This chapter represents a revised version of a chapter by Dr. Norton J. Greenberger, Dr. Phillip P. Toskes, and Dr. Kurt J. Isselbacher that was in the previous editions of this textbook.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART TWELVE -DISORDERS OF THE IMMUNE SYSTEM, CONNECTIVE TISSUE, AND JOINTS

SECTION 1 - DISORDERS OF THE IMMUNE SYSTEM

305. INTRODUCTION TO THE IMMUNE SYSTEM - *Barton F. Haynes*, *Anthony S. Fauci*

DEFINITIONS

·*Adaptive immune system* -- recently evolved system of immune responses mediated by T and B lymphocytes. Immune responses by these cells are based on specific antigen recognition by clonotypic receptors that are products of genes that rearrange during development and throughout the life of the organism. Additional cells of the adaptive immune system include various types of antigen-presenting cells.

· *Antibody* -- B cell-produced molecules encoded by genes that rearrange during B cell development consisting of immunoglobulin heavy and light chains that together form the central component of the B cell receptor for antigen. Antibody can exist as B cell surface antigen-recognition molecules or as secreted molecules in plasma and other body fluids.

· *Antigens* -- foreign or self molecules that are recognized by the adaptive and innate immune systems resulting in innate immune cell triggering, T cell activation, and/or B cell antibody production.

· *Antimicrobial peptides* -- small peptides<100 amine acids in length that are produced by cells of the innate immune system and have anti-infectious agent activity.

· *B lymphocytes* -- bone marrow-derived or bursal-equivalent lymphocytes that express surface immunoglobulin (the B cell receptor for antigen) and secrete specific antibody after interaction with antigen.

· *B cell receptor for antigen* -- complex of surface molecules that rearrange during postnatal B cell development, made up of surface immunoglobulin (Ig) and associated Ig ab chain molecules that recognize nominal antigen via Ig heavy and light chain variable regions, and signal the B cell to terminally differentiate to make antigen-specific antibody.

· *Complement* -- cascading series of plasma enzymes and effector proteins whose function is to lyse pathogens and/or target them to be phagocytized by neutrophils and monocyte/macrophage lineage cells of the reticuloendothelial system.

· *Co-stimulatory molecules* -- molecules of antigen-presenting cells (such as B7-1 and B7-2 or CD40) that lead to T cell activation when ligated by ligands on activated T cells (such as CD28 or CD40 ligand).

· *Cytokines* -- soluble proteins that interact with specific cellular receptors that are involved in the regulation of the growth and activation of immune cells and mediate normal and pathologic inflammatory and immune responses.

· *Dendritic cells* -- myeloid and/or lymphoid lineage antigen-presenting cells of the adaptive immune system. Immature dendritic cells, or dendritic cell precursors, are key components of the innate immune system by responding to infections with production of high levels of cytokines. Dendritic cells are key initiators both of innate immune responses via cytokine production and of adaptive immune responses via presentation of antigen to T lymphocytes.

· *Innate immune system* -- ancient immune recognition system of host cells bearing germline encoded pattern recognition receptors (PRRs) that recognize pathogens and trigger a variety of mechanisms of pathogen elimination. Cells of the innate immune system include natural killer (NK) cell lymphocytes, monocytes/macrophages, immature or dendritic cell precursors, neutrophils, basophils, eosinophils, tissue mast cells, and epithelial cells.

· *Large granular lymphocytes* -- lymphocytes of the innate immune system with azurophilic cytotoxic granules that haveNK cell activity capable of killing foreign and host cells with little or no self major histocompatibility complex (MHC) class I molecules.

· *Natural killer cells* -- large granular lymphocytes that kill target cells that express little or no HLA class I molecules, such as malignantly transformed cells and virally infected cells. NK cells express receptors that inhibit killer cell function when self MHC class I is present.

· *Pathogen-associated molecular patterns* -- Invariant molecular structures expressed by large groups of microorganisms that are recognized by host cellularPRRs in the mediation of innate immunity.

· *Pattern recognition receptors* -- germline-encoded receptors expressed by cells of the innate immune system that recognize pathogen-associated molecular patterns (PAMPs).

· *T cells* -- thymus-derived lymphocytes that mediate adaptive cellular immune responses including T helper and cytotoxic T lymphocyte effector cell functions.

· *T cell receptor for antigen* -- complex of surface molecules that rearrange during postnatal T cell development made up of clonotypic T cell receptor (TCR) a andb chains that are associated with the CD3 complex composed of invariant g, d,e, z, andh chains. The clonotypic TCRa and b chains recognize peptide fragments of protein antigen physically bound in antigen-presenting cell MHC class I or II molecules, leading to signaling via the CD3 complex to mediate effector functions.

· *Tolerance* -- recognition of foreign or self antigens by B and T lymphocytes in the absence of expression of antigen-presenting cell co-stimulatory molecules that leads to B and T cell nonresponsiveness to antigens. Active T lymphocyte tolerance can be achieved through blockade of the B7/CD28 co-stimulatory pathway.

INTRODUCTION

The human immune system has evolved over millions of years from both invertebrate and vertebrate organisms to develop sophisticated defense mechanisms highly specific for invading pathogens. Immune systems evolved to protect the host from microbes and their virulence factors. From invertebrates, humans have inherited the innate immune system, an ancient defense system that uses germ line-encoded proteins to recognize pathogens. Cells of the innate immune system, such as macrophages andNKlymphocytes, recognize pathogen molecular motifs that are highly conserved among many microbes (PAMPs) and use a diverse set of receptor molecules (PRRs). Important components of the recognition of microbes by the innate immune system are: (1) recognition by germ line-encoded host molecules, (2) recognition of key microbe virulence factors but not recognition of self molecules, and (3) nonrecognition of benign foreign molecules or microbes. It is particularly important for the innate immune system to not recognize foreign nonpathogenic molecules that are common in the environment, since reaction against them would cause continuous inflammatory disease. Upon contact with pathogens, macrophages and NK cells may kill pathogens directly or may activate a series of events that both slows the infection and recruits the more recently evolved arm of the human immune system, the adaptive immune system.

Adaptive immunity is found only in vertebrates and is based on the generation of antigen receptor T and B lymphocytes by germ-line gene rearrangements that occur during the development of each person. By a complex series of molecular mechanisms of gene rearrangement, individual T or B cells express unique antigen receptors on their surface, such that taken together the pools of adult human T and B lymphocytes contain cells capable of specifically recognizing the diverse antigens of the myriad of infectious agents in the environment. Coupled with finely tuned specific recognition mechanisms that maintain tolerance to self antigen, T and B lymphocytes of the adaptive immune system with their postnatally rearranged clonotypic antigen receptors bring both *specificity* and *immune memory* to vertebrate host defenses.

This chapter describes the cellular components, molecules, and mechanisms that make up the innate and adaptive immune systems and describes how adaptive immunity is recruited to the defense of the host by innate immune responses. An appreciation of the cellular and molecular bases of innate and adaptive immune responses is critical to understanding the pathogenesis of inflammatory, autoimmune, infectious, and immunodeficiency diseases.

THE CD CLASSIFICATION OF HUMAN LYMPHOCYTE DIFFERENTIATION ANTIGENS

The development of monoclonal antibody technology led to the discovery of a large number of new leukocyte surface molecules. In 1982, the First International Workshop on Leukocyte Differentiation Antigens was held to establish a nomenclature for cell-surface molecules of human leukocytes. From this and subsequent leukocyte differentiation workshops has come the cluster of differentiation (CD) classification of leukocyte antigens (Table 305-1). The data presented inTable 305-1 establish a context to facilitate discussion and study of the complex series of events that transpire during normal and aberrant innate and adaptive human immune responses.

THE INNATE IMMUNE SYSTEM

All multicellular organisms, including humans, have developed the use of a limited number of germ line-encoded molecules that recognize large groups of pathogens. Because of the myriad human pathogens, host molecules of the human innate immune system must recognizePAMPs, the common molecular structures shared by many pathogens. PAMPs must be conserved structures vital to pathogen virulence and survival, such as bacterial endotoxin, so that pathogens cannot mutate molecules of PAMPs to evade human innate immune responses. In addition, one major end product of innate immunity is the destruction of the invading pathogen, thus necessitating that PAMPs recognized by innate immune responses be completely distinct from self molecules.PPRs are host proteins of the innate immune system that recognize PAMPs and are human molecules whose ancestors are evolutionarily ancient (Tables 305-2,305-3). Thus, recognition of pathogen molecules by hematopoietic and nonhematopoietic cell types leads to activation/production of the complement cascade, cytokines, and antimicrobial peptides as effector molecules.

PATTERN RECOGNITION

MajorPRRfamilies of proteins include C-type lectins, leucine-rich proteins, macrophage scavenger receptor proteins, plasma pentraxins, lipid transferase, and integrins (Table 305-3). A major group of PRR collagenous glycoproteins with C-type lectin domains are termed *collectins* and include the serum protein, mannose-binding lectin. Mannose-binding lectin and other collectins, as well as two other protein families -- the pentraxins (such as C-reactive protein and serum amyloid P) and macrophage scavenger receptors -- all have the property of opsonizing (coating) bacteria for phagocytosis by macrophages and can also activate the complement cascade to lyse bacteria. Integrins are cell-surface adhesion molecules that signal cells after cells bind bacterial lipopolysacchride (LPS) and activate phagocytic cells to ingest pathogens.

A remarkable series of recent discoveries has revealed the mechanisms of connection between the innate and adaptive immune systems; these include (1) the plasma protein,LPS-binding protein, which binds and transfers LPS to the macrophage LPS receptor, CD14; and (2) a human family of proteins called *Toll proteins*, which are associated with CD14, bind LPS, and signal the macrophage to produce cytokines and upregulate cell-surface molecules that signal the initiation of adaptive immune responses (Fig. 305-1,Table 305-3, andTable 305-4). Proteins in the Toll family are expressed on macrophages (Toll 2 and Toll 4) and on dendritic cells and B cells (RP105). Upon ligation these receptors activate a series of intracellular events that lead to the killing of bacteria as well as to the recruitment and ultimate activation of antigen-specific T and B lymphocytes (Fig. 305-1). Importantly, signaling by massive amounts of LPS through Toll receptors leads to the release of large amounts of cytokines that mediate LPS-induced shock. Mutations in Toll proteins in mice protect from LPS shock, and mutations in Toll proteins in humans similarly protect from LPS-induced inflammatory diseases such as LPS-induced asthma.

Cells of invertebrates and vertebrates produce antimicrobial small peptides containing fewer than 100 amino acids that can act as endogenous antibodies (Table 305-2). Some of these peptides are produced by epithelia that line various organs, while others are found in macrophages or neutrophils that ingest pathogens. Antimicrobial peptides have been identified that kill bacteria such as *Pseudomonas* spp., *Escherichia coli*, and *Mycobacterium tuberculosis*.

EFFECTOR CELLS OF INNATE IMMUNITY

Cells of the innate immune system and their roles in the first line of host defense are described in Table 305-4. Equally important as their roles in the mediation of innate immune responses are the roles that each cell type plays in recruiting T and B lymphocytes of the adaptive immune system to engage in specific antipathogen responses.

Monocytes-Macrophages Monocytes arise from precursor cells within bone marrow (Fig. 305-2) and circulate with a half-life ranging from 1 to 3 days. Monocytes leave the peripheral circulation by marginating in capillaries and migrating into a vast extravascular pool. Tissue macrophages arise from monocytes that have migrated out of the circulation and by in situ proliferation of macrophage precursors in tissue. Common locations where tissue macrophages (and certain of their specialized forms) are found are lymph node, spleen, bone marrow, perivascular connective tissue, serous cavities such as the peritoneum, pleura, skin connective tissue, lung (alveolar macrophage), liver (Kupffer cell), bone (osteoclast), central nervous system (microglia), and synovium (type A lining cell).

In general, monocytes-macrophages are on the first line of defense associated with innate immunity; however, they also play a major role in recruitment of adaptive immune responses by mediation of functions such as bindingLPS, the presentation of antigen to T lymphocytes, and the secretion of factors such as interleukin (IL) 1, tumor necrosis factor (TNF), IL-12, and IL-6, which are central to antigen-specific activation of T and B lymphocytes (Fig. 305-1). Although monocytes-macrophages were originally thought to be the major antigen-presenting cells (APCs) of the immune system, it is now clear that dendritic/Langerhans cells are the most potent and effective APCs in the body (see below). Monocytes-macrophages mediate innate immune effector functions such as destruction of antibody-coated bacteria, tumor cells, or even normal hematopoietic cells in certain types of autoimmune cytopenias. Activated macrophages can also mediate antigen-nonspecific lytic activity and eliminate cell types such as tumor cells in the absence of antibody. This activity is largely mediated by cytokines (i.e., TNF-a and IL-1). Monocytes-macrophages express lineage-specific molecules (e.g., the cell-surface LPS receptor, CD14) as well as surface receptors for a number of molecules, including the Fc region of IgG (CD16, CD32, CD64), activated complement components (CD35) (Table 305-1), and various cytokines (Table 305-5). Finally, macrophage secretory products are more diverse than those of any other cell of the immune system. Among monocyte-macrophage-secreted products are hydrolytic enzymes, products of oxidative metabolism, TNF-a, IL-1, -6, -10, -12, -15, -18, and a number of chemoattractant cytokines (chemokines) involved in the orchestration of an immune response in tissues (Table 305-5).

Dendritic/Langerhans Cells Dendritic/Langerhans cells are bone marrow-derivedAPCsthat are distinct from monocytes-macrophages and are derived from both lymphoid and myeloid lineages. They generally lack the standard T, B,NK, and monocyte cell markers but do express CD83 and other molecules that aid in their identification. They can be expanded in culture, and their function is enhanced by the cytokines granulocyte-macrophage colony stimulating factor (GM-CSF),IL-1, IL-4, andTNF-a. They are distinguished by an exceptional ability to present antigen, by expression of high levels ofMHCclass II and co-stimulatory molecules, and by dendritic morphology with multiple thin membrane projections (veils).

Dendritic cells are referred to as Langerhans cells when they are present in the skin and beneath the mucosal surface. They comprise the dendritic cells of the blood and the spleen and the veil cells of afferent lymphatics, and they form part of the interdigitating cell network of lymphoid organs. In responses involving the innate immune system, bacterialLPSbinds to dendritic cell RP105 Toll-like protein, upregulating dendritic cell molecules, such asMHCclass II, B7-1 (CD80), and B7-2 (CD86), which enhance specific antigen presentation and induce dendritic cell cytokine production.

A critical cell type of the innate immune system is the dendritic cell precursor that, in response to viral infections, produces high levels of interferon (IFN)a. IFN-a in turn activatesNK cells to kill virally infected cells and activates monocytes-macrophages and otherAPCs to recruit antigen-specific T and B cells to respond to viral infections. thus, immature dendritic cells are important components of innate immunity, while mature dendritic cells, as APCs, are important components of adaptive immunity.

Follicular Dendritic Cells Follicular dendritic cells (FDCs) areAPCs for B cells and their lineage is distinct from that of dendritic/Langerhans cells, the major APCs for T cells. FDCs are located in the germinal centers of follicles of secondary lymphoid organs. Their main function is to trap and retain antigens in the germinal centers of lymphoid organs and to present these antigens to B cells. Antigen is retained on their membranes in the form of antigen-antibody complexes that bind to the cell via the cellular receptor for C3. FDCs have extensive, thin, finger-like projections that surround the B cells in the germinal centers, allowing for maximal exposure of trapped antigen. The retention of antigen on the surface of FDC membranes is critical for the selection and growth of high-affinity clones of B cells and for the maintenance of B cell memory. Of note, HIV is trapped in large quantities on the processes of FDCs in lymphoid organs, allowing the lymphoid tissue to serve as a reservoir of virus and a source of infection for CD4+ T cells migrating into the area to provide help to B cells in the initiation and propagation of an HIV-specific humoral response (Chap. 309).

Large Granular Lymphocytes/Natural Killer Cells Large granular lymphocytes (LGLs) account for approximately 5 to 10% of peripheral blood lymphocytes. LGLs are nonadherent, nonphagocytic cells with large azurophilic cytoplasmic granules. LGLs express surface receptors for the Fc portion of IgG (CD16) and for NCAM-I (CD56), and many LGLs express some T lineage markers, particularly CD8, and proliferate in response toIL-2. LGLs arise in both bone marrow and thymic microenvironments (Fig. 305-2).

Functionally,LGLsshare features with both monocytes-macrophages and neutrophils in that LGLs mediate both antibody-dependent cellular cytotoxicity (ADCC) andNKactivity. ADCC is the binding of an opsonized (antibody-coated) target cell to an Fc receptor-bearing effector cell via the Fc region of antibody, resulting in lysis of the target by the effector cell. NK cell activity is the nonimmune (i.e., effector cell never having had previous contact with the target),MHC-unrestricted, non-antibody-mediated killing of target cells, which are usually malignant cell types, transplanted foreign cells, or virus-infected cells. Thus, LGLs that mediate NK cell activity may play an important role in immune surveillance and destruction of cells that spontaneously undergo malignant transformation in vivo. Subsets of NK cells may play a role in hematopoietic cell engraftment; some subsets stimulate bone marrow stem cells, and others stimulate engraftment. Lymphokine-activated killer (LAK) cells are NK lymphocytes that proliferate in vitro to high concentrations ofIL-2 and develop the ability to kill tumor cells more efficiently than unstimulated NK cells. Rare patients with complete absence of NK cells have been described who lack both NK cell activity and CD56+, CD16+ lymphocytes but have normal T and B cell function. NK cell hyporesponsiveness is also observed in patients with the *Chediak-Higashi syndrome*, an autosomal recessive disease associated with fusion of cytoplasmic granules and defective degranulation of neutrophil lysosomes.

The ability ofNK cells to kill target cells is inversely related to target cell expression ofMHCclass I molecules. Thus, NK cells kill target cells with low or no levels of MHC class I expression and are prevented from killing target cells with high levels of class I expression. Recent studies have demonstrated the presence of NK receptors (NK-Rs) or killer cell inhibitory receptors (KIRs) that bind to either classic MHC class I molecules in a polymorphic way or the MHC-class Ib molecule HLA-E (Fig. 305-3). In every person, NK cells express at least one NK-R that recognizes a self-MHC class I allele. NK-Rs of the Ig superfamily bind specific MHC class I molecules; for example, the NK-R p140 binds HLA-A3, and another NK-R, p70, binds HLA-B27 (Fig. 305-3). A second NK-R of the C-type lectin family of proteins is termed *CD94/NKG2A* and binds the MHC-related protein HLA-E (Fig. 305-3). HLA-E has an MHC class I structure but exclusively binds the leader sequence peptides of classic MHC class I molecules in the HLA-E MHC-like "notch" (see "Molecular Basis of T Cell Recognition of Antigen," below). In this manner, CD94/NKG2A NK cell molecules survey and monitor the total level of classic MHC class I molecules on the surface of host cells. When cell-surface levels of host MHC class I molecules decrease, such as occurs during malignant transformation or viral infection of host cells, the altered host cell with diminished MHC class I expression is recognized by NK-Rs, and the NK cell is activated to kill the host tumor or virally infected cells. The ability of NK-Rs to bind to self-MHC and inhibit NK killing of normal host cells is a key protective mechanism for prevention of NK cell-mediated autoimmune disease.

SomeNK cells express CD3 and are termed *NK/T cells*. NK/T cells can also express oligoclonal forms of theTCR for antigen that can recognize lipid molecules of intracellular bacteria when presented in the context of CD1 molecules onAPCs. This mode of recognition of intracellular bacteria such as *Listeria monocytogenese* and *M. tuberculosis* by NK/T cells is thought to be an important defense mechanism against these organisms that, via usage of a clonal form of TCRs for antigen, incorporates components of both the innate and adaptive immune systems.

Neutrophils, Eosinophils, and Basophils Granulocytes are present in nearly all forms of inflammation and are amplifiers and effectors of innate immune responses. Unchecked accumulation and activation of granulocytes can lead to host tissue damage, as seen in neutrophil- and eosinophil-mediated *systemic necrotizing vasculitis*. Granulocytes are derived from stem cells in bone marrow (Fig. 305-2). Each type of granulocyte (neutrophil, eosinophil, or basophil) is derived from a different subclass of progenitor cell, which is stimulated to proliferate by colony stimulating factors (Table 305-5). During terminal maturation of granulocytes, class-specific nuclear morphology and cytoplasmic granules appear that allow for histologic identification of granulocyte type.

Neutrophils express Fc receptors for IgG (CD16) and receptors for activated complement components (C3b or CD35) (Table 305-1). Upon interaction of neutrophils with opsonized bacteria or immune complexes, azurophilic granules (containing myeloperoxidase, lysozyme, elastase, and other enzymes) and specific granules (containing lactoferrin, lysozyme, collagenase, and other enzymes) are released, and microbicidal superoxide radicals (O_{2}-) are generated at the neutrophil surface. The generation of superoxide leads to inflammation by direct injury to tissue and by alteration of macromolecules such as collagen and DNA.

Eosinophils express Fc receptors for IgG (CD32) and are potent cytotoxic effector cells for various parasitic organisms. Intracytoplasmic contents of eosinophils, such as major basic protein, eosinophil cationic protein, and eosinophil-derived neurotoxin, are capable of directly damaging tissues and may be responsible in part for the organ system dysfunction in the *hypereosinophilic syndromes* (Chap. 64). Since the eosinophil granule contains anti-inflammatory types of enzymes (histaminase, arylsulfatase, phospholipase D), eosinophils may homeostatically downregulate or terminate ongoing inflammatory responses.

The normal functions of basophils and tissue mast cells are not completely understood; they are potent reservoirs of cytokines such asIL-4. The capacity of basophil cytokines and mediators to increase local delivery of antibodies and complement by increasing vascular permeability is hypothetical. Thus, the basophil is identified principally with allergic reactions and some delayed cutaneous hypersensitivity states. Certainly, the promotion of increased vascular permeability by basophils is important in the genesis of inflammatory lesions in some vasculitis syndromes (Chap. 317). Basophils express high-affinity surface receptors for IgE (FcRI) and, upon cross-linking of basophil-bound IgE by antigen, release histamine, eosinophil chemotactic factor of anaphylaxis, and neutral protease -- all mediators of immediate (anaphylaxis) hypersensitivity responses (Table 305-6). In addition, basophils express surface receptors for activated complement components (C3a, C5a), through which mediator release can be directly effected.**For further discussion of tissue mast cells, see Chap. 310.*

THE COMPLEMENT SYSTEM

The complement system, an important soluble component of the innate immune system, is a series of plasma enzymes, regulatory proteins, and proteins that are activated in a cascading fashion, resulting in cell lysis. There are two arms of the complement system (Fig. 305-4). Activation of the classic complement pathway via C1, C4, and C2 and activation of the alternative complement pathway via factor D, C3, and factor B both lead to cleavage and activation of C3. C3 is a protein whose activation fragments, when bound to target surfaces such as bacteria and other foreign antigens, are critical for opsonization (coating by antibody and complement) in preparation for phagocytosis.

The protein fragment C3b, split from C3, is necessary for activation of the terminal complement components C5 through C9. These form the membrane attack complex, which, when inserted into cell membranes, brings about osmotic lysis of the cell.

C3b also joins with a cleavage product of factor B (called Bb) to form C3bBb, also known as the *alternative pathway C3 convertase*. Activation of the classic complement pathway results in cleavage of C4 and C2 with a resulting complex of fragments, C4b2a, also called the *classic pathway C3 convertase*. Both the classic pathway C3 convertase (C4b2a) and the alternative pathway C3 convertase (C3bBb) function to cleave C3 to form active C3b, thus driving activation of the C5-9 membrane attack complex. The fact that C3b can combine with Bb to form the alternative pathway C3 convertase gives rise to a potent positive-feedback loop for production of C3b and thus continued activation of terminal complement components.

The classic complement pathway is activated by interaction of antigen and antibody to form immune complexes that bind C1q, a subunit of C1. Immunoglobulin isotypes that bind C1q and activate the classic pathway are IgM, IgG1, IgG2, and IgG3. In contrast, IgA1, IgA2, and IgD activate complement via the alternative pathway. Activation of the complement cascade via the classic pathway by IgG- or IgM-containing immune complexes is a rapid and efficient pathway to activation of terminal complement components. In contrast, activation of the alternative complement pathway via IgA-containing immune complexes or by bacterial endotoxin is a slower and less efficient pathway to terminal component activation. Thus the immunoglobulin isotype composition of immune complexes is a critical factor in determining complement activation and the efficiency of clearance of immune complexes by C3 receptor-bearing cells.

In addition to the role of complement in opsonization of bacteria and cell lysis, several complement fragments are potent mediators of immune cell activation. C3a and C5a bind to receptors on mast cells and basophils, resulting in release of histamine and other mediators of anaphylaxis. C5a is also a potent chemoattractant for neutrophils and monocytes-macrophages (Table 305-7).

CYTOKINES

Cytokines are soluble proteins produced by a wide variety of hematopoietic and nonhematopoietic cell types (Table 305-5). They are critical for both normal innate and adaptive immune responses, and their expression may be perturbed in most immune, inflammatory, and infectious disease states.

Cytokines are involved in the regulation of the growth, development, and activation of immune system cells and in the mediation of the inflammatory response. In general, cytokines are characterized by considerable redundancy in that different cytokines have similar functions. In addition, many cytokines are pleiotropic in that they are capable of acting on many different cell types. This pleiotropism results from the expression on multiple cell types of receptors for the same cytokine (see below), leading to the formation of "cytokine networks." The action of cytokines may be: (1) autocrine when the target cell is the same cell that secretes the cytokine, (2) paracrine when the target cell is nearby, and (3) endocrine when the cytokine is secreted into the circulation and acts distal to the source. A number of classifications have been proposed for the grouping of cytokines according to functions; however, these are all imperfect because of the fact that a number of cytokines overlap these groupings. One empirical classification divides the cytokines into the following three groups:

1. Immunoregulatory cytokines involved in the activation, growth, and differentiation of lymphocytes and monocytes, e.g.,IL-2, IL-4, IL-10,IFN-g, and transforming growth factor (TGF) b

2. Proinflammatory cytokines produced predominantly by mononuclear phagocytes in response to infectious agents (e.g.,IL-1,TNF-a, and IL-6) and the chemokine family of inflammatory cytokines, within which are included IL-8, monocyte chemotactic protein (MCP)-1, MCP-2, MCP-3, macrophage inflammatory protein (MIP)-1a, MIP-1b, and regulation-upon-activation, normal T expressed and secreted (RANTES) (Chap. 64)

3. Cytokines that regulate immature leukocyte growth and differentiation, e.g.,IL-3, IL-7, andGM-CSF.

In general, cytokines exert their effects by influencing gene activation that results in cellular activation, growth, differentiation, functional cell-surface molecule expression, and cellular effector function. In this regard, cytokines can have dramatic effects on the regulation of immune responses and the pathogenesis of a variety of diseases. Indeed, T cells have been categorized on the basis of the pattern of cytokines that they secrete that results in either humoral immune response (T_H2) or a cell-mediated immune response (T_H1).

Cytokine receptors can be grouped into five general families based on similarities in their extracellular amino acid sequences and conserved structural domains (Fig. 305-5). The *immunoglobulin (Ig) superfamily* represents a large number of cell-surface and secreted proteins. All members of the Ig superfamily must have at least one common domain in their protein structure. TheIL-1 receptors (type 1, type 2) are examples of cytokine receptors with extracellular Ig domains.

The hallmark of the *hematopoietic growth factor (type 1) receptor* family is that the extracellular regions of each receptor contain two conserved motifs. One motif located at the N terminus is rich in cysteine residues. The other motif is located at the C terminus proximal to the transmembrane region and comprises five amino acid residues, tryptophan-serine-X-tryptophan-serine (WSXWS). Cytokine receptors expressing the WSXWS motif are also referred to as "type I family of cytokine receptors." This family can be further grouped on the basis of the number of receptor subunits they have and on the utilization of shared subunits. The shared common receptors often have a critical role in signal transduction. A number of cytokine receptors, i.e.,IL-6, IL-11, IL-12, and leukemia inhibitory factor, are paired with gp130. There is also a common 150-kDa subunit shared by IL-3, IL-5, andGM-CSFreceptors. The gamma chain (g_c) of the IL-2 receptor is common to the IL-2, IL-4, IL-7, IL-9, and IL-15 receptors. Thus, the specific cytokine receptor is responsible for ligand-specific binding, while the subunits such as gp130, the 150-kDa subunit, andg_c are important in signal transduction. Theg_c gene is on the X chromosome, and mutations in the$g_c$protein result in the X-linked form of severe combined immune deficiency syndrome (X-SCID) (Chap. 308).

The members of the *interferon (type II) receptor* family include the receptors forIFN-g, and -b, which share a similar 210-amino-acid binding domain with conserved cysteine pairs at both the amino and carboxy termini. The receptors for the interferons consist of at least two distinct subunits.

The members of the *TNF (type III) receptor family* share a common binding domain composed of repeated cysteine-rich regions. Members of this family include the p55 and p75 receptors for TNF(TNFR1 and TNFR2, respectively); CD40 antigen, which is an important B cell-surface marker involved in immunoglobulin isotype switching; fas/Apo-1, whose triggering induces apoptosis (programmed cell death); CD27 and CD30, which are found on activated T cells and B cells; and nerve growth factor receptor.

The common motif for the *seven transmembrane helix family* was originally found in receptors linked to GTP-binding proteins. This family includes receptors for chemokines,b-adrenergic receptors, and retinal rhodopsin. It is important to note that two members of the chemokine receptor family, CXC chemokine receptor type 4 (CXCR4) and b chemokine receptor type 5 (CCR5), have recently been found to serve as the two major coreceptors for binding and entry of HIV into CD4-expressing host cells (Chap. 309). Both cytokines and their receptors share similar structures and functions. For example, ligands for theTNFreceptor family of receptors regulate and determine activation for programmed cell death (*apoptosis*) and all ligate molecules of the same structural family. SimilarlyIL-3, IL-5, andGM-CSFare all produced by T helper (T_H) 2 cells, and the receptors of these cytokines share commong chains. Thus, cytokines and their receptors may have diversified together during evolution.

Significant advances have been made in defining the signaling pathways through which cytokines exert their effects intracellularly. This is particularly true with regard to the diverse family of hematopoietin receptors. The Janus family of protein tyrosine kinases (JAK) is a critical element involved in signaling via the hematopoietin receptors. There are four known JAK kinases, JAK1, JAK2, JAK3, and Tyk2, which preferentially bind different receptor subunits. Cytokine binding to its receptor brings the cytokine receptor subunits into apposition and allows a pair of JAKs to transphosphorylate and activate one another. The JAKs then phosphorylate the receptor on the tyrosine residues and allow signaling molecules to bind to the receptor, where these molecules in turn can become phosphorylated. These signaling molecules can bind the receptor because they have domains (SH2, or src homology 2 domains) that can bind phosphorylated tyrosine residues. There are a number of these important signaling molecules that bind the receptor, such as the adapter molecule SHC, which can couple the receptor to the activation of the mitogen-activated protein kinase pathway. In addition, a very important class of substrate of the JAKs is the signal transducers and activators of transcription (STAT) family of transcription factors. STATs have SH2 domains that enable them to bind to phosphorylated receptors, where they are then phosphorylated by the JAKs. It appears that different STATs have specificity for different receptor subunits. The STATs then dissociate from the receptor and translocate to the nucleus, bind to DNA motifs that they recognize, and regulate gene expression. The STATs preferentially bind DNA motifs that are slightly different from one another and thereby presumably control transcription of specific genes. The importance of this pathway is particularly relevant to lymphoid development. Mutations of JAK3 itself also result in a disorder identical toX-SCID; however, since JAK3 is found on chromosome 19 and not on the X chromosome, JAK3 deficiency occurs in boys and girls (Chap. 308). In this chapter the cytokines that affect various cell types are discussed in the context of each of the cell types.

THE ADAPTIVE IMMUNE SYSTEM

Adaptive immunity is characterized by antigen-specific responses to a foreign antigen or pathogen and, compared to innate immunity which occurs immediately (1 to 2 days), generally takes several days or longer to materialize. A key feature of adaptive immunity is memory for the antigen such that subsequent antigen exposures lead to more rapid and often more vigorous immune responses. The adaptive immune system consists of dual limbs of cellular and humoral immunity. The principal effectors of cellular immunity are T lymphocytes, while the principal effectors of humoral immunity are B lymphocytes (Table 305-8). Both B and T lymphocytes derive from a common stem cell (Fig. 305-2).

The proportion and distribution of immunocompetent cells in various tissues reflect cell traffic, homing patterns, and functional capabilities. Bone marrow is the major site of maturation of B cells, monocytes-macrophages, and granulocytes and contains pluripotent stem cells which, under the influence of various colony stimulating factors, are capable of giving rise to all hematopoietic cell types (Fig. 305-2). T cell precursors also arise from hematopoietic stem cells and home to the thymus for maturation. Mature T lymphocytes, B lymphocytes, monocytes, and dendritic/Langerhans cells enter the circulation and home to peripheral lymphoid organs (lymph nodes, spleen) and the gut-associated lymphoid tissue (tonsil, Peyer's patches, and appendix) as well as the skin and mucous membranes and await activation by foreign antigen.

T CELLS

The pool of effector T cells is established in the thymus early in life and is maintained throughout life both by new T cell production in the thymus and by antigen-driven expansion of virgin peripheral T cells into "memory" T cells that reside in peripheral lymphoid organs. The thymus exports approximately 2% of the total number of thymocytes per day throughout life, with the total number of daily thymic emigrants decreasing by approximately 3% per year during the first four decades of life. Thymic emigrants can be identified by the expression of certain combinations of T cell surface markers and by the presence in nuclei of excised (deleted) pieces of rearrangedTCR DNA, called *T cell receptor excision circles*.

Mature T lymphocytes constitute 70 to 80% of normal peripheral blood lymphocytes (only 2% of the total-body lymphocytes are contained in peripheral blood), 90% of thoracic duct lymphocytes, 30 to 40% of lymph node cells, and 20 to 30% of spleen lymphoid cells. In lymph nodes, T cells occupy deep paracortical areas around B cell germinal centers, and in the spleen, they are located in periarteriolar areas of white pulp (Chap. 63). T cells are the primary effectors of cell-mediated immunity, with subsets of T cells maturing into CD8+ cytotoxic T cells capable of lysis of virus-infected or foreign cells. In general, CD4+ T cells are also the primary regulatory cells of T and B lymphocyte and monocyte function by the production of cytokines and by direct cell contact. In addition, T cells regulate erythroid cell maturation in bone marrow, and through cell contact (CD40 ligand) have an important role in activation of B cells and induction of Ig isotype switching.

Human T cells express cell-surface proteins that mark stages of intrathymic T cell maturation or identify specific functional subpopulations of mature T cells. Many of these molecules mediate or participate in important T cell functions (Table 305-1;Fig. 305-6).

A number of cytokines regulate the process of T cell proliferation and differentiation (Table 305-5). The earliest identifiable T cell precursors in bone marrow are CD34+ pro-T cells (i.e., cells in whichTCRgenes are neither rearranged nor expressed). In the thymus, CD34+ T cell precursors begin cytoplasmic (c) synthesis of components of the CD3 complex of TCR-associated molecules (Fig. 305-6.) Within T cell precursors, TCR for antigen gene rearrangement begins under the influence ofIL-7 and yields two T cell lineages, expressing either TCRabchains or TCRgd chains. T cells expressing the TCRab chains comprise the majority of peripheral T cells in blood, lymph node, and spleen and terminally differentiate into either CD4+ or CD8+ cells. Cells expressing TCRgd chains circulate as a minor population in blood; their functions, although not fully understood, have been postulated to be those of immune surveillance at epithelial surfaces and cellular defenses against mycobacterial organisms and other intracellular bacteria (see below). Immature cortical thymocytes express, in addition to CD1, both CD4 and CD8 (i.e., they are double positive); however, upon reaching functional maturity, T cell expression of CD1 ceases, and CD4 and CD8 are reciprocally expressed. (i.e., T cells become single positive for either CD4 or CD8).

In the thymus, the recognition of self-peptides on thymic epithelial cells, thymic macrophages, and dendritic cells plays an important role in shaping the T cell repertoire to recognize foreign antigen (*positive selection*) and in eliminating highly autoreactive T cells (*negative selection*). As immature cortical thymocytes begin to express surfaceTCR for antigen, autoreactive thymocytes are destroyed (negative selection), thymocytes with TCRs capable of interacting with foreign antigen peptides in the context of self\en\MHCantigens are activated and develop to maturity (positive selection), and thymocytes with TCR that are incapable of binding to self-MHC antigens die of attrition (*no selection*). Mature thymocytes that are positively selected are either CD4+ helper T cells or MHC class II-restricted cytotoxic (killer) T cells, or they are CD8+ T cells destined to become MHC class I-restricted cytotoxic T cells. For T cells to be *MHC class I-* or *class II-restricted* means that T cells recognize antigen peptide fragments only when they are presented in the antigen-recognition site of a class I or class II MHC molecule, respectively (see below).

After thymocyte maturation and selection, mature CD4 and CD8 thymocytes leave the thymus and migrate to all sites of the peripheral immune system. It is important to note that the adult thymus continues to function, albeit with decreasing output, well into adult life, with detectable function though age 50. Thus, the thymus continues to be a contributor to the peripheral immune system, both normally and when the peripheral T cell pool is damaged, such as occurs in AIDS and cancer chemotherapy.

MOLECULAR BASIS OF T CELL RECOGNITION OF ANTIGEN

TheTCR for antigen is a complex of molecules consisting of an antigen-binding heterodimer of either ab or gd chains noncovalently linked with five CD3 subunits (g,d, e,z, and h) (Fig. 305-7). The CD3 z chains are either disulfide-linked homodimers (CD3-z_2) or disulfide-linked heterodimers composed of one zchain and one h chain. TCRabor TCRgd molecules must be associated with CD3 molecules to be inserted into the T cell surface membrane, TCRa being paired with TCRband TCRg being paired with TCRd. Molecules of the CD3 complex mediate transduction of T cell activation signals via TCRs, while TCRa and b org and d molecules combine to form the TCR antigen-binding site.

Thea, b,g, and dTCR for antigen molecules have amino acid sequence homology and structural similarities to immunoglobulin heavy and light chains and are thus members, along with other important molecules of immune cells of the *immunoglobulin gene superfamily* of molecules (e.g., MHC class I or II, CD3, CD4, CD8). The genes encoding the TCR molecules are encoded as clusters of gene segments that rearrange during the course of T cell maturation. This creates an efficient and compact mechanism for housing the diversity requirements of antigen receptor molecules. The TCRa chain is on chromosome 14 and consists of a series of V (variable), J (joining), and C (constant) regions. The TCRb chain is on chromosome 7 and consists of multiple V, D (diversity), J, and C TCRb loci. The TCRg chain is on chromosome 7, and the TCRd chain is in the middle of the TCRa locus on chromosome 14. Thus, molecules of the TCR for antigen have constant (framework) and variable regions, and the gene segments encoding the a, b,g, and d chains of these molecules are recombined and selected in the thymus, culminating in synthesis of the completed molecule. In both T and B cell precursors (see below), DNA rearrangements of antigen receptor genes involves the same enzymes, recombinase activating gene (RAG)1 and RAG2, both DNA-dependent protein kinases.

TCRdiversity is created by the different V, (D), and J segments that are possible for each receptor chain by the many permutations of V, D, and J segment combinations, by "N-region diversification" due to the addition of nucleotides at the junction of rearranged gene segments, and the pairing of individual chains to form a TCR dimer. As T cells mature in the thymus, the repertoire of antigen-reactive T cells is modified by selection processes that eliminate many autoreactive T cells, enhance the proliferation of cells that function appropriately with self-MHCmolecules and antigen, and allow T cells with nonproductive TCR rearrangements to die. Like B cell antigen receptors (Ig molecules), TCRs may also undergo affinity maturation by somatic mutation of the receptor, once they leave the thymus.

T cells do not recognize native protein or carbohydrate antigens. Instead, T cells recognize only short (approximately 9 to 13 amino acids) peptide fragments derived from protein antigens taken up or produced inAPCs. Foreign antigens may be taken up by endocytosis into acidified intracellar vesicles and degraded into small peptides that associate withMHCclass II molecules (exogenous antigen-presentation pathway). Other foreign antigens arise endogenously in the cytosol (such as from replicating viruses) and are broken down into small peptides that associate with MHC class I molecules (endogenous antigen-presenting pathway). Thus, APCs proteolytically degrade foreign proteins and display peptide fragments embedded in the MHC class I or II antigen-recognition site on the MHC molecule surface, where foreign peptide fragments are available to bind toTCRab or TCRgd chains of reactive T cells (Fig. 305-8). CD4 molecules act as an adhesive and, by direct binding to MHC class II (DR, DQ, or DP) molecules, stabilize the interaction of TCR with peptide antigen (Fig. 305-7). Similarly, CD8 molecules also act as adhesives to stabilize the TCR-antigen interaction by direct CD8 molecule binding to MHC class I (A, B, or C) molecules.

Antigens that arise in the cytosol and are processed via the endogenous antigen-presentation pathway are cleaved into small peptides by a 28-subunit complex of proteases called the *proteasome*. From the proteasome, antigen peptide fragments are transported from the cytosol into the lumen of the endoplasmic reticulum by a heterodimeric complex termed *transporters associated with antigen processing*, or TAP proteins. There,MHCclass I molecules in the endoplasmic reticulum membrane physically associate with processed cytosolic peptides. Following peptide association with class I molecules, peptide-class I complexes are exported to the Golgi apparatus, and then to the cell surface, for recognition by CD8+ T cells.

Antigens taken up from the extracellular space via endocytosis into intracellular acidified vesicles are degraded by vesicle proteases into peptide fragments. Intracellular vesicles containingMHCclass II molecules fuse with peptide-containing vesicles, thus allowing peptide fragments to physically bind to MHC class II molecules. Peptide-MHC class II complexes are then transported to the cell surface for recognition by CD4+ T cells.

Whereas it is generally agreed that theTCRabreceptor recognizes peptide antigens in the context ofMHCclass I or class II molecules, recent data suggest that lipids in the cell wall of intracellular bacteria such as *M. tuberculosis* can also be presented to a wide variety of T cells, including subsets of CD4-, CD8-TCRab T cells, TCRgd T cells, and a subset of CD8+ TCRab T cells. Importantly, bacterial lipid antigens are not presented in the context of MHC class I or II molecules, but rather are presented in the context of MHC-related CD1 molecules. Some gd T cells that recognize lipid antigens via CD1 molecules have very restricted TCR usage, do not need antigen priming to respond to bacterial lipids, and may actually be a form of innate rather than acquired immunity to intracellular bacteria.

Just as foreign antigens are degraded and their peptide fragments presented in the context ofMHCclass I or class II molecules onAPCs, endogenous self-proteins are also degraded and self-peptide fragments are presented to T cells in the context of MHC class I or class II molecules on APCs. In peripheral lymphoid organs, T cells are present that are capable of recognizing self-protein fragments but normally are *anergic* or *tolerant*, i.e., nonresponsive to self-antigenic stimulation, due to lack of self-antigen upregulating APC *co-stimulatory molecules* such as B7-1 and B7-2 (see below).

Once engagement of mature T cellTCR by foreign peptide occurs in the context of self-MHCclass I or class II molecules, binding of non-antigen-specific adhesion ligand pairs such as CD54-CD11/CD18 and CD58-CD2 stabilizes MHC peptide-TCR binding and the expression of these adhesion molecules is upregulated (Fig. 305-7). Once antigen ligation of the TCR occurs, the T cell membrane is partitioned into *lipid membrane microdomains*, or *lipid rafts*, that coalesce the key signaling molecules

TCR/CD3 complex, CD28, CD2, LAT (linker for activation of T cells), intracellular activated (dephosphorylated) src family protein tyrosine kinases (PTKs), and the key CD3z-associated protein-70 (ZAP-70) PTK (Fig. 305-7). Importantly, during T cell activation, the dephosphorylating molecule, CD45, with protein tyrosine phosphatase activity is partitioned away from the TCR complex to allow activating phosphorylation events to occur. The coalescence of signaling molecules of activated T lymphocytes in *microdomains* has suggested that T cell-APCinteractions can be considered *immunologic synapses*, analogous in function to neuronal synapses.

AfterTCR-MHCbinding is stabilized, activation signals are transmitted through the cell to the nucleus that lead to the expression of gene products important in mediating the wide diversity of T cell functions such as the secretion ofIL-2. The TCR does not have intrinsic signaling activity but is linked to a variety of signaling pathways via immunoreceptor tyrosine-based activation motifs (ITAMs) expressed on the various CD3 chains that bind to proteins that mediate signal transduction. Each of the pathways results in the activation of particular transcription factors that control the expression of cytokine and cytokine receptor genes. Thus, antigen-MHC binding to the TCR induces the activation of the src family ofPTKs, fyn and lck (lck is associated with CD4 or CD8 co-stimulatory molecules); phosphorylation of CD3z chain; activation of the related tyrosine kinases ZAP-70 and syk; and downstream activation of the calcium-dependent calcineurin pathway, the ras pathway, and the protein kinase C pathway. Each of these pathways leads to activation of specific families of transcription factors (including NF-AT, fos and jun, and rel/NF-kB) that form heteromultimers capable of inducing expression of IL-2, IL-2 receptor, IL-4,TNF-a, and other T cell mediators. The src family kinases require dephosphorylation of an inactivation site by CD45 phosphatase before they can be phosphorylated on an activation site. Furthermore, the activity through the receptor is downregulated by the csk-PEP enzyme, a phosphatase that inactivates the src family kinases.

In addition to the signals delivered to the T cell from theTCR and CD4 and CD8 molecules, co-stimulatory receptors [such as CD28 activated by CD80 (B7-1) and/or CD86 (B7-2)] also deliver important signals that upregulate the function of the T cell. The CD28 signal transduction pathway appears particularly important. CD28 signals through phosphoinositide-3-phosphate kinase; its downstream effects are not completely clear. However, if signal transduction through CD28 does not occur in concert with TCR ligation, or if CD28 is blocked, the T cell becomes inactivated or anergic (nonresponsive or tolerant) rather than activated.

CTLA-4 (CD52) is an Ig superfamily molecule on T cells that, like CD28, is a ligand for B7-1 and B7-2 but has a higher affinity for B7-1 and B7-2 than does CD28. T cell CTLA-4 ligation sends a negative signal to the T cell to become tolerant or nonresponsive to antigen stimulation afterTCR-MHCligation. Blocking of CD28-mediated co-stimulation occurs when a second ligand for B7-1 and B7-2 ligates anAPCwhile the TCR is bound to MHC. Thus, a convergence of molecular and biochemical events involving co-stimulatory molecules is required for normal T cell recognition of antigen and consequent T cell activation. In order to exploit this biology for therapeutic purposes, one clinical strategy currently being tested is administration of soluble recombinant CTLA-4 protein to patients at the time of organ transplantation in order to induce a cohort of organ transplant-specific tolerant T cells and thereby reduce the rejection of organ allografts. Alternatively, blocking CTLA-4/B7 interactions with soluble CD28 or CTLA-4 monoclonal antibodies might possibly be therapeutically useful to enhance immune responses to human cancers (see "Immunotherapy," below).

T CELL SUPERANTIGENS

Conventional antigens bind toMHCclass I or II molecules in the groove of the abheterodimer and bind to T cells via the V regions of theTCRa and -b chains (Fig. 305-7). In contrast, superantigens bind directly to the lateral portion of the TCRb chain and MHC class II b chain and stimulate T cells based solely on the Vb gene segment utilized independent of the D, J, and Va sequences present. *Superantigens* are protein molecules capable of activating up to 20% of the peripheral T cell pool, whereas conventional antigens activate fewer than 1 in 10,000 T cells. T cell superantigens include staphylococcal enterotoxins, other bacterial products, and certain nonhuman retroviral proteins. Superantigen stimulation of human peripheral T cells occurs in the clinical setting of the *staphylococcal toxic shock syndrome*, leading to massive overproduction of T cell cytokines (Chap. 139).

B CELLS

Mature B cells comprise 10 to 15% of human peripheral blood lymphocytes, 50% of splenic lymphocytes, and approximately 10% of bone marrow lymphocytes. B cells express on their surface intramembrane immunoglobulin (Ig) molecules that function as B cell receptors (BCR) for antigen in a complex of Ig-associated a andb signaling molecules with properties similar to those described in T cells (Fig. 305-9). Unlike T cells, which recognize only processed peptide fragments of conventional antigens embedded in the notches ofMHCclass I and class II antigens ofAPCs, B cells are capable of recognizing and proliferating to whole unprocessed native antigens via antigen binding to B cell surface Ig (sIg) receptors. B cells also express surface receptors for the Fc region of IgG molecules (CD32) as well as receptors for activated complement components (C3d or CD21, C3b or CD35). The primary function of B cells is to produce antibodies. B cells also serve as APCs and are highly efficient at antigen processing. Their antigen-presenting function is enhanced by a variety of cytokines. Mature B cells are derived from bone marrow precursor cells that arise continuously throughout life (Figs. 305-2,305-10).

B lymphocyte development can be separated into antigen-independent and antigen-dependent phases. Antigen-independent B cell development occurs in primary lymphoid organs, including fetal liver and bone marrow, and includes all stages of B cell maturation up to the sIg+ mature B cell. Antigen-dependent B cell maturation is driven by the interaction of antigen with the mature B cell sIg, leading to memory B cell induction, Ig class switching, and plasma cell formation. Antigen-dependent stages of B cell maturation occur in secondary lymphoid organs, including lymph node, spleen, and gut Peyer's patches. In contrast to the T cell repertoire that is for the most part generated intrathymically before contact with foreign antigen, the repertoire of B cells expressing diverse antigen-reactive sites is modified by further alteration of Ig genes after stimulation by antigen -- a process called *somatic mutation* -- which occurs in lymph node germinal centers.

During B cell development, diversity of the antigen-binding variable region of Ig is generated by an ordered set of Ig gene rearrangements that are similar to the rearrangements undergone byTCRa, b,g, and d genes. Heavy chain rearrangements precede those for light chains. For the heavy chain, there is first a rearrangement of D segments to J segments, followed by a second rearrangement between a V gene segment and the newly formed D-J sequence; the C segment is aligned to the V-D-J complex to yield a functional Ig heavy chain gene (V-D-J-C). During later stages, a functional k orl light chain gene is generated by rearrangement of a V segment to a J segment, ultimately yielding an intact Ig molecule composed of heavy and light chains.

The process of Ig gene rearrangement is regulated to result in a single antibody specificity produced by each B cell, with each Ig molecule comprising one type of heavy chain and one type of light chain. Although each B cell contains two copies of Ig light and heavy chain genes, only one gene of each type is productively rearranged and expressed in each B cell, a process termed *allelic exclusion*.

There are approximately 300 V_k genes and 5 J_k genes, resulting in the pairing of $V_k$and J_k genes to create over 1500 different light chain combinations. The number of distinctk light chains that can be generated is increased by somatic mutations within the V_k and J_k genes, thus creating large numbers of possible specificities from a limited amount of germ-line genetic information. As noted above, in heavy chain Ig gene rearrangement, the VH domain is created by the joining of three types of germ-line genes called V_H, D_H, and J_H, thus allowing for even greater diversity in the variable region of heavy chains than of light chains.

The most immature B cell precursors (early pro-B cells) lack cytoplasmic Ig (cIg) and sIg (Fig. 305-10). The large pre-B cell is marked by the acquisition of the surface pre-BCRcomposed of u heavy (H) chains and a pre-B light chain, termedyLC (Fig. 305-9). yLC is a surrogate light chain receptor encoded by the nonrearranged V pre-B and thel5 light chain locus (the pre-BCR). Pro- and pre-B cells are driven to proliferate and mature by signals from bone marrow stroma, in particular,IL-7. Light chain rearrangement occurs in the small pre-B cell stage such that the full BCR is expressed at the immature B cell stage. Immature B cells have rearranged Ig light chain genes and express sIgM. As immature B cells develop into mature B cells, sIgD is expressed as well as sIgM. At this point, B lineage development in bone marrow is complete, and B cells exit into the peripheral circulation and migrate to secondary lymphoid organs to encounter specific antigens.

Random rearrangements of Ig genes occasionally generates self-reactive antibodies, and mechanisms must be in place to correct these mistakes. One such mechanism isBCRediting, whereby autoreactive BCRs are mutated to not react with self-antigens. If receptor editing is unsuccessful in eliminating autoreactive B cells, then autoreactive B cells undergo negative selection in the bone marrow through induction of apoptosis after BCR engagement of self-antigen.

After leaving the bone marrow, B cells populate peripheral B cell sites, such as lymph node and spleen, and await contact with foreign antigens that react with each B cell's clonotypic receptor. As antigen-driven B cell activation occurs through theBCR, a process known as *somatic hypermutation* takes place whereby point mutations in rearranged H- and L-genes give rise to mutant sIg molecules, some of which bind antigen better than the original sIg molecules. Somatic hypermutation, therefore, is a process whereby memory B cells in peripheral lymph organs have the best binding, or the highest affinity antibodies. This overall process of generating the best antibodies is called *affinity maturation of antibody*.

Lymphocytes that synthesize IgG, IgA, and IgE are derived from sIgM+, sIgD+ mature B cells. Ig class switching occurs in lymph node and other peripheral lymphoid tissue germinal centers. Pairs of CD40+ B cells and CD40 ligand+ T cells bind and drive B cell Ig switching via T cell-produced cytokines such asIL-4 andTGF-b. IL-1, -2, -4, -5, and -6 synergize to drive mature B cells to proliferate and differentiate into Ig-secreting cells.

Humoral Mediators of Adaptive Immunity: Immunoglobulins Immunoglobulins are the products of differentiated B cells and mediate the humoral arm of the immune response. The primary functions of antibodies are to bind specifically to antigen and bring about the inactivation or removal of the offending toxin, microbe, parasite, or other foreign substance from the body. The structural basis of Ig molecule function and Ig gene organization has provided insight into the role of antibodies in normal protective immunity, pathologic immune-mediated damage by immune complexes, and autoantibody formation against host determinants.

All immunoglobulins have the basic structure of two heavy and two light chains (Figs. 305-9 and305-11). Immunoglobulin isotype (i.e., G, M, A, D, E) is determined by the type of Ig heavy chain present. IgG and IgA isotypes can be divided further into subclasses (G1, G2, G3, G4, and A1, A2) based on specific antigenic determinants on Ig heavy chains. The characteristics of human immunoglobulins are outlined inTable 305-9. The four chains are covalently linked by disulfide bonds. Each chain is made up of a V region and C regions (also called *domains*), themselves made up of units of approximately 110 amino acids. Light chains have one variable (V_L) and one constant (C_L) unit; heavy chains have one variable unit (V_H) and three or four constant (C_H) units, depending on isotype. As the name suggests, the constant, or C, regions of Ig molecules are made up of homologous sequences and share the same primary structure as all other Ig chains of the same isotype and subclass. Constant regions are involved in biologic functions of Ig molecules. The C_{H2}domain of IgG and the C_{H4}units of IgM are involved with the binding of the C1q portion of C1. The $C_H$region at the carboxy-terminal end of the IgG molecule, the Fc region (Fig. 305-11), binds to surface Fc receptors (CD16, CD32, CD64) of macrophages,LGLs, B cells, neutrophils, and eosinophils.

Variable regions (V_L and V_H) constitute the antibody-binding (Fab) region of the molecule. Within the V_L and $V_H$regions are hypervariable regions (extreme sequence variability) that constitute the antigen-binding site unique to each Ig molecule. The idiotype is defined as the specific region of the Fab portion of the Ig molecule to which antigen binds. Antibodies against the idiotype portion of an antibody molecule are called *anti-idiotype antibodies*. The formation of such antibodies in vivo during a normal B cell antibody response may generate a negative (or "off") signal to B cells to terminate antibody production.

IgG comprises approximately 75 to 85% of total serum immunoglobulin. The four IgG subclasses are numbered in order of their level in serum, IgG1 being found in greatest amounts and IgG4 the least. IgG subclasses have clinical relevance in their varying ability to bind macrophage and neutrophil Fc receptors and to activate complement (Table 305-9). Moreover, selective deficiencies of certain IgG subclasses give rise to clinical syndromes in which the patient is inordinately susceptible to bacterial infections. IgG antibodies are frequently the predominant antibody made after rechallenge of the host with antigen (secondary antibody response).

IgM antibodies normally circulate as a 950-kDa pentamer with 160-kDa bivalent monomers joined by a molecule called the *J chain*, a 15-kDa nonimmunoglobulin molecule that also effects polymerization of IgA molecules. IgM is the first immunoglobulin to appear in the immune response (primary antibody response) and is the initial type of antibody made by neonates. Membrane IgM in the monomeric form also functions as a major antigen receptor on the surface of mature B cells (Fig. 305-9). IgM is an important component of immune complexes in autoimmune diseases. For example, IgM antibodies against IgG molecules (rheumatoid factors) are present in high titers in *rheumatoid arthritis*, other collagen diseases, and some infectious diseases (*subacute bacterial endocarditis*). IgM antibody binds the C1 component of complement via the CH4 domain and thus is a potent activator of the complement cascade.

IgA comprises only 7 to 15% of total serum immunoglobulin but is the predominant class of immunoglobulin in secretions. IgA in secretions (tears, saliva, nasal secretions, gastrointestinal tract fluid, and human milk) is in the form of secretory IgA (sIgA), a polymer consisting of two IgA monomers, a joining molecule, again called the J chain, and a glycoprotein called the *secretory protein*. Of the two IgA subclasses, IgA1 is primarily found in serum, whereas IgA2 is more prevalent in secretions. IgA fixes complement via the alternative complement pathway and has potent antiviral activity in humans by prevention of virus binding to respiratory and gastrointestinal epithelial cells.

IgD is found in minute quantities in serum and, together with IgM, is a major receptor for antigen on the B cell surface (Table 305-9). IgE, which is present in serum in very low concentrations, is the major class of immunoglobulin involved in arming mast cells and basophils by binding to these cells via the Fc region. Antigen cross-linking of IgE molecules on basophil and mast cell surfaces results in release of mediators of the immediate hypersensitivity response (Table 305-6).

CELLULAR INTERACTIONS IN REGULATION OF NORMAL IMMUNE RESPONSES

The net result of activation of the humoral (B cell) and cellular (T cell) arms of the adaptive immune system by foreign antigen is the elimination of antigen directly by specific effector T cells or in concert with specific antibody. In addition, regulatory T cells are activated that modulate effector T cell activation and B cell antibody production. Figure 305-12 is a simplified schematic diagram of the T and B cell responses indicating some of these cellular interactions.

The expression of adaptive immune cell function is the result of a complex series of immunoregulatory events that occur in phases. Both T and B lymphocytes mediate immune functions, and each of these cell types, when given appropriate signals, passes through stages, from activation and induction through proliferation, differentiation, and ultimately effector functions. The effector function expressed may be at the end point of a response, such as secretion of antibody by a differentiated plasma cell, or it might serve a regulatory function that modulates other functions, such as is seen with CD4+ inducer or CD8+ regulatory T lymphocytes, which modulate both differentiation of B cells and activation of CD8+ or CD4+ cytotoxic T cells.

CD4 helper T cells can be subdivided on the basis of cytokines produced (Fig. 305-13). Activated TH1-type helper T cells secreteIL-2,IFN-g, IL-3,TNF-a,GM-CSF, and TNF-b, while activated TH2-type helper T cells secrete IL-3, -4, -5, -6, -10, and -13. TH1 CD4+ T cells, through elaboration of IFN-g, have a central role in mediating intracellular killing by a variety of pathogens. TH1 CD4+ T cells also provide T cell help for generation of cytotoxic T cells and some types of opsonizing antibody, and generally respond to antigens that lead to delayed hypersensitivity types of immune responses for many intracellular viruses and bacteria (such as HIV or *M. tuberculosis*). In contrast, TH2 cells have a primary role in regulatory humoral immunity and isotype switching. In addition, TH2 cells, through production of IL-4 and IL-10, have a regulatory role in limiting proinflammatory responses mediated by TH1 cells (Table 305-5). In addition, TH2 CD4+ T cells provide help to B cells for specific Ig production and respond to antigens that require high antibody levels for foreign antigen elimination (extracellular encapsulated bacteria such as *Streptococcus pneumoniae* and certain parasite infections). Different cytokines can drive the immune response preferentially towards a TH1 or a TH2 response. For example,APC-derived IL-12 induces CD4+ T cell differentiation towards a TH1 type cell, whereas IL-4 drives differentiation towards a TH2 type cell (Fig. 305-13).

As shown in Fig. 305-12, upon activation byAPCssuch as dendritic cells, regulatory T cell subsets that produceIL-2, IL-3,IFN-g, and/or IL-4, -5, -6, -10, and -13 are generated that exert positive and negative influences on effector T and B cells. For B cells, trophic effects are mediated by a variety of cytokines, particularly T cell-derived IL-3, -4, -5, and -6, which act at sequential stages of B cell maturation, resulting in B cell proliferation, differentiation, and ultimately antibody secretion (Table 305-5). For cytotoxic T cells, trophic factors include inducer T cell secretion of IL-2, IFN-g, and IL-12 (Table 305-5). In addition, B cells themselves are capable of serving as APCs, processing and presenting antigens to T cells, and secretingTNF-aand IL-6.

Although B cells recognize native antigen via B cell surface Ig receptors, B cells require T cell help to produce high-affinity antibody of multiple isotypes that are the most effective in eliminating foreign antigen. This T cell dependence likely functions in the regulation of B cell responses and in protection against excessive autoantibody production. T cell-B cell interactions that lead to high-affinity antibody production require: (1) processing of native antigen by B cells and expression of peptide fragments on the B cell surface for presentation to THcells, (2) the ligation of B cells both by the T cell receptor complex and the CD40 ligand, (3) induction of the process termed *antibody isotype switching* in antigen-specific B cell clones, and (4) induction of the process of *affinity maturation* of antibody in the germinal centers of B cell follicles of lymph node and spleen.

Naive B cells express cell-surface IgD and IgM, and initial contact of naive B cells with antigen is via binding of native antigen to B cell-surface IgM. T cell cytokines, released following TH2 cell contact with B cells or by a "bystander" effect, induce changes in Ig gene conformation that promote recombination of Ig genes. These events then result in the "switching" of expression of heavy chain exons in a triggered B cell, leading to the secretion of IgG, IgA, or, in some cases, IgE antibody with the same V region antigen specificity as the original IgM antibody, for response to a wide variety of extracellular bacteria, protozoa, and helminths. CD40 ligand expression by activated T cells is critical for induction of B cell antibody isotype switching and for B cell responsiveness to cytokines. Patients with mutations in T cell CD40 ligand have B cells that are unable to undergo isotype switching, resulting in lack of memory B cell generation and the immunodeficiency syndrome of *X-linked hyper-IgM syndrome* (Chap. 308).

MECHANISMS OF IMMUNE-MEDIATED DAMAGE TO MICROBES OR HOST TISSUES

Several responses by the host innate and adaptive immune systems to foreign microbes culminate in rapid and efficient elimination of microbes. In these scenarios, the classic weapons of the adaptive immune system (T cells, B cells) interface with cells (macrophages, dendritic cells,NKcells, neutrophils, eosinophils, basophils) and soluble products (microbial peptides, pentraxins, complement and coagulation systems) of the innate immune system (Chaps. 64 and310).

There are five general phases of host defenses: (1) migration of leukocytes to sites of antigen localization; (2) antigen nonspecific recognition of pathogens by macrophages and other cells and systems of the innate immune system; (3) specific recognition of foreign antigens mediated by T and B lymphocytes; (4) amplification of the inflammatory response with recruitment of specific and nonspecific effector cells by complement components, cytokines, kinins, arachidonic acid metabolites, and mast cell-basophil products; and (5) macrophage, neutrophil, and lymphocyte participation in destruction of antigen with ultimate removal of antigen particles by phagocytosis (by macrophages or neutrophils) or by direct cytotoxic mechanisms (involving macrophages, neutrophils, and lymphocytes). Under normal circumstances, orderly progression of host defenses through these phases results in a well-controlled immune and inflammatory response that protects the host from the offending antigen. However, dysfunction of any of the host defense systems can damage host tissue and produce clinical disease. Furthermore, for certain pathogens or antigens, the normal immune response itself might contribute substantially to the tissue damage. For example, the immune and inflammatory response in the brain to certain pathogens such as *M. tuberculosis* may be responsible for much of the morbidity of this disease in that organ system (Chap. 169). In addition, the morbidity associated with certain pneumonias such as that caused by *Pneumocystis carinii* may be associated more with inflammatory infiltrates than with the tissue destructive effects of the microorganism itself (Chap. 209). Thus, it is important to appreciate how normally protective proinflammatory responses that mediate intracellular killing are regulated. What follows are brief discussions of mechanisms of leukocyte migration to sites of inflammation, immune complex formation, immediate-type hypersensitivity responses, cytotoxic reactions of antibody, delayed cellular types of hypersensitivity responses, and programmed cell death of immune competent cells.

The Molecular Basis of Lymphocyte-Endothelial Cell Interactions The control of lymphocyte circulatory patterns between the bloodstream and peripheral lymphoid organs operates at the level of lymphocyte-endothelial cell interactions to control the specificity of lymphocyte subset entry into organs. Similarly, lymphocyte-endothelial cell interactions regulate the entry of lymphocytes into inflamed tissue. Adhesion molecule expression on lymphocytes and endothelial cells regulates the retention and subsequent egress of lymphocytes within tissue sites of antigenic stimulation, delaying cell exit from tissue and preventing reentry into the circulating lymphocyte pool. All types of lymphocyte migration begin with lymphocyte attachment to specialized regions of vessels, termed *high endothelial venules* (HEVs). An important concept for many of the adhesion molecules listed in Table 305-10 is that the molecules do not generally bind their ligand until a conformational change (ligand activation) occurs in the adhesion molecule that allows ligand binding. Induction of a conformation-dependent determinant on an adhesion molecule can be accomplished by cytokines or via ligation of other adhesion molecules on the cell.

The first stage of lymphocyte-endothelial cell interactions, *attachment and rolling*, occurs when lymphocytes leave the stream of flowing blood cells in a postcapillary venule and roll along venule endothelial cells (Fig. 305-14). Lymphocyte rolling is mediated by the L-selectin molecule (LECAM-1, LAM-1) and slows cell transit time through venules, allowing time for activation of adherent cells.

The second stage of lymphocyte-endothelial cell interactions, *adhesion triggering*, requires stimulation of lymphocytes by chemoattractants or by endothelial cell-derived cytokines. Cytokines thought to participate in adherent cell triggering include members of theIL-8 family, platelet-activation factor, leukotriene B_4, and C5a. Following activation by chemoattractants, lymphocytes shed L-selectin from the cell surface and upregulate cell CD11b/18 (MAC-1) or CD11a/18 (LFA-1) molecules, resulting in firm attachment of lymphocytes toHEVs.

Lymphocyte homing to peripheral lymph nodes involves adhesion of L-selectin to carbohydrate of peripheral nodeHEVs, whereas homing of lymphocytes to intestine Peyer's patches primarily involves adhesion of the a4,b7 integrin to MAdCAM-1 oligosaccharides on the Peyer's patch HEVs. However, for migration to mucosal Peyer's patch lymphoid aggregates, naive lymphocytes primarily use L-selectin, whereas memory lymphocytes usea4,b7 integrin.a4,b1 integrin (CD49d/CD29, VLA-4)-VCAM-1 interactions are important in the initial interaction of memory lymphocytes with HEVs of multiple organs in sites of inflammation.

The third stage of leukocyte emigration inHEVs, *sticking and arrest*, is sticking of the lymphocyte and arrest at the site of sticking, mediated predominantly by ligation ofaL,b2 integrin LFA-1 to the integrin ligands ICAM-1 and ICAM-2 on HEVs. While the first three stages of lymphocyte attachment to HEVs takes only a few seconds, the fourth stage of lymphocyte emigration, *transendothelial migration*, takes approximately 10 min. Although the molecular mechanisms that control lymphocyte transendothelial migration are not fully characterized, the HEV CD44 molecule and molecules of the HEV glycocalyx (extracellular matrix) are thought to play important regulatory roles in this process (Fig. 305-14). Finally, expression of matrix metalloproteases capable of digesting the subendothelial basement membrane, rich in nonfibrillar collagen, appears to be required for the penetration of lymphoid cells into the extravascular sites.

Abnormal induction ofHEVformation and use of the molecules discussed above have been implicated in the induction and maintenance of inflammation in a number of chronic inflammatory diseases. In animal models of insulin-dependent diabetes mellitus (IDDM), MAdCAM-1 and GlyCAM-1 have been shown to be highly expressed on HEVs in inflamed pancreatic islets, and treatment of these animals with inhibitors of L-selectin and a4 integrin function blocked the development of IDDM (Chap. 333). A similar role for abnormal induction of the adhesion molecules of lymphocyte emigration has been suggested in rheumatoid arthritis (Chap. 312), Hashimoto's thyroiditis (Chap. 330), Graves' disease (Chap. 330), multiple sclerosis (Chap. 371), Crohn's disease (Chap. 287), and ulcerative colitis (Chap. 287).

Immune-Complex Formation Clearance of antigen by immune-complex formation between antigen and antibody is a highly effective mechanism of host defense. However, depending on the level of immune complexes formed and their physicochemical properties, immune complexes may or may not result in host and foreign cell damage. After antigen exposure, certain types of soluble antigen-antibody complexes freely circulate and, if not cleared by the reticuloendothelial system, can be deposited in blood vessel walls and in other tissues such as renal glomeruli (Chap. 317).

Immediate-Type Hypersensitivity Helper T cells that drive antiallergen IgE responses are usually T_H2-type inducer T cells that secreteIL-4, IL-5, IL-6, and IL-10. Mast cells and basophils have high-affinity receptors for the Fc portion of IgE (FcRI), and cell-bound antiallergen IgE effectively "arms" basophils and mast cells. Mediator release is triggered by antigen (allergen) interaction with Fc receptor-bound IgE; the mediators released are responsible for the pathophysiologic changes of allergic diseases (Table 305-6). Mediators released from mast cells and basophils can be divided into three broad functional types: (1) those that increase vascular permeability and contract smooth muscle (histamine, platelet-activating factor, SRS-A, BK-A), (2) those that are chemotactic for or activate other inflammatory cells (ECF-A, NCF, leukotriene B_4), and (3) those that modulate the release of other mediators (BK-A, platelet-activating factor) (Chap. 310).

Cytotoxic Reactions of Antibody In this type of immunologic injury, complement-fixing (C1-binding) antibodies against normal or foreign cells or tissues (IgM, IgG1, IgG2, IgG3) bind complement via the classic pathway and initiate a sequence of events similar to that initiated by immune-complex deposition, resulting in cell lysis or tissue injury. Examples of antibody-mediated cytotoxic reactions include red cell lysis in *transfusion reactions*, *Goodpasture's syndrome* with anti-glomerular basement membrane antibody formation, and *pemphigus vulgaris* with antiepidermal antibodies inducing blistering skin disease.

Classic Delayed-Type Hypersensitivity Reactions Inflammatory reactions initiated by mononuclear leukocytes and not by antibody alone have been termed *delayed-type hypersensitivity reactions*. The term *delayed* has been used to contrast a secondary cellular response that appears 48 to 72 h after antigen exposure with an *immediate* hypersensitivity response generally seen within 12 h of antigen challenge and initiated by basophil mediator release or preformed antibody. For example, in an individual previously infected with *M. tuberculosis* organisms, intradermal placement of tuberculin purified-protein derivative as a skin test challenge results in an indurated area of skin at 48 to 72 h, indicating previous exposure to tuberculosis.

The cellular events that result in classic delayed-type hypersensitivity responses are centered around T cells (predominantly, though not exclusively,IFN-g,IL-2, andTNF-a-secreting TH1-type helper T cells) and macrophages. First, local immune and inflammatory responses at the site of foreign antigen upregulate endothelial cell adhesion molecule expression, promoting the accumulation of lymphocytes at the tissue site. In the general scheme outlined in Fig. 305-12, antigen is processed by dendritic/Langerhans cells or monocytes-macrophages and presented to small numbers of CD4+ T cells expressing aTCRspecific for the antigen. IL-12 produced byAPCsinduces T cells to produce IFN-g(TH-1 response). IL-1 and IL-6 secreted by APCs amplify the clonal expansion of antigen-specific T cells, and other cytokines (primarily IL-2, IFN-g, and TNF-b) are secreted that promote recruitment of diverse populations of T cells and macrophages to the site of the cellular inflammatory response. In particular, CD8+ cytotoxic T cells are induced by IL-2 to become active killer cells. Once recruited, macrophages frequently undergo epithelioid cell transformation and fuse to form multinucleated giant cells. This type of mononuclear cell infiltrate is termed *granulomatous inflammation*. Examples of diseases in which delayed-type hypersensitivity plays a major role are fungal infections (*histoplasmosis*) (Chap. 201), mycobacterial infections (*tuberculosis*, *leprosy*) (Chaps. 169 and170), chlamydial infections (*lymphogranuloma venereum*) (Chap. 179), helminth infections (*schistosomiasis*) (Chap. 224), reactions to toxins (*berylliosis*) (Chap. 254), and hypersensitivity reactions to organic dusts (*hypersensitivity pneumonitis*) (Chap. 253). In addition, delayed-type hypersensitivity responses play important roles in tissue damage in autoimmune diseases such as *rheumatoid arthritis*, *temporal arteritis*, and *Wegener's granulomatosis* (Chaps. 312 and317).

The Cellular and Molecular Control of Programmed Cell Death (Apoptosis) The process of apoptosis plays a crucial role in regulating normal immune responses to antigen. In general, a wide variety of stimuli trigger cell surface receptors (e.g.,TNFreceptor family members or related proteins) or cytoplasmic receptors (e.g., ceramide, glucocorticoids) that activate groups of proteases such as FADD-likeIL-1b-converting enzyme (FLICE) or Caspase 8 (Fig. 305-15). These proteases either cleave molecules that lead to cell death themselves or activate other enzymes to cleave molecules that eventuate in cell death (Fig. 305-15). The end stages of this sequence of events lead to cell death characterized by degradation of cytoplasmic (actin) and nuclear cytoskeletal proteins as well as cleavage of DNA at regular intervals (nucleosomes), leading to nuclear disintegration seen on electron microscopy and "laddering" of DNA when analysed by agarose gel electrophoresis. The level of expression of certain cytosolic proteins, such as Bcl-2 and Bcl-XL, negatively regulates the process of apoptosis by inhibiting activation of cytosolic proteases that induce cell death. For example, T cells that are negatively selected in the thymus are induced to undergo apoptosis and have low levels of proteins such as Bcl-2, whereas medullary thymocytes that have been triggered to proliferate and survive thymocyte selection (positive selection) have high levels of Bcl-2.

Thus, in the immune system, apoptosis is a mechanism induced to remove autoreactive T cells from the thymus during negative selection, to remove autoreactive B and T cells from peripheral lymphoid organs upon contact with antigen or antigen-reactive helper T cells in spleen and lymph node, and to remove virus-infected or malignant cells after contact with antigen-specific CD8+ cytotoxic T lymphocytes. Induction of apoptosis is one of two principal mechanisms of target cell lysis by cytotoxic T lymphocytes, the other consisting of the release of cytotoxic perforin molecules.

CLINICAL EVALUATION OF IMMUNE FUNCTION

Clinical assessment of immunity requires investigation of the four major components of the immune system that participate in host defense and in the pathogenesis of autoimmune diseases: (1) humoral immunity (B cells); (2) cell-mediated immunity (T cells, monocytes); (3) phagocytic cells of the reticuloendothelial system (macrophages), as well as polymorphonuclear leukocytes; and (4) complement. Clinical problems that require an evaluation of immunity include chronic infections, recurrent infection, unusual infecting agents, and certain autoimmune syndromes. The type of clinical syndrome under evaluation can provide information regarding possible immune defects (Chap. 308). Defects in cellular immunity generally result in viral, mycobacterial, and fungal infections. An extreme example of deficiency in cellular immunity is AIDS (Chap. 309). Antibody deficiencies result in recurrent bacterial infections, frequently with organisms such as *S. pneumoniae* and *Haemophilus influenzae* (Chap. 308). Disorders of phagocyte function frequently are manifested by recurrent skin infections, often due to *Staphylococcus aureus* (Chap. 64). Finally, deficiencies of early and late complement components are associated with autoimmune phenomena and recurrent *Neisseria* infections (Table 305-10).**For further discussion of useful initial screening tests of immune function, see Chap. 308.*

IMMUNOTHERAPY

Most current therapies for autoimmune and inflammatory diseases involve the use of nonspecific immune-modulating or immunosuppressive agents such as glucocorticoids or cytotoxic drugs. The goal of development of new treatments for immune-mediated diseases is to design ways to specifically interrupt pathologic immune responses, leaving nonpathologic immune responses intact. Novel ways to interrupt pathologic immune responses that are under investigation include: the use of anti-inflammatory cytokines or specific cytokine inhibitors as anti-inflammatory agents; the use of monoclonal antibodies against T or B lymphocytes as therapeutic agents; the induction of anergy by administration of soluble CTLA-4 protein, the use of intravenous Ig for certain infections and immune complex-mediated diseases, and the use of specific cytokines to reconstitute components of the immune system (Table 305-11) (Chaps. 64,308, and309).

Cytokines and Cytokine Inhibitors Recently a humanized mouse anti-TNF-amonoclonal antibody (MAB) has been tested in both rheumatoid arthritis and ulcerative colitis. Use of anti-TNF-a antibody therapy has resulted in clinical improvement in patients with these diseases and has opened the way for targeting TNF-a to treat other severe forms of autoimmune and/or inflammatory disease. Anti-TNF-a MAB has been approved for treatment of patients with rheumatoid arthritis.

Other cytokine inhibitors under investigation are recombinant solubleTNF-areceptor (R) fused to human Ig and solubleIL-1 receptor (termed *IL-1 receptor antagonist*, or IL-1 ra). Soluble TNF-aR and IL-1 ra act to inhibit the activity of pathogenic cytokines in rheumatoid arthritis, i.e., TNF-a and IL-1 respectively. Similarly, anti-IL-6,IFN-b, and IL-11 act to inhibit pathogenic proinflammatory cytokines. Anti-IL-6 inhibits IL-6 activity, while IFN-b and IL-11 decrease IL-1 and TNF-aproduction.

Recent studies have identified mutations in theIL-12 gene in patients susceptible to severe myobacterial infections. IL-12 is a critical cytokine for induction ofIFN-gand cytotoxic T lymphocytes (CTLs) against intracellular organisms; it is under study for treatment of severe infections such as that caused by *M. tuberculosis* and for treatment of various cancers. In this latter setting, IL-12 is being studied for its ability to enhance antitumor cellular immunity by enhancing the induction of antitumor CTL.

Of particular note has been the successful use ofIFN-gin the treatment of the phagocytic cell defect in *chronic granulomatous disease* (Chap. 64). Intermittent infusions ofIL-2 in HIV-infected individuals in the early or intermediate stages of disease have resulted in substantial and sustained increases in CD4+ T cells.

Monoclonal Antibodies to T and B Cells The OKT3MABagainst human T cells has been used for several years as a T cell-specific immunosuppressive agent that can substitute for horse anti-thymocyte globulin (ATG) in the treatment of solid organ transplant rejection. OKT3 produces fewer allergic reactions than ATG but does induce human anti-mouse Ig antibody -- thus limiting its use. Anti-CD4 MAB therapy has been used in trials to treat patients with rheumatoid arthritis. While inducing profound immunosuppression, anti-CD4 MAB treatment also induces considerable susceptibility to severe infections. Treatment of patients with a MAB against the T cell molecule CD40 ligand (CD154) is under investigation to induce tolerance to organ transplants, with promising results reported in animal studies.

Tolerance Induction Specific immunotherapy has moved into a new era with the introduction of soluble CTLA-4 protein into clinical trials. Use of this molecule to block T cell activation viaTCR/CD28 ligation during organ or bone marrow transplantation has showed promising results in animals and in early human clinical trials. Specifically, treatment of bone marrow with CTLA-4 protein reduces rejection of the graft in HLA-mismatched bone marrow transplantation. In addition, promising results with soluble CTLA-4 have been reported in the downmodulation of autoimmune T cell responses in the treatment of psoriasis.

Intravenous Immunoglobulin (IVIg) IVIg has been successfully used to block reticuloendothelial cell function and immune complex clearance in various immune cytopenias such as immune thrombocytopenia (Chap. 116). In addition, IVIg is useful for prevention of tissue damage in certain inflammatory syndromes such as Kawasaki's disease (Chap. 317) and as Ig replacement therapy for certain types of immunoglobulin deficiencies (Chap. 308). In addition, controlled clinical trials support the use of IVIg in selected patients with graft-versus-host disease, multiple sclerosis, myasthenia gravis, Guillain-Barre syndrome, and chronic demyelinating polyneuropathy (Table 305-11).

Thus, a number of recent insights into immune system function have spawned a new field of interventional immunotherapy and have enhanced the prospect for development of specific and nontoxic therapies for immune and inflammatory diseases.

(Bibliography omitted in Palm version)

Back to Table of Contents

306. THE MAJOR HISTOCOMPATIBILITY GENE COMPLEX - ***Gerald T. Nepom*, *Joel D. Taurog***

THE HLA COMPLEX AND ITS PRODUCTS

The human major histocompatibility complex (MHC), commonly called the human leukocyte antigen (HLA) complex, is a 4-megabase (Mb) region on chromosome 6 (6p21.3) that is densely packed with expressed genes. The best known of these genes are the HLA class I and class II genes, whose products are critical for immunologic specificity and transplantation histocompatibility; they play a major role in susceptibility to a number of autoimmune diseases. Many other genes in the HLA region are also essential to the innate and antigen-specific functioning of the immune system. The HLA region shows extensive conservation with the MHC of other mammals in terms of genomic organization, gene sequence, and protein structure and function. Much of our understanding of the MHC has come from investigation of the MHC in mice, where it is termed the *H-2 complex*, and to a lesser degree from other species as well. Nonetheless, in this chapter the discussion will be confined to information applicable to the MHC in humans.

The *HLA class I genes* are located in a 2-Mb stretch of DNA at the telomeric end of the HLA region (Fig. 306-1). The classic (MHCclass Ia) HLA-A, -B, and -C loci, the products of which are integral participants in the immune response to intracellular infections, tumors, and allografts, are expressed in all nucleated cells and are highly polymorphic in the population. *Polymorphism* refers to a high degree of allelic variation within a genetic locus that leads to extensive variation between different individuals expressing different alleles. Over 100 alleles at HLA-A, 200 at HLA-B, and 50 at HLA-C have been identified in different human populations. Each of the alleles at these loci encodes a *heavy chain* (also called an*a chain*) that associates noncovalently with the nonpolymorphic light chain*b*$_2$*-microglobulin*, encoded on chromosome 15.

The *nomenclature* of HLA genes and their products reflects the grafting of newer DNA sequence information on an older system based on serology. Among class I genes, alleles of the HLA-A, -B, and -C loci were originally identified in the 1950s, 1960s, and 1970s by alloantisera, derived primarily from multiparous women, who in the course of normal pregnancy produce antibodies against paternal antigens expressed on fetal cells. The serologic allotypes were designated by consecutive numbers, e.g., HLA-A1, HLA-B8. The HLA-C locus alleles were designated HLA-Cw, rather than HLA-C, partly to distinguish them from the HLA-encoded complement loci C2 and C4. With the application of DNA sequence analysis and other molecular techniques in the 1980s, and particularly polymerase chain reaction-based techniques since the late 1980s, most serologically defined specificities were found to include a number of closely related alleles differing by only a few amino acids. These are commonly termed *subtypes* of the parent specificity. Currently, under World Health Organization nomenclature, class I alleles are given a single designation that indicates locus, serologic specificity, and sequence-based subtype. For example, HLA-A*0201 indicates subtype 1 of the serologically defined allele HLA-A2. As new alleles are discovered, they are named and numbered based on sequence homology to known alleles or, in the absence of strong homology, designated as consecutively numbered separate alleles, irrespective of serologic reactivity. Subtypes that differ from each other at the nucleotide but not the amino acid sequence level are designated by an extra numeral; e.g., HLA-B*07021 and HLA-B*07022 are two variants of the HLA-B7 subtype of HLA-B*0702. The nomenclature of class II genes, discussed below, is made more complicated by the fact that both chains of a class II molecule are encoded by closely linked HLA-encoded loci, both of which may be polymorphic, and by the presence of differing numbers of isotypic DRB loci in different individuals. It has become clear that accurate HLA genotyping requires DNA sequence analysis, and the identification of alleles at the DNA sequence level has contributed greatly to the understanding of the role of HLA molecules as peptide-binding ligands, to the analysis of associations of HLA alleles with certain diseases, to the study of the population genetics of HLA, and to a clearer understanding of the contribution of HLA differences in allograft rejection and graft-vs.-host disease. Current databases of HLA class I and class II sequences can be accessed by internet (e.g., from the American Society for Histocompatibility and Immunogenetics,http://www.swmed.edu/home_pages/ASHI/sequences/seq3.htm), and frequent updates of HLA gene lists are published in several journals.

As shown inFig. 306-2 and discussed below in detail, a characteristic structural feature of class I and class II HLA molecules is the *peptide-binding groove* that enables these molecules to form highly stable complexes with a wide array of peptide sequences that can be recognized as antigens by T cells. In the case of class I molecules, peptide binding provides a display on the cell surface of peptides derived from intracellular proteins, and this serves as a readout to CD8+ T cells of the proteins being produced within somatic cells. The polymorphism at the loci encoding these molecules predominantly affects the amino acid residues that make up the peptide-binding groove, further amplifying the array of peptides that can be bound by different HLA molecules and generating important functional immune differences and transplantation incompatibility among different individuals.

The nonclassic, or class Ib,MHCmolecules, HLA-E, -F, and -G, are much less polymorphic than MHC Ia and, except for HLA-E, have a more limited tissue distribution. The HLA-E molecule, which has a peptide repertoire restricted to signal peptides cleaved from classic MHC class I molecules, is the major self-recognition target for the natural killer (NK) cell inhibitory receptors NKG2A or NKG2C paired with CD94 (see below andChap. 305). HLA-G is expressed selectively in extravillous trophoblasts, the fetal cell population directly in contact with maternal tissues. It binds a wide array of peptides, is expressed in six different alternatively spliced forms, and provides inhibitory signals to both NK cells and T cells, presumably in the service of maintaining maternofetal tolerance. The function of HLA-F remains largely unknown. Although HLA-C is considered a classic class I molecule, its degree of polymorphism and level of surface expression are significantly lower than those of HLA-A and HLA-B. Moreover, unlike HLA-A and -B molecules, which function primarily by presenting antigen to CD8+ T cells expressing ab T cell receptors (TCRs), the primary function of HLA-C molecules appears to be to serve as targets of NK cell recognition (see below).

Additional class I-like genes have been identified, some HLA-linked and some encoded on other chromosomes, that show only distant homology to the class Ia and Ib molecules but that share the three-dimensional class I structure. Those on chromosome 6p21 include MIC-A and MIC-B, which are encoded centromeric to HLA-B; and HLA-HFE, located 3 to 4 cM (centi-Morgan) telomeric of HLA-F. MIC-A and MIC-B do not bind peptide but are expressed on gut and other epithelium in a stress-inducible manner and serve as activation signals for certain gd T cells andNKcells, whereas HLA-HFE encodes the gene defective in hereditary hemochromatosis (Chap. 345). Among the non-HLA, class I-like genes, CD1 refers to a family of molecules that present glycolipids or other nonpeptide ligands to certain T cells, including T cells with NK activity; FcRn binds IgG within lysosomes and protects it from catabolism (Chap. 305); and Zn-a$_2$-glycoprotein 1 binds a nonpeptide ligand and promotes catabolism of triglycerides in adipose tissue. Like the HLA-A, -B, -C, -E, -F, and -G heavy chains, each of which forms a heterodimer withb$_2$-microglobulin (Fig. 306-2), the class I-like molecules HLA-HFE, FcRn, and CD1 also bind tob$_2$-microglobulin, but MIC-A, MIC-B, and Zn-a$_2$-glycoprotein 1 do not.

The *HLA class II region* is also illustrated in Fig. 306-1. Multiple class II genes are arrayed within the centromeric 1 Mb of the HLA region, forming distinct haplotypes. A *haplotype* refers to an array of alleles at polymorphic loci along a chromosomal segment. In the context of HLA, haplotype can refer either to a large segment of the HLA region encompassing many of the polymorphic HLA loci (also called an *extended haplotype* or to a more restricted segment such as the tightly linked DR and DQ loci. Multiple class II genes are present on a single haplotype, clustered into three major subregions: HLA-DR, -DQ, and -DP. Each of these subregions contains at least one functional alpha (A) locus and one functional beta (B) locus. Together these encode proteins that form thea and b polypeptide chains of a mature class II HLA molecule. Thus, the DRA and DRB genes encode an HLA-DR molecule; products of the DQA1 and DQB1 genes form an HLA-DQ molecule; and the DPA1 and DPB1 genes encode an HLA-DP molecule. There are several DRB genes (DRB1, DRB2, and DRB3, etc.), so that two expressed DR molecules are encoded on most haplotypes by combining the a-chain product of the DRA gene with separate bchains.

The class II region was originally termed the *D-region*. The allelic gene products were first detected by their ability to stimulate lymphocyte proliferation by *mixed lymphocyte reaction*, and were named Dw1, Dw2, etc. Subsequently, serology was used to identify gene products on peripheral blood B cells, and the antigens were termed *DR* (D-related). After additional class II loci were identified, these came to be known as DQ and DP.

The HLA class II DRB and DPB loci are extremely polymorphic, with over 200 DR alleles and over 75 DP alleles, respectively. In the DQ region, both DQA1 and DQB1 are polymorphic, with 20 DQA1 alleles and over 40 DQB1 alleles. The current nomenclature is largely analogous to that discussed above for class I, using the convention "locus*allele." Thus, for example, subtypes of the serologically defined specificity DR4, encoded by the DRB1 locus, are termed DRB1*0401, -0402, etc. In addition to allelic polymorphism, products of different DQA1 alleles can, with some limitations, pair with products of different DQB1 alleles through both *cis* and *trans* pairing to create combinatorial complexity and expand the number of expressed class II molecules. Because of the enormous allelic diversity in the general population, most individuals are heterozygous at all of the class I and class II loci. Thus, most individuals express six classic class I molecules (two each of HLA-A, -B, and -C) and approximately eight class II molecules -- two DP, two DR (more in the case of haplotypes with additional functional DRB genes), and up to four DQ (two *cis* and two *trans*).

The localization of polymorphic residues in class II molecules is similar to that for class I, i.e., it is predominantly in sites that affect peptide binding (see below). In the case of class II molecules, the peptides displayed on the cell surface are primarily derived from proteins acquired from the extracellular environment, processed through the endosomal-lysosomal pathway, and presented to CD4+ T cells.

OTHER GENES IN THE MHC

Immunologically Relevant Genes In addition to the class I and class II genes themselves, there are numerous genes interspersed among the HLA loci that have interesting and important immunologic functions. The current concept of the function ofMHCgenes now encompasses many of these additional genes. As discussed in more detail below, TAP and LMP genes encode molecules that participate in intermediate steps in the HLA class I biosynthetic pathway, and deficiencies of the TAP or LMP genes can markedly alter class I-mediated immune recognition. Another set of HLA genes, DMA and DMB, performs an analogous function for the class II pathway. These genes encode an intracellular molecule that facilitates the proper complexing of HLA class II molecules with antigen (see below). The HLA class III region is a name given to a cluster of genes between the class I and class II complexes, which includes genes for the two closely related cytokines tumor necrosis factor (TNF) a and lymphotoxin (TNF-b); the complement components C2, C4, and Bf; heat shock protein (HSP)70; and the enzyme 21-hydroxylase.

The class I genes HLA-A, -B, and -C are expressed in all nucleated cells, although generally to a higher degree on leukocytes than on other cells. In contrast, the class II genes show a more restricted distribution: HLA-DR and HLA-DP genes are constitutively expressed on most cells of the myeloid cell lineage, whereas all three class II gene families (HLA-DR, -DQ, and -DP) are inducible by certain stimuli provided by inflammatory cytokines such as interferon g. Within the lymphoid lineage, expression of these class II genes is constitutive on B cells and inducible on human T cells. Most endothelial and epithelial cells in the body, including the vascular endothelium and the intestinal epithelium, are also inducible for class II gene expression. Thus, while these somatic tissues normally express only class I and not class II genes, during times of local inflammation they are recruited by cytokine stimuli to express class II genes as well, thereby becoming active participants in ongoing immune responses. Other HLA genes involved in the immune response, such as TAP and LMP, are also susceptible to upregulation by signals such as interferong.

Other Genes and Genetic Elements Large-scale genomic sequencing projects have recently yielded sequence data for the entire HLA region, which can be accessed on the internet (e.g.,http://www.sanger.ac.uk/HGP/Chr6/). As a result, many new genes have been discovered, the functions of which remain to be determined, as well as numerous microsatellite regions and other genetic elements. The gene density of the class II region is high, with approximately one protein encoded every 30 kb; that of the class I and class III regions is even higher, with approximately one protein encoded every 15 kb. It is also of interest that these regions also differ with respect to the GC (guanidine + cytosine) content. Vertebrate genomes have a long-range mosaic structure with regard to relative GC content that is related to chromosome banding. Regions of homogeneous

GC content are termed *isochores.* The HLA class I and class III regions belong to the H3 (highest GC) isochore, with 53% GC, whereas the class II region belongs to the L or H1 isochores (low GC), with 40 to 45% GC. An abrupt demarcation between these two isochores occurs near the boundary separating the class II and class III regions.

LINKAGE DISEQUILIBRIUM

In addition to extensive polymorphism at the class I and class II loci, another characteristic feature of the HLA complex is *linkage disequilibrium*. This is formally defined as a deviation from Hardy-Weinberg equilibrium for alleles at linked loci. This is reflected in the very low recombination rates between certain loci within the HLA. For example, recombination between DR and DQ loci is almost never observed in family studies, and characteristic haplotypes with particular arrays of DR and DQ alleles are found in every population. Similarly, the complement components C2, C4, and Bf are almost invariably inherited together, and the alleles at these loci are found in characteristic haplotypes. In contrast, there is a recombinational hotspot between DQ and DP, which are separated by 1 to 2 cM of genetic distance, despite their close physical proximity. Certain extended haplotypes encompassing the interval from DQ into the class I region are commonly found, the most notable being the haplotype DR3-B8-A1, which is found, in whole or in part, in 10 to 30% of northern European Caucasians. The genetic mechanisms that account for linkage disequilibrium in HLA have not been determined. It has been hypothesized that selective pressures may maintain certain haplotypes, but this remains to be demonstrated. As discussed below under HLA and immunologic disease, one consequence of the phenomenon of linkage disequilibrium has been the difficulty it produces in assigning HLA disease associations to a single allele at a single locus.

MHC STRUCTURE AND FUNCTION

Class I and class II molecules display a distinctive structural architecture that contains specialized functional domains responsible for the unique genetic and immunologic properties of the HLA complex. The principal known function of both class I and class II HLA molecules is to bind antigenic peptides in order to present antigen to an appropriate T cell. The ability of a particular peptide to bind to an individual HLA molecule satisfactorily is a direct function of the molecular fit between the amino acid residues on the peptide with respect to the amino acid residues of the HLA molecule. The bound peptide forms a tertiary structure called the *MHC-peptide complex*, which communicates with T lymphocytes through binding to theTCRmolecule. The first site of TCR-MHC-peptide interaction in the life of a T cell occurs in the thymus, where self-peptides are presented to developing thymocytes by MHC molecules expressed on thymic epithelium and hematopoietically derived antigen-presenting cells, which are primarily responsible for positive and negative selection, respectively (seeChap. 305 for details of thymic selection of the T cell repertoire). Mature T cells encounter MHC molecules in the periphery both in the maintenance of tolerance (Chap. 305) and in the initiation of immune responses. Because most antibody responses and all T cell responses are T cell dependent (Chap. 305), the MHC-peptide-TCR interaction is the central event in the initiation of most antigen-specific immune responses, since it is the event that actually confers the specificity. Thus, the population of MHC-T cell complexes expressed in the thymus shapes the TCR repertoire. For potentially immunogenic peptides, the ability of a given peptide to be generated and bound by an HLA molecule is a primary determinant of whether or not an immune response to that peptide can be generated; the repertoire of peptides that a particular individual's HLA molecules can bind exerts a major influence over the specificity of that individual's immune response.

When aTCRmolecule binds to an HLA-peptide complex, it forms intermolecular contacts with both the antigenic peptide and with the HLA molecule itself. The outcome of this recognition event depends on the density and duration of the binding interaction, accounting for a dual specificity requirement for activation of the T cell. That is, the TCR must be specific both for the antigenic peptide and for the HLA molecule. The polymorphic nature of the presenting molecules, and the influence that this exerts on the peptide repertoire of each molecule, results in the phenomenon of *MHC restriction* of the T cell specificity for a given peptide. The binding of CD8 or CD4 molecules, respectively, to the class I or class II molecule also contributes to the interaction between T cell and the HLA-peptide complex, by providing for the selective activation of the appropriate T cell.

CLASS I STRUCTURE (Fig. 306-2*A*)

As noted above,MHCclass I molecules provide a cell-surface display of peptides derived from intracellular proteins; they also provide the signal for self-recognition byNKcells. Surface-expressed class I molecules consist of an MHC-encoded 44-kDa glycoprotein heavy chain, a non-MHC-encoded 12-kDa light chainb$_2$-microglobulin; and an antigenic peptide, typically 8 to 11 amino acids in length and derived from intracellularly produced protein. The heavy chain contains three domains, termeda$_1$,a$_2$, anda$_3$. Thea$_1$ anda$_2$domains form an "intrachain dimer," which together form a peptide-binding groove. In HLA-A and -B molecules, the groove is approximately 3 nm in length by 1-2 nm in maximum width (30 A ´ 12 A), whereas it is apparently somewhat wider in HLA-C. In cell surface-expressed class molecules, each domain contributes four of the eight strands of antiparallelb sheet, the membrane-distal side of which forms the floor of the groove, and one of the pair of a helices, the two coils of which form the walls of the groove (Fig. 306-2*A*). The membrane-anchoreda$_3$domain and noncovalently associatedb$_2$-microglobulin chain reside on the membrane-proximal side of the bsheet, each folded in the conformation of an immunoglobulin domain. The peptide is noncovalently bound in an extended conformation within the peptide-binding groove, with both N- and C-terminal ends anchored in pockets within the groove (A and F pockets, respectively) and, in many cases, with a prominent kink, or arch, approximately one-third of the way from the N-terminus that elevates the peptide main chain off the floor of the groove.

A remarkable property of peptide binding byMHCmolecules is the ability to form highly stable complexes with a wide array of peptide sequences. This is accomplished by a combination of peptide sequence-independent and -dependent bonding. The former consists of hydrogen bond and van der Waals interactions between conserved residues in the peptide-binding groove and charged or polar atoms along the peptide backbone. The latter are dependent upon the six side pockets that are formed by the irregular surface produced by protrusion of amino acid side chains from within the binding groove. The side chains lining the pockets interact with some of the peptide side chains. The sequence polymorphism among different class I alleles and isotypes predominantly affects the residues that line these pockets, and the interactions of these residues with peptide residues constitute the sequence-dependent bonding that confers a particular sequence "motif" on the range of peptides that can bind any given MHC molecule.

CLASS I BIOSYNTHESIS (Fig. 306-3*A*)

The biosynthesis of the classicalMHCclass I molecules reflects their role in presenting endogenous peptides. The heavy chain is cotranslationally inserted into the membrane of the endoplasmic reticulum (ER), where it becomes glycosylated and associates sequentially with the chaperone proteins calnexin and ERp57. It then forms a complex withb$_2$-microglobulin, and this complex associates with the chaperone calreticulin and the MHC-encoded molecule tapasin. Meanwhile, peptides generated within the cytosol from intracellular proteins by the multisubunit, multicatalytic proteasome complex are actively transported into the ER by MHC-encoded TAP (transporter associated with antigen processing) heterodimer. Following association with chaperones and trimming by peptidases within the ER, peptides bind to nascent class I molecules for which they have requisite affinity, to form complete, folded heavy chain-b$_2$-microglobulin-peptide trimer complexes. These are transported rapidly from the ER, through the *cis*- and *trans*-Golgi, where the N-linked oligosaccharide is further processed, and thence to the cell surface. Other proteins have been implicated in MHC class I assembly, e.g., the chaperones BiP and HSP70, but their roles in the pathway are not clear. The pathways for surface MHC class I degradation are poorly understood. A small proportion of heavy chains within properly assembled MHC class I molecules apparently become unfolded and subsequently degraded in the lysosomal pathway.

Most of the peptides transported by TAP are produced in the cytosol by proteolytic cleavage of intracellular proteins by the multisubunit, multicatalytic proteasome. Inhibitors of the proteasome dramatically reduce expression of class I-presented antigenic peptides, but other proteolytic systems may also generate peptides bound to class I. TheMHC-encoded proteasome subunits LMP2 and LMP7 may influence the spectrum of peptides produced, but they are not essential for proteasome function. Under certain circumstances, peptides derived from extracellular proteins in particulate form can become associated with class I molecules, but not necessarily by entering the class I pathway. Peptides are apparently bound to chaperones in the cytosol, including HSP90 and HSP70.

CLASS I FUNCTION

Peptide Antigen Presentation It is estimated that on any given cell, a given class I allele binds several hundred to several thousand distinct peptide species. The vast majority of these peptides are self peptides to which the host immune system is tolerant by one or more of the mechanisms that maintain tolerance, e.g., clonal deletion in the thymus or clonal anergy or clonal ignorance in the periphery (Chaps. 305 and307). However, class I molecules bearing foreign peptides expressed in a permissive immunologic context activate CD8 T cells, which, if naive, will then differentiate into cytolytic T lymphocytes (CTL). These T cells and their progeny, through their ab T cell receptors, are then capable of Fas/CD95- and/or perforin-mediated cytotoxicity and/or cytokine secretion (Chap. 305) upon further encounter with the class I-peptide combination that originally activated it, and also with other combinations of class I molecules plus peptide that present a similar immunochemical stimulus to theTCR. As alluded to above, this phenomenon by which T cells recognize foreign antigens in the context of specificMHCalleles is termed *MHC restriction*, and the specific MHC molecule is termed the *restriction element*. The most common source of foreign peptides presented by class I molecules is viral infection, in the course of which peptides from viral proteins enter the class I pathway. The generation of a strong CTL response that destroys virally infected cells represents an important antigen-specific defense against many viral infections (Chap. 305). In the case of some viral infections -- hepatitis B, for example -- CTL-induced target cell apoptosis is thought to be a more important mechanism of tissue damage than any direct cytopathic effect of the virus itself. The importance of the class I pathway in the defense against viral infection is underscored by the identification of a number of viral products that interfere with the normal class I biosynthetic pathway and thus block the immunogenetic expression of viral antigens.

Other examples of intracellularly generated peptides that can be presented by class I molecules in an immunogenic manner include peptides derived from nonviral intracellular infectious agents (e.g., *Listeria*, *Plasmodium*), tumor antigens, minor histocompatibility antigens, and presumably certain autoantigens. There are also situations in which cell surface-expressed class I molecules are thought to acquire and present exogenously derived peptides.

The role of class I HLA molecules in transplantation and in infectious and autoimmune diseases is discussed below.

NK Cell Recognition (See alsoChap. 305) NKcells, which play an important role in innate immune responses, are activated to cytotoxicity and cytokine secretion by contact with cells that lackMHCclass I expression, and NK cell activation is inhibited by cells that express MHC class I. In humans, the recognition of class I molecules by NK cells is carried out by two classes of receptor families, the killer cell-inhibitory cell receptor (KIR) family and the CD94/NKG2 family. The KIR family, encoded on chromosome 19q13.4, comprises glycoproteins of the immunoglobulin (Ig) superfamily that bind HLA class I molecules and inhibit NK cell-mediated cytotoxicity. An estimated 40 genes are divided into two subfamilies, KIR2D and KIR3D, which contain either two or three Ig domains, respectively. The KIR2D molecules primarily recognize alleles of HLA-C. The latter all possess either asparagine at position 77 and lysine at position 80, or serine at 77 and asparagine at 80 in thea$_1$domain of the heavy chain. Different members of the KIR2D family recognize the alternative forms of this polymorphism as well as other residues of the HLA-C heavy chain. The KIR3D molecules predominantly recognize HLA-B alleles. The latter carry a supertypic polymorphism defined serologically by two allotypes, HLA-Bw4 and -Bw6, that are determined by residues 77 to 83 in thea$_1$domain of the heavy chain. It is primarily alleles of the Bw4 supertype that bind KIR3D molecules. Although there is KIR recognition of some HLA-A and -Bw6 alleles, many of these alleles appear not to have a corresponding KIR ligand.

The second family of inhibitoryNKreceptors for HLA is encoded in the NK complex on chromosome 12p12.3-13.1 and consists of CD94 and four NKG2 genes: A, C, E, and D/F. These molecules are C-type (calcium-binding) lectins and are thought to exist as disulfide-bonded heterodimers between CD94 and the various NKG2 glycoproteins. CD94/NKG2A apparently binds to HLA-E and -G and several alleles of HLA-A, -B, and

-C. CD94/NKG2C binds primarily to HLA-E. The specificities of the other NKG2 molecules are not yet established._*The function of NK cells in immune responses is discussed in Chap. 305._

CLASS II STRUCTURE

A specialized functional architecture similar to that of the class I molecules can be seen in the example of a class II molecule depicted inFig. 306-2_B_, with an antigen-binding cleft arrayed above a supporting scaffold that extends the cleft toward the external cellular environment. However, in contrast to the HLA class I molecular structure,b_2-microglobulin is not associated with class II molecules. Rather, the class II molecule is a heterodimer, composed of a 29-kDa b chain and a 34-kDaa chain. The amino-terminal domains of each chain form the antigen-binding elements, which, like the class I molecule, cradle a bound peptide in a groove bounded by extended a-helical loops, one encoded by the A (a chain) gene and one by the B (b chain) gene. Like the class I groove, the class II antigen-binding groove is punctuated by pockets that contact the side chains of amino acid residues of the bound peptide; unlike the class I groove, it is open at both ends. Therefore, peptides bound by class II molecules vary greatly in length, since both the N- and C-terminal ends of the peptides can extend through the open ends of this groove. Approximately 11 amino acids within the bound peptide form intimate contacts with the class II molecule itself, with backbone hydrogen bonds and specific side chain interactions combining to provide stability and specificity, respectively, to the binding (Fig. 306-4).

The genetic polymorphisms that distinguish different class II genes correspond to changes in the amino acid composition of the class II molecule, and these variable sites are clustered predominantly around the pocket structures within the antigen-binding groove. As with class I, this is a critically important feature of the class II molecule, which explains how genetically different individuals have functionally different HLA molecules.

As noted above, the class I-peptide complex is preferentially recognized by CD8 T cells, and the class II-peptide complex is preferentially recognized by CD4 T cells. These interactions provide an important signal for activation of specific T cell lineages during antigen-recognition events. The CD8 recognition site is located on the$a_3$domain of theMHCclass I molecule, and the CD4 recognition site is located on the$b_2$domain of the class II molecule, in both cases remote from the peptide-binding site.

BIOSYNTHESIS AND FUNCTION OF CLASS II MOLECULES

The intracellular assembly of class II molecules occurs within a specialized compartmentalized pathway that differs dramatically from the class I pathway described above. As illustrated inFig. 306-3_B_, the class II molecule assembles in the ERin association with a chaperone molecule, known as the _invariant chain_. The invariant chain performs at least two roles. First, it binds to the class II molecule and blocks the peptide-binding groove, thus preventing antigenic peptides from binding. This role of the invariant chain appears to account for one of the important differences between class I and class IIMHCpathways, since it can explain why class I molecules present endogenous peptides from proteins newly synthesized in the ER, but class II molecules generally do not. Second, the invariant chain contains molecular localization signals that direct the class II molecule to traffic into post-Golgi compartments known as *endosomes*, which develop into specialized acidic compartments where proteases cleave the invariant chain, and antigenic peptides can now occupy the class II groove. It is at this stage in the intracellular pathway that the MHC-encoded DM molecule catalytically facilitates the exchange of peptides within the class II groove to help optimize the specificity and stability of the MHC-peptide complex.

Once thisMHC-peptide complex is deposited in the outer cell membrane, it becomes the target for T cell recognition via a specificTCRexpressed on lymphocytes. Because the endosome environment contains internalized proteins retrieved from the extracellular environment, the class II-peptide complex often contains bound antigens that were originally derived from extracellular proteins. In this way, the class II peptide loading pathway provides a mechanism for immune surveillance of the extracellular space. This appears to be an important feature that permits the class II molecule to bind foreign peptides, distinct from the endogenous pathway of class I-mediated presentation.

ROLE OF HLA IN TRANSPLANTATION

The development of modern clinical transplantation in the decades since the 1950s provided a major impetus for elucidation of the HLA system, as allograft survival is highest when donor and recipient are HLA-identical. Although many molecular events participate in transplantation rejection, allogeneic differences at class I and class II loci play a major role. Class I molecules can promote T cell responses in several different ways. In the cases of allografts in which the host and donor are mismatched at one or more class I loci, host T cells can be activated by classical *direct alloreactivity*, in which the antigen receptors on the host T cells react with the foreign class I molecule expressed on the allograft. In this situation, the response of any givenTCR may be dominated by the allogeneicMHCmolecule, the peptide bound to it, or some combination of the two. Another type of host antigraft T cell response involves the uptake and processing of donor MHC antigens by host antigen-presenting cells and the subsequent presentation of the resulting peptides by host MHC molecules. This mechanism is termed *indirect alloreactivity*, or *cross-priming*. It appears to play a quantitatively significant role in allograft rejection, although the molecular and cellular basis for the antigen processing remain to be completely elucidated. In the case of class I molecules on allografts that are shared by the host and the donor, a host T cell response may still be triggered because of peptides that are presented by the class I molecules of the graft but not of the host. The most common basis for the existence of these endogenous antigenic peptides, called *minor histocompatibility antigens*, is a genetic difference between donor and host at a non-MHC locus encoding the structural gene for the protein from which the peptide is derived. These loci are termed *minor histocompatibility loci*, and nonidentical individuals typically differ at many such loci, although only a few provide peptides for any given HLA allele. In recent years, the peptides and parent proteins for a number of human and experimental rodent minor histocompatibility antigens have been identified. In many of these cases of allograft rejection, T cell help for the generation of class I-restricted CD8 cells is provided by CD4 T cells reacting to analogous II differences. Moreover, class II differences alone are sufficient to drive allograft rejection.

ASSOCIATION WITH INFECTIOUS DISEASE

It has long been postulated that infectious agents provide the driving force for the allelic diversification seen in the HLA system. This has been difficult to confirm definitively, but one corollary of this hypothesis, namely, that it would be unusual to find HLA alleles strongly associated with susceptibility to any particular infectious disease, has generally been observed. Some modest associations of susceptibility to tuberculosis and leprosy have been found for several subtypes of HLA-DR2 (DRB1*15), and progression of HIV has been associated with several HLA haplotype including HLA-B35, HLA-CW*04, and HLA-A1-B8-DR3 in some studies (Chap. 309). With regard to resistance to infectious disease, the best documented example has been shown for malaria, in which B*5301, DRB1*1302, and DRB1*0101 have been shown to exert varying degrees of protection against severe disease. Slow progression of HIV has been associated with several HLA haplotypes (Chap. 309), and reduced persistence of hepatitis B and C viruses has been associated, respectively, with DRB1*1302 and with DR5.

A polymorphism in the promoter of theTNF-agene in the HLA class III region, which is associated with quantitative variation in the production of TNF, has recently been shown to have an association with the manifestations of a number of infectious diseases, including cerebral malaria, mucocutaneous leishmaniasis, lepromatous leprosy, scarring trachoma, persistent hepatitis B infection, and fatal meningicoccal meningitis.

ASSOCIATION OF HLA ALLELES WITH SUSCEPTIBILITY TO IMMUNOLOGICALLY MEDIATED DISEASES

Because of the immense polymorphism of HLA loci and strong linkage disequilibrium within the HLA region, it became possible, once a sufficient number of alleles had been defined by the early 1970s, to find associations of particular HLA alleles with certain disease states by comparing allele frequencies in patients with any particular disease and in control populations. A large number of such associations were identified during the 1970s. Most subsequent work in this field has been devoted to refining these associations to molecularly defined alleles and attempting to elucidate the contribution of HLA to disease pathogenesis.Table 306-1 lists the major diseases associated with HLA class I and class II genes. The strength of genetic association is reflected in the term *relative risk*, which is a statistical odds ratio representing the risk of disease in an individual carrying a particular genetic marker compared with the risk in individuals in that population without that marker. The nomenclature shown in Table 306-1 reflects both the HLA serotype (e.g., DR3, DR4) and the HLA genotype (e.g., DRB1*0301, DRB1*0401). Both because of the strong linkage disequilibrium within HLA and because the serologically identified loci represent only a small fraction of the total genes within the region, for many years it could not be established whether the associated alleles themselves participated in disease pathogenesis or were merely markers that were in linkage disequilibrium with the true disease allele. In recent years, it has become clear that it is very likely that the class I and class II alleles themselves are the true disease alleles for most of these associations. However, as discussed below, because of the extremely strong linkage disequilibrium between the DR and DQ loci, in some cases it has been difficult to determine the specific locus or combination of class II loci involved. At a minimum, different populations with different DR-DQ haplotypes need to be compared.

As might be predicted from the known function of the class I and class II gene products, almost all of the diseases associated with specific HLA alleles have an immunologic component to their pathogenesis. In some cases, as discussed below, specific protein and even peptide antigens have been implicated, but in no case is the molecular and cellular pathogenesis well understood. From a genetic point of view, strong HLA associations with disease (those associations with a relative risk of 10 or greater) are unusual because the implicated HLA alleles are normal, rather than defective, alleles. However, even in diseases with very strong HLA associations such as ankylosing spondylitis (AS) and type I diabetes mellitus, the non-HLA contribution exceeds 50% of the genetic predisposition, and the concordance of disease in monozygotic twins is considerably higher than in HLA-identical dizygotic twins or other sibling pairs. Genome-wide linkage analyses in these two diseases have found that the non-HLA genetic contribution comes from several other regions, although the linkage to HLA is by far the strongest.

Another group of diseases is genetically linked to HLA, not because of the immunologic function of HLA alleles, but rather because they are caused by autosomal dominant or recessive abnormal alleles at loci that happen to reside in or near the HLA region. Examples of these are 21-hydroxylase deficiency, hemochromatosis, and spinocerebellar ataxia (Chaps. 338,345, and364, respectively).

CLASS I DISEASE ASSOCIATIONS

Although the associations of human disease with particular HLA alleles or haplotypes predominantly involve the class II region, there are also several prominent disease associations with class I alleles. These include the association of Behcet's disease (Chap. 316) with HLA-B51, psoriasis vulgaris (Chap. 56) with HLA-Cw6, and, most prominently, the spondyloarthropathies (Chap. 315) with HLA-B27. The latter is among the strongest of all HLA associations with disease.

HLA-B27 was originally defined as a serologic determinant. It currently includes a family of 15 HLA-B locus alleles, designated HLA-B*2701-2715, as determined by nucleotide sequencing. HLA-B*2705 is the predominant subtype in Caucasians and most other non-Asian populations, and this subtype has been subjected to the most extensive investigation. All of the subtypes share a common B pocket in the peptide-binding groove, containing characteristic residues His9, Thr24, Glu45, and Cys67, and almost all share adjacent residues Ala69, Lys70, and Ala71. This deep, negatively charged pocket shows a strong preference for binding the arginine side chain, explaining the preference of the B27 binding group for peptides with Arg at P2 (peptide residue 2), as suggested by the crystal structure and confirmed by peptide isolation and sequencing. In addition, B27 is among the most negatively charged of HLA class I heavy chains, and the overall preference is for positively charged peptides. B27 is distinguished among class I alleles as a dominant restricting element forCTLrecognition of antigens from a wide variety of viruses, including HIV, and it is associated with prolonged survival in HIV infection (Chap. 309).

HLA-B27 and Disease HLA-B27 is very highly associated withAS(Chap. 315), both in its idiopathic form and in association with chronic inflammatory bowel disease or psoriasis vulgaris. It is also associated with reactive arthritis (ReA;Chap. 315), with other idiopathic forms of peripheral arthritis (undifferentiated spondyloarthropathy), and with recurrent acute anterior uveitis. B27 is found in 50 to 90% of individuals with these conditions, compared with a prevalence of ~7% in North American Caucasians. The prevalence of B27 in patients with idiopathic AS is 90%, and in AS complicated by iritis or aortic insufficiency is close to 100%. The absolute risk of spondyloarthropathy in unselected B27+ individuals has been variously estimated at 2 to 13% and >20% if a B27+ first-degree relative is affected. The concordance rate of AS in identical twins is very high, approximately 75%. It can be concluded that the B27 molecule itself is involved in disease pathogenesis, based on strong evidence from clinical epidemiology and on the occurrence of a spondyloarthropathy-like disease in HLA-B27 transgenic rats. A well-established association with both AS and ReA exists for subtypes B*2702, -04, and -05, and anecdotal association has been reported for subtypes B*2701, -03, -07, -08, -10, and -11. The propensity of the B27 molecule to induce disease thus presumably derives from one or more unique features of its structure that are shared by several B27 subtypes. It remains a central unanswered question whether the pathogenesis of B27-associated disease derives from the specificity of a particular peptide or family of peptides bound to B27 or whether another mechanism is involved that is independent of the peptide specificity of B27. The first alternative can be further subdivided into mechanisms that involve T cell recognition of B27-peptide complexes, and those that do not. A variety of other roles for B27 in disease pathogenesis have been postulated, including molecular or antigenic mimicry between B27 and certain bacteria and reduced killing of intracellular bacteria in cells expressing B27. However, the most straightforward possibility is the presentation of peptides to CD8 T cells in a way that somehow promotes joint inflammation, i.e., the "arthritogenic peptide" hypothesis. This concept has been supported in ReA by the finding of CD8-restricted antigen-specific T cells, particularly in *Yersinia*-induced ReA. It is of particular interest that the HSP60 molecule from *Y. enterocolitica* has been shown to give rise to dominant antigens recognized by both synovial B27-restricted CD8 and class II-restricted CD4 T cells. This suggests a T cell pathogenesis involving intramolecular help and/or epitope spreading in which a B27-restricted response could well be primary.

In contrast toReA, there is little direct evidence regarding the molecular role of HLA-B27 in ankylosing spondylitis. However, correlations between disease susceptibility and the peptide-binding specificity of the B27 subtypes have been found that support the "arthritogenic peptide" hypothesis inAS. Specifically, a lack of disease susceptibility has been documented for the subtypes B*2706, found mainly amongst Southeast Asians, and B*2709, found mainly amongst Sardinians. B*2709 differs from B*2705 only at residue 116, carrying His instead of Asp. B*2706 differs from B*2704 at 114 and 116, carrying Asp and Tyr instead of His and Asp. These residues, which lie in the floor of the peptide-binding groove, interact with the C-terminal end of the bound peptide. Unlike the disease-associated subtypes, B*2706 and B*2709 have been found not to carry peptides with C-terminal Tyr. This has led to the hypothesis of a disease-prone B27-bound peptide with C-terminal Tyr.

CLASS II DISEASE ASSOCIATIONS

The majority of associations between HLA and disease are with class II alleles (Table 306-1). Several diseases have complex HLA genetic associations.

Celiac Disease In the case of celiac disease (Chap. 286), it is probable that the HLA-DQ genes are the primary basis for the disease association. HLA-DQ genes present on both the celiac-associated DR3 and DR7 haplotypes include the DQB1*0201 gene, and further detailed studies have documented a specific class IIab dimer encoded by the DQA1*0501 and DQB1*0201 genes, which appears to account for the HLA genetic contribution to celiac disease susceptibility. This specific HLA association with celiac disease may have a straightforward explanation: peptides derived from the wheat gluten component gliaden are bound to the molecule encoded by DQA1*0501 and DQB1*0201 and presented to T cells. A gliaden-derived peptide that has been implicated in this immune activation binds the DQ class II dimer best when the peptide contains a glutamine to glutamic acid substitution. It has been proposed that tissue transglutaminase, an enzyme present at increased levels in the intestinal cells of celiac patients, converts glutamine to glutamic acid in gliadin, creating peptides that are capable of being bound by the DQ2 molecule and presented to T cells.

Pemphigus Vulgaris In the case of pemphigus vulgaris (Chap. 58), there are two HLA haplotypes associated with disease: DRB1*0402-DQB1*0302 and DRB1*1401-DQB1*0503. Peptides derived from epidermal autoantigens have been implicated that preferentially bind to the DRB1*0402-encoded molecule, suggesting that specific peptide binding by this disease-associated class II molecule is important in disease. However, there are no class II genes in common between the disease-associated DR4 and DR14 haplotypes, and there is no evidence for any interaction of the latter haplotype interacting with the epidermal peptides that bind the DRB1*0402-encoded molecule. Thus, the most likely interpretation is that each of these class II associations with pemphigus represents a different pathway to a comparable clinical outcome.

Juvenile Arthritis Pauciarticular juvenile arthritis (Chap. 312) is an autoimmune disease associated with genes at the DRB1 locus and also with genes at the DPB1 locus. Patients with both DPB1*0201 and a DRB1 susceptibility allele (usually DRB1*08 or -*05) have a higher relative risk than expected from the additive effect of those genes alone. In juvenile patients with rheumatoid factor-positive polyarticular disease, heterozygotes carrying both DRB1*0401 and -*0404 have a relative risk>100, reflecting an apparent synergy in individuals inheriting both of these susceptibility genes.

Type 1 Diabetes Mellitus There are several aspects of the genetics of type 1 diabetes (Chap. 333) that illustrate the complex nature of HLA associations with autoimmune diseases. First, type 1 (autoimmune) diabetes mellitus is associated with both DR3 and DR4 serotypes and their corresponding genes. The presence of both the DR3 and DR4 haplotypes in one individual confers the highest known genetic risk for type diabetes, and individuals carrying either of these haplotypes also carry some increased risk. Specific class II genes on each haplotype have been thoroughly studied, and the strongest association is with DQB1*0302, a specific gene on the diabetes-associated DR4 haplotypes. Thus, all DR4 haplotypes that carry a DQB1*0302 gene are associated with type 1 diabetes, whereas related DR4 haplotypes that carry a different DQB1 gene are not. The primary class II determinant of susceptibility, therefore, is HLA-DQB1*0302. However, the relative risk associated with inheritance of this gene can be modified, depending on other HLA genes present either on the same or a second haplotype. For example, just as the presence of a second haplotype containing DR3 is associated with an increased risk of diabetes, the presence of a DR2-positive haplotype containing a DQB1*0602 gene is associated with decreased risk. This gene, DQB1*0602, is considered "protective" for type 1 diabetes. Even some DRB1 genes that can occur on the same haplotype as DQB1*0302 may modulate risk, so that individuals with the DR4 haplotype that contains DRB1*0403 are less susceptible to type 1 diabetes than individuals with other DR4-DQB1*0302 haplotypes.

Although the presence of a DR3 haplotype in combination with the DR4-DQB1*0302 haplotype is a very high risk combination for diabetes susceptibility, the specific gene on the DR3 haplotype that is responsible for this synergy has not yet been identified. This is because the predominant HLA-DR3 haplotype in Caucasians has very tight linkage with other genes within theMHC, including HLA-A1, -B8, -Cw7, and -C4A, as discussed above. Thus, any of a large variety of genes within the HLA region on this DR3 haplotype may be the primary gene(s) responsible for contributing to diabetes susceptibility. An example that more directly implicates other genes linked to DR3 is the association between HLA genes and systemic lupus erythematosus (SLE;Chap. 311). The C4A null alleles that are present on the HLA-DR3 haplotypes in SLE are also often present in patients without DR3, notably those with HLA-DR2. This implies the presence of a C4A silent allele, which is a defective structural gene for the C4 complement component, rather than the expression of any particular class II gene, as a potential susceptibility gene within HLA associated with SLE.

HLA and Rheumatoid Arthritis The HLA genes most highly associated with rheumatoid arthritis (RA) are DRB1*0401 and DRB1*0404 (Chap. 312). These genes encode a distinctive sequence of amino acids from codons 67 to 74 of the DRb molecule: RA-associated class II molecules carry the sequence LeuLeuGluGlnArgArgAlaAla or LeuLeuGluGlnLysArgAlaAla in this region, while non-RA-associated genes carry one or more differences in this region. These residues form a portion of the molecule that lies in the middle of the a-helical portion of the DRB1-encoded class II molecule, termed the *shared epitope*.

These DR4+RA-associated alleles are most frequent among patients with more severe, erosive disease. The frequency of these DR4+ alleles is lower among patients with RA who are rheumatoid factor-negative and those with nonerosive forms of the disease. Although the frequency of these DRB1 susceptibility alleles in RA patients is high, the same genes are also prevalent in the unaffected population, and thus the absolute risk associated with these susceptibility alleles is low. The highest risk for susceptibility to RA comes in individuals who carry both a DRB1*0401 and DRB1*0404 gene. Some forms of RA are associated with other HLA genes, such as DRB1*01, -*1001, and -*1402, which also carry the shared epitope sequence, strongly suggesting that this part of the class II molecule contributes directly to disease pathogenesis.

MOLECULAR MECHANISMS FOR HLA DISEASE ASSOCIATIONS

As noted above, HLA molecules play a key role in the selection and establishment of the antigen-specific T cell repertoire and a major role in the subsequent activation of those T cells during the initiation of an immune response. Precise genetic polymorphisms characteristic of individual alleles dictate the specificity of these interactions and thereby instruct and guide antigen-specific immune events. These same genetically determined pathways are therefore implicated in disease pathogenesis when specific HLA genes are responsible for autoimmune disease susceptibility.

The fate of developing T cells within the thymus is determined by the affinity of interaction betweenTCR and HLA molecules bearing self peptides; thus, the particular HLA types of each individual control the precise specificity of the T cell repertoire (Chap. 305). The primary basis for HLA-associated disease susceptibility may well lie within this thymic maturation pathway. The positive selection of potentially autoreactive T cells, based on the presence of specific HLA susceptibility genes, may establish the threshold for disease risk in a particular individual.

At the time of onset of a subsequent immune response, the primary role of the HLA molecule is to bind peptide and present it to antigen-specific T cells. The HLA complex can therefore be viewed as encoding genetic determinants of precise immunologic activation events. Antigenic peptides that bind particular HLA molecules are capable of stimulating T cell immune responses; peptides that do not bind are not presented to T cells and are not immunogenic. This genetic control of the immune response is mediated by the polymorphic sites within the HLA antigen-binding groove that interact with the bound peptides. In autoimmune and immune-mediated diseases, it is likely that specific tissue antigens that are targets for pathogenic lymphocytes are complexed with the HLA molecules encoded by specific susceptibility alleles. In autoimmune diseases with an infectious etiology, it is likely that immune responses to peptides derived from the initiating pathogen are bound and presented by particular HLA molecules to activate T lymphocytes that play a triggering or contributory role in disease pathogenesis. The concept that early events in disease initiation are triggered by specific HLA-peptide complexes offers some prospects for therapeutic intervention, since it may be possible to design compounds that interfere with the formation or function of specific HLA-peptide-TCRinteractions.

When considering mechanisms of HLA associations with the immune response and with disease it is well to remember that just as HLA genetics are complex, so are the mechanisms likely to be heterogeneous. Immune-mediated disease is a multistep process in which one of the HLA-associated functions is to establish a repertoire of potentially reactive T cells, while another HLA-associated function is to provide the essential peptide-binding specificity for T cell recognition. For diseases with multiple HLA genetic associations, it is possible that both of these interactions occur and synergize to advance an accelerated pathway of disease.

(Bibliography omitted in Palm version)

Back to Table of Contents

307. AUTOIMMUNITY AND AUTOIMMUNE DISEASES - *Peter E. Lipsky*, *Betty Diamond*

One of the classically accepted features of the immune system is the capacity to distinguish self from non-self. Although they are able to recognize and generate reactions to a vast array of foreign materials, most animals do not mount immune responses to self-antigens under ordinary circumstances and thus are tolerant to self. Although recognition of self plays an important role in generating both the T cell and B cell repertoires of immune receptors and plays an essential role in the recognition of nominal antigen by T cells, the development of potentially harmful immune responses to self-antigens is, in general, precluded. Autoimmunity, therefore, represents the end result of the breakdown of one or more of the basic mechanisms regulating immune tolerance.

The presence or absence of pathologic consequences resulting from self-reactivity determines whether autoimmunity leads to the development of an autoimmune disease. The essential feature of an autoimmune disease is that tissue injury is caused by the immunologic reaction of the organism with its own tissues. Autoimmunity, on the other hand, refers merely to the presence of antibodies or T lymphocytes that react with self-antigens and does not necessarily imply that the development of self-reactivity has pathogenic consequences.

Autoimmunity may occur as an isolated event or in the setting of specific clinical syndromes. Autoimmunity may be seen in normal individuals and in higher frequency in normal older people. In addition, autoreactivity may develop during various infectious conditions. The expression of autoimmunity may be self-limited, as occurs with many infectious processes, or persistent. In both circumstances there is a tendency to develop autoreactivity directed against a variety of different tissues or organs. As mentioned above, autoimmunity does not necessarily lead to tissue damage, and even in the presence of organ pathology, it may be difficult to determine whether the damage is mediated by autoreactivity. Thus, the presence of self-reactivity may be either the cause or a consequence of an ongoing pathologic process. Furthermore, when autoreactivity is induced by an inciting event, such as infection or tissue damage from trauma or infarction, there may or may not be ensuing pathology.

MECHANISMS OF AUTOIMMUNITY

Since Ehrlich first postulated the existence of mechanisms to prevent the generation of self-reactivity in 1900, ideas concerning the nature of this inhibition have developed in parallel with the progressive increase in understanding of the immune system. Burnet's clonal selection theory included the idea that interaction of lymphoid cells with their specific antigens during fetal or early postnatal life would lead to elimination of such "forbidden clones." This idea became untenable, however, when it was shown by a number of investigators that autoimmune diseases could be induced by simple immunization procedures, that autoantigen-binding cells could be demonstrated easily in the circulation of normal individuals, and that self-limited autoimmune phenomena frequently developed during infections. These observations indicated that clones of cells capable of responding to autoantigens were present in the repertoire of antigen-reactive cells in normal adults and suggested that mechanisms in addition to clonal deletion were responsible for preventing their activation.

Currently, three general processes are thought to be involved in the maintenance of selective unresponsiveness to autoantigens (Table 307-1): (1) sequestration of self-antigens, rendering them inaccessible to the immune system; (2) specific unresponsiveness (tolerance or anergy) of relevant T or B cells; and (3) limitation of potential reactivity by regulatory mechanisms. These mechanisms permit the host to respond to the vast universe of foreign antigens but preclude responses to autoantigens that might have pathogenic consequences.

Derangements of these normal processes may predispose to the development of autoimmunity (Table 307-2). In general, these abnormal responses relate to stimulation by exogenous agents, usually bacterial or viral, or endogenous abnormalities in the cells of the immune system. Thus, autoreactivity can result from exogenous stimulation of the immune system in a manner that overcomes the regulated unresponsiveness to self-antigens. Microbial superantigens, such as staphylococcal protein A and staphylococcal enterotoxins, are substances that can stimulate a broad range of T and B cells based upon specific interactions with selected families of immune receptors irrespective of their antigen specificity. If autoantigen reactive T and/or B cells express these receptors, autoimmunity might develop. Alternatively, molecular mimicry or cross-reactivity between a microbial product and a self-antigen might lead to activation of autoreactive lymphocytes. One of the best examples of autoreactivity and autoimmune disease resulting from molecular mimicry is rheumatic fever, in which antibodies to the M protein of streptococci cross-react with myosin, laminin, and other matrix proteins. Deposition of these autoantibodies in the heart initiates an inflammatory response. Molecular mimicry between microbial proteins and host tissues has been reported in insulin-dependent diabetes mellitus (IDDM), rheumatoid arthritis, and multiple sclerosis. The capacity of nonspecific stimulation of the immune system to predispose to the development of autoimmunity has been explored in a number of models; one is provided by the effect of adjuvants on the production of autoimmunity. Autoantigens become much more immunogenic when administered with adjuvant. It is presumed that infectious agents may be able to overcome self-tolerance because they possess molecules, such as bacterial endotoxin, that have adjuvant-like effects on the immune system.

Endogenous derangements of the immune system may also contribute to the loss of immunologic tolerance, or anergy, to self-antigens and the development of autoimmunity (Table 307-2). Many autoantigens reside in immunologically privileged sites, such as the brain or the anterior chamber of the eye. These sites are characterized by the inability of engrafted tissue to elicit immune responses. Immunologic privilege results from a number of events, including the limited entry of proteins from those sites into lymphatics, the local production of immunosuppressive cytokines such as transforming growth factor (TGF) b, and the local expression of molecules such as Fas ligand that can induce apoptosis of activated T cells. Lymphoid cells remain in a state of immunologic ignorance (neither activated nor anergized) to proteins expressed uniquely in immunologically privileged sites. If the privileged site is damaged by trauma or inflammation, or if T cells are activated elsewhere, proteins expressed at this site can become the targets of immunologic assault. Such an event may occur in multiple sclerosis and sympathetic ophthalmia, in which antigens uniquely expressed in the brain and eye, respectively, become the target of activated T cells.

Alterations in antigen presentation may also contribute to autoimmunity. This may occur by epitope spreading, in which protein determinants (*epitopes*) not routinely seen by lymphocytes (*cryptic epitopes*) are recognized as a result of immunologic reactivity to associated molecules. For example, animals immunized with one protein component of the spliceosome may be induced to produce antibodies to multiple other spliceosome proteins. Finally, inflammation, drug exposure, or normal senescence may cause a primary chemical alteration in proteins, resulting in the generation of immune responses that cross-react with normal self-proteins. Alterations in the availability and presentation of autoantigens may be important components of immunoreactivity in certain models of organ-specific autoimmune diseases. In addition, these factors may be relevant in understanding the pathogenesis of various drug-induced autoimmune conditions. However, the diversity of autoreactivity manifest in non-organ-specific systemic autoimmune diseases suggests that these conditions might result from a more general activation of the immune system rather than from an alteration in individual self-antigens.

A number of experimental models have suggested that intense stimulation of T lymphocytes can produce nonspecific signals that bypass the need for antigen-specific helper T cells and lead to polyclonal B cell activation with the formation of multiple autoantibodies. For example, antinuclear, antierythrocyte, and antilymphocyte antibodies are produced during the chronic graft-versus-host reaction. In addition, true autoimmune diseases, including autoimmune hemolytic anemia and immune complex-mediated glomerulonephritis, can also be induced in this manner. While it is clear that such diffuse activation of helper T cell activity can cause autoimmunity, nonspecific stimulation of B lymphocytes can also lead to the production of autoantibodies. Thus, the administration of polyclonal B cell activators, such as bacterial endotoxin, to normal mice leads to the production of a number of autoantibodies, including those directed to DNA and IgG (rheumatoid factor).

Primary alterations in the activity of T and/or B cells, cytokine imbalances, or defective immunoregulatory circuits may also contribute to the emergence of autoimmunity. Although the biochemical bases of many of these abnormalities have not been documented, they may contribute to the emergence of autoimmunity either alone or in concert. For example, decreased apoptosis, as can be seen in animals with defects in Fas (CD95) or Fas ligand or in patients with related abnormalities, can be associated with the development of autoimmunity. Similarly, diminished production of tumor necrosis factor (TNF) a and interleukin (IL)10 has been reported to be associated with the development of autoimmunity.

An alternative explanation for the development of autoimmunity is that self-reactivity results not from overstimulation of the immune system but rather from an abnormality of immunoregulatory mechanisms. Observations made in both human autoimmune disease and animal models suggest that defects in the generation and expression of regulatory T cell activity may allow for the production of autoantibodies. The importance of defects in immunoregulatory cells is confirmed by the finding that administration of normal suppressor T cells or factors derived from them can prevent the development of autoimmune disease in rodent models of autoimmunity.

One of the mechanisms that regulates normal humoral immune responses is the production of anti-idiotype antibodies. These are immunoglobulin molecules directed against antigen-binding determinants of the specific antibodies originally elicited by the immunogen. Production of anti-idiotype antibodies may be dependent on helper T cell activity even when the initial immunogen is T cell independent. Therefore, it is possible that abnormalities in the generation of appropriate anti-idiotype antibodies, either at the B or T cell level, are responsible for the development of autoimmunity in certain circumstances.

It should be apparent that no single mechanism can explain all the varied manifestations of autoimmunity. Indeed, it appears likely, especially in systemic autoimmune diseases, that a number of abnormalities may converge to induce the complete syndrome. Moreover, one abnormality may cause a second, which, in concert with the first, facilitates the expression of autoimmunity. This possibility is consistent with recent findings in murine models ofIDDM; systemic lupus erythematosus (SLE), rheumatoid arthritis, and multiple sclerosis in which multiple genetic regions, many of which are involved in controlling immune reactivity, appear to contribute to the development of autoimmune disease.

Despite the plethora of immunologic derangements identified in systemic autoimmune diseases such asSLE, the primary abnormality causing the disease remains unclear. In fact, detailed examination of a number of murine strains that spontaneously develop a lupus-like syndrome has failed to demonstrate a common immunologic abnormality. Additional factors that appear to be important determinants in the induction of autoimmunity include age, sex, genetic background, exposure to infectious agents, and environmental contacts. How all of these disparate factors affect the capacity to develop self-reactivity is currently being intensively investigated.

GENETIC CONSIDERATIONS

Evidence in humans that there are susceptibility genes for autoimmunity comes from family studies and especially from studies of twins. Studies inIDDM, rheumatoid arthritis, multiple sclerosis, andSLE have shown that approximately 15 to 30% of pairs of monozygotic twins show disease concordance, compared with <5% of dizygotic twins. The occurrence of different autoimmune diseases within the same family has suggested that certain susceptibility genes may predispose to a variety of autoimmune diseases. In addition to this evidence from humans, certain inbred mouse strains reproducibly develop specific spontaneous or experimentally induced autoimmune diseases, whereas others do not. These findings have led to an extensive search for genes that determine susceptibility to autoimmune disease.

The most consistent association for susceptibility to autoimmune disease has been with the major histocompatibility complex (MHC). Many human autoimmune diseases are associated with particular HLA alleles (Chap. 306). It has been suggested that the association of MHC genotype with autoimmune disease relates to differences in the ability of different allelic variations of MHC molecules to present autoantigenic peptides to autoreactive T cells. An alternative hypothesis involves the role of MHC alleles in shaping the T cell receptor repertoire during T cell ontongeny in the thymus.

Additionally, specific MHC gene products themselves may be the source of peptides that can be recognized by T cells. Cross-reactivity between such MHC peptides and peptides derived from proteins produced by common microbes may trigger autoimmunity by molecular mimicry. However, MHC genotype alone does not determine the development of autoimmunity. Identical twins are far more likely to develop the same autoimmune disease than MHC-identical nontwin siblings, suggesting that genetic factors other than the MHC also affect disease susceptibility. Recent studies of the genetics ofIDDM,SLE, and multiple sclerosis in humans and mice have shown that there are several independently segregating disease susceptibility loci in addition to the MHC.

There is evidence that several other genes are important in increasing susceptibility to autoimmune disease. In humans, inherited homozygous deficiency of the early proteins of the classic pathway of complement (C1, C4, or C2) is very strongly associated with the development ofSLE. In mice and humans, abnormalities in the genes encoding proteins involved in the regulation of apoptosis, including Fas (CD95) and Fas ligand (CD95 ligand), are strongly associated with the development of autoimmunity. There is also evidence that inherited variation in the level of expression of certain cytokines, such asTNF-aorIL-10, may also increase susceptibility to autoimmune disease.

A further important factor in disease susceptibility is the hormonal status of the patient. Many autoimmune diseases show a strong sex bias, which appears in most cases to relate to the hormonal status of women.

IMMUNOPATHOGENIC MECHANISMS IN AUTOIMMUNE DISEASES

The mechanisms of tissue injury in autoimmune diseases can be divided into antibody-mediated and cell-mediated processes. Representative examples are listed inTable 307-3.

The pathogenicity of autoantibodies can be mediated through several mechanisms, including opsonization of soluble factors or cells, activation of an inflammatory cascade via the complement system, and interference with the physiologic function of soluble molecules or cells.

In autoimmune thrombocytopenic purpura, opsonization of platelets targets them for elimination by phagocytes. Likewise, in autoimmune hemolytic anemia, binding of immunoglobulin to red cell membranes leads to phagocytosis and lysis of the opsonized cell. Goodpastures' syndrome, a disease characterized by lung hemorrhage and severe glomerulonephritis, represents an example of antibody binding leading to local activation of complement and neutrophil accumulation and activation. The autoantibody in this disease binds to thea_3 chain of type IV collagen in the basement membrane. InSLE, activation of the complement cascade at sites of immunoglobulin deposition in renal glomeruli is considered to be a major mechanism of renal damage.

Autoantibodies can also interfere with normal physiologic functions of cells or soluble factors. Autoantibodies against hormone receptors can lead to stimulation of cells or to inhibition of cell function through interference with receptor signaling. For example, long-acting thyroid stimulators (LATS), which are autoantibodies that bind to the receptor for thyroid-stimulating hormone (TSH), are present in Graves' disease and function as agonists, causing the thyroid to respond as if there were an excess of TSH. Alternatively, antibodies to the insulin receptor can cause insulin-resistant diabetes mellitus through receptor blockade. In myasthenia gravis, autoantibodies to the acetylcholine receptor can be detected in 85 to 90% of patients and are responsible for muscle weakness. The exact location of the antigenic epitope, the valence and affinity of the antibody, and perhaps other characteristics determine whether activation or blockade results from antibody binding.

Antiphospholipid antibodies are associated with thromboembolic events in primary and secondary antiphospholipid syndrome and have also been associated with fetal wastage. The major antibody is directed to the phospholipid-b_2-glycoprotein I complex and appears to exert a procoagulant effect. In pemphigus vulgaris, autoantibodies bind to a component of the epidermal cell desmosome, desmoglein 3, and play a role in the induction of the disease. They exert their pathologic effect by disrupting cell-cell junctions through stimulation of the production of epithelial proteases, leading to blister formation. Cytoplasmic antineutrophil cytoplasmic antibody (c-ANCA), found in Wegener's granulomatosis, is an antibody to an intracellular antigen, the 29-kDa serine protease (proteinase-3). In vitro experiments have shown that IgG c-ANCA causes cellular activation and degranulation of primed neutrophils.

It is important to note that autoantibodies of a given specificity may cause disease only in genetically susceptible hosts, as has been shown in experimental models of myasthenia gravis. Finally, some autoantibodies seem to be markers for disease but have as yet no known pathogenic potential.

AUTOIMMUNE DISEASE

Manifestations of autoimmunity are found in a large number of pathologic conditions. However, their presence does not necessarily imply that the pathologic process is an autoimmune disease. A number of attempts to establish formal criteria for the diagnosis of autoimmune diseases have been made, but none is universally accepted. One set of criteria is shown in Table 307-4; however, this should be viewed merely as a guide in consideration of the problem.

To classify a disease as autoimmune, it is necessary to demonstrate that the immune response to a self-antigen causes the observed pathology. Initially, the demonstration that antibodies against the affected tissue could be detected in the serum of patients suffering from various diseases was taken as evidence that these diseases had an autoimmune basis. However, such autoantibodies are also found when tissue damage is caused by trauma or infection, and the autoantibody is secondary to tissue damage. Thus, it is necessary to show that autoimmunity is pathogenic before classifying a disease as autoimmune.

If the autoantibodies are pathogenic, it may be possible to transfer disease to experimental animals by the administration of autoantibodies, with the subsequent development of pathology in the recipient similar to that seen in the patient from whom the antibodies were obtained. This has been shown, for example, in Graves' disease. Some autoimmune diseases can be transferred from mother to fetus and are observed in the newborn babies of diseased mothers. The symptoms of the disease in the newborn usually disappear as the levels of the maternal antibody decrease. An exception is congenital heart block, in which damage to the developing conducting system of the heart as a result of transfer of anti-Ro antibody from the mother results in permanent heart block.

In most situations, the critical factors that determine when the development of autoimmunity results in autoimmune disease have not been delineated. The relationship of autoimmunity to the development of autoimmune disease may relate to the fine specificity of the antibodies or T cells or their specific effector capabilities. In many circumstances a mechanistic understanding of the pathogenic potential of autoantibodies has not been established. In some autoimmune diseases, biased production of cytokines by helper T (T_H) cells may play a role in pathogenesis. In this regard, T cells can differentiate into specialized effector cells that predominantly produce interferong(T_H1) orIL-4 (T_H2) (Chap. 305). The former facilitate macrophage activation and classic cell-mediated immunity, whereas the latter are thought to have regulatory functions and are involved in the resolution of normal immune responses and also the development of responses to a variety of parasites. In a number of autoimmune diseases, such as rheumatoid arthritis, multiple sclerosis,IDDM, and Crohn's disease, there appears to be biased differentiation of T_H1 cells, with resultant organ damage.

ORGAN-SPECIFIC VERSUS SYSTEMIC AUTOIMMUNE DISEASES

Autoimmune diseases form a spectrum, from those specifically affecting a single organ to systemic disorders with involvement of many organs (Table 307-5). Hashimoto's autoimmune thyroiditis is probably the best studied example of an organ-specific autoimmune disease (Chap. 330). In this disorder, there is a specific lesion in the thyroid associated with infiltration of mononuclear cells and damage to follicular cells. Antibody to thyroid constituents can be demonstrated in nearly all cases. Other organ- or tissue-specific autoimmune disorders include pemphigus vulgaris, autoimmune hemolytic anemia, idiopathic thrombocytopenic purpura, Goodpasture's syndrome, myasthenia gravis, and sympathetic ophthalmia. One important feature of some organ-specific autoimmune diseases is the tendency for overlap, such that an individual with one specific syndrome is more likely to develop a second syndrome. For example, there is a high incidence of pernicious anemia in individuals with autoimmune thyroiditis. More striking is the tendency for individuals with an organ-specific autoimmune disease to develop multiple other manifestations of autoimmunity without the development of associated organ pathology. Thus, as many as 50% of individuals with pernicious anemia have non-cross-reacting antibodies to thyroid constituents, whereas patients with myasthenia gravis may develop antinuclear antibodies, antithyroid antibodies, rheumatoid factor, antilymphocyte antibodies, and polyclonal hypergammaglobulinemia. Part of the explanation for this may relate to the genetic elements shared by individuals with these different diseases.

SYSTEMIC AUTOIMMUNE DISEASE

Systemic autoimmune diseases differ from organ-specific diseases in that pathologic lesions are found in multiple, diverse organs and tissues. The hallmark of these conditions is the demonstration of associated relevant autoimmune manifestations that are likely to be etiologic in the organ pathology.SLE is the best example of such a disorder. Although a number of other diseases such as Sjogren's syndrome possess certain of the features of a systemic autoimmune disease, SLE represents the prototype of these disorders because of its abundance of autoimmune manifestations.

SLEis a disease of protean manifestations that characteristically involves the kidneys, joints, skin, serosal surfaces, blood vessels, and central nervous system (Chap. 311). The disease is associated with a vast array of autoantibodies whose production appears to be a part of a generalized hyperreactivity of the humoral immune system. Other features of SLE include generalized B cell hyperresponsiveness, polyclonal hypergammaglobulinemia, and increased titers of antibodies to commonly encountered viral antigens.

A number of the autoantibodies found in SLEhave clearly been implicated in certain of the pathologic features of the disease. Classically, SLE has been considered a disorder in which immune complexes are the major pathogenic entity. While immune-complex deposition or in situ formation of complexes appears to be a major pathogenic mechanism in lupus renal disease, additional autoimmune processes may be implicated in the pathogenesis of other features of the disease (Table 307-6).

The etiology ofSLEremains unknown, and the interplay of a number of factors appears to be involved in its pathogenesis. Race and gender play an important role as evidenced by the increased incidence in young black females. The role of environmental factors is suggested by the high incidence of autoantibodies, especially antilymphocyte antibodies, in nonconsanguineous household contacts of individuals with SLE. The importance of genetic influences is suggested by family studies indicating that first-degree relatives of SLE patients have an increased likelihood of developing autoimmunity and autoimmune disease. The very high concordance rate for SLE and even higher rate for autoimmunity in monozygotic twins supports this concept. Finally, the association of SLE withMHCgenes confirms the importance of immunogenetic factors in its pathogenesis. Current hypotheses concerning the immunopathogenesis of SLE suggest that autoantibody formation may result from a combination of exaggerated B cell activation owing either to excessive exogenous stimulation or endogenous hyperactivity and inadequate regulatory T cell or anti-idiotype regulation. Genetic elements may contribute to each of these abnormalities.

(Bibliography omitted in Palm version)

Back to Table of Contents

308. PRIMARY IMMUNE DEFICIENCY DISEASES - *Max D. Cooper*, *Harry W. Schroeder, Jr.*

Immunologic functions are mediated by developmentally independent, but functionally interacting, families of lymphocytes. The activities of B and T lymphocytes and their products in host defense are closely integrated with the functions of other cells of the reticuloendothelial system. Dendritic cells, Langerhans' cells in the skin, and macrophages play an important role in the trapping and presentation of antigens to T and B cells to initiate the immune response. Macrophages also become effector cells, especially when activated by cytokine products of lymphocytes. The scavenger activity of macrophages and polymorphonuclear leukocytes is directed and made specific by antibodies in concert with cytokines and the complement system. Natural killer (NK) cells, a population of granular lymphocytes with receptors specific for MHC class I molecules, may spontaneously kill tumor and virus-infected cells, activities that are enhanced by the cytokine products of immune and inflammatory cells. Killing by NK cells can also be targeted by IgG antibodies for which NK cells have cell-surface receptors. The interaction of basophils and tissue mast cells with IgE antibodies in causation of immediate-type hypersensitivity is discussed in Chap. 310. Consideration of these interrelationships is an important part of the analysis of patients with suspected immune deficiency.

DIFFERENTIATION OF T AND B CELLS

The functional deficits that occur in both congenital and acquired immunodeficiencies are usefully viewed as being caused by defects at various points along the differentiation pathways of immunocompetent cells. A subpopulation of hematopoietic stem cells become restricted to lymphoid differentiation prior to migration to the thymus, where T cells are generated, and in the fetal liver and adult bone marrow, where B cell development begins (Fig. 308-1). Immature T and B cells then migrate through the circulation to the spleen, lymph nodes, intestine and other peripheral lymphoid organs. In these sites, they may encounter antigens presented by dendritic cells or macrophages and respond with proliferation, differentiation and mediation of immune responses.Chap. 305 provides a general account of their roles in cellular and humoral immunity.

Differentiation of T or B cells may be arrested at either the primary or secondary stages. Reflecting the complex cellular interactions involved in immune responses and the pivotal role played by T lymphocytes, immune deficiencies primarily involving T cells are usually also associated with abnormal B cell function. Conversely, immunodeficiencies manifested primarily by inability to produce antibodies may be caused by T cell defects not associated with abnormal cell-mediated immunity.

CLINICAL DISEASE FEATURES COMMON TO IMMUNE DEFICIENCY

Immunodeficiency syndromes, whether congenital, spontaneously acquired, or iatrogenic, are characterized by unusual susceptibility to infection and not infrequently to autoimmune disease and lymphoreticular malignancies. The types of infection often provide the first clue to the nature of the immunologic defect.

Patients with *defects in humoral immunity* have recurrent or chronic sinopulmonary infection, meningitis, and bacteremia, most commonly caused by pyogenic bacteria such as *Haemophilus influenzae*, *Streptococcus pneumoniae*, and staphylococci. These and other pyogenic organisms also cause frequent infections in individuals who have either neutropenia or a deficiency of the pivotal third component of complement (C3). The tripartite collaboration of antibody, complement, and phagocytes in host defense against pyogenic organisms makes it important to assess all three systems in individuals with unusual susceptibility to bacterial infections.

Antibody-deficient patients in whom cell-mediated immunity is intact have an interesting response to viral infections. The clinical course of primary infection with viruses such as varicella zoster or rubeola, unless complicated by bacterial infection, does not differ significantly from that of the normal host. However, long-lasting immunity may not develop, and as a result, multiple bouts of chickenpox and measles may occur. Such observations suggest that intact T cells may be sufficient for control of established viral infections, while antibodies play an important role in limiting the initial dissemination of virus and in providing long-lasting protection. Exceptions to this generalization are becoming more widely recognized. Agammaglobulinemic patients fail to clear hepatitis B virus from their circulation and have a progressive, and often fatal, course. Poliomyelitis has occurred following live-virus vaccination in some patients. Chronic encephalitis, which may progress over a period of months to years, is a particular threat in congenitally agammaglobulinemic boys. Echoviruses and adenoviruses have been isolated from brain, spinal fluid, or other sites in such patients.

The occurrence of an unusually serious infection, for example, *H. influenzae* meningitis in an older child or adult, warrants consideration of humoral immune deficiency. Recurrent bacterial pneumonias also suggest this possibility. Chronic otitis media occurs frequently in patients with hypogammaglobulinemia and is significant because of its relative rarity in normal adults. Pansinusitis, although almost invariably present in immunoglobulin deficiency, is a less helpful finding because it is not rare in apparently normal people. Bacterial infections of the skin or urinary tract are less frequent problems in hypogammaglobulinemic patients.

Infestation with the intestinal parasite *Giardia lamblia* is a frequent cause of diarrhea in antibody-deficient patients.

Abnormalities of cell-mediated immunity predispose to disseminated virus infections, particularly with latent viruses such as herpes simplex (Chap. 182), varicella zoster (Chap. 183), and cytomegalovirus (Chap. 185). In addition, patients so affected almost invariably develop mucocutaneous candidiasis and frequently acquire systemic fungal infections. Pneumonia caused by *Pneumocystis carinii* is also common (Chap. 209). Severe enteritis caused by *Cryptosporidium* infection may extend to the biliary tract to result in sclerosing cholangitis.

T cell deficiency is always accompanied by some abnormality of antibody responses (Fig. 308-1), although this may not be reflected by hypogammaglobulinemia. This explains in part why patients with primary T cell defects are also subject to overwhelming bacterial infection.

The most severe form of immune deficiency occurs in individuals, often infants, who lack both cell-mediated and humoral immune functions. Individuals with severe combined immunodeficiency (SCID) are susceptible to the whole range of infectious agents including organisms not ordinarily considered pathogenic. Multiple infections with viruses, bacteria, and fungi occur, often simultaneously. Because donor lymphocytes cannot be rejected by these recipients, blood transfusions can produce fatal graft-versus-host disease.

EVALUATION OF IMMUNODEFICIENT PATIENTS

A careful history and physical examination will usually indicate whether the major problem involves the antibody-complement-phagocyte system or cell-mediated immunity. A history of contact dermatitis due to poison ivy suggests intact cellular immunity. Persistent mucocutaneous candidiasis suggests deficient cell-mediated immunity. Lymphopenia and the absence of palpable lymph nodes may be important findings. However, patients with profound immunodeficiency may have diffuse lymphoid hyperplasia. Most immunodeficiencies may be diagnosed by thoughtful use of tests available in local or regional clinical laboratories. More precise evaluation of immunologic functions and treatment may require referral to specialized centers.Table 308-1 presents a resume of widely available laboratory investigations.

Humoral Immunity With rare exceptions, deficiency of humoral immunity is accompanied by diminished serum concentration of one or more classes of immunoglobulin. Normal values vary with age, and adult concentrations of IgM (1.0 ± 0.4 g/L) are reached at about 1 year, of IgG (10.0 ± 3.0 g/L) at 5 to 6 years, and of IgA (2.5 ± 1.0 g/L) by puberty (Chap. 305). The wide range of values among normal adults creates difficulty in defining the lower limits of normal. Reasonable estimates for low values are below 0.4 g/L for IgM, 5 g/L for IgG, and 0.5 g/L for IgA. In the presence of borderline hypogammaglobulinemia, assessing the patient's capacity to produce specific antibodies becomes particularly important. Isohemagglutinins, anti-streptolysin O, and "febrile agglutinins" are valuable standard assays, and measurements of pre- and postimmunization titers to tetanus toxoid, diphtheria toxoid, *H. influenzae* capsular polysaccharide, and *S. pneumoniae* serotypes provide a comprehensive assessment of humoral responsiveness.

Estimation of numbers of circulating B and T lymphocytes is of value in determining the pathogenesis of certain types of immune deficiency. B lymphocytes are identified by the presence of membrane-bound immunoglobulins, their associated a- and b-chain units, and other lineage-specific molecules on the B cell surface (Table 308-1), which can be identified and enumerated by specific monoclonal antibodies.

Since antibody deficiency may be mimicked clinically by deficiency of complement components, measurement of total hemolytic complement (CH_{50}) should be a part of the evaluation of host defense. Measurement of C3 alone is inadequate for screening, since deficiencies of both early and late complement components may predispose to bacterial infection (Chap. 305).

Cellular Immunity T lymphocytes may be enumerated by their expression of the TCR/CD3 complex of surface molecules. The CD4 molecule serves as a marker for helper T cells, although macrophages also express this molecule in relatively low levels. Conversely, CD8ab heterodimers are expressed by cytotoxic T cells. CD8 is also expressed by some gd T cells and byNKcells, although usually as CD8aa homodimeric molecules.

Normal levels of serum immunoglobulins and antibody responsiveness are reliable indices of intact helper T cell function. T lymphocyte function can be measured directly by delayed hypersensitivity skin testing using a variety of antigens to which the majority of older children and adults have been sensitized. A generally useful skin test antigen is a 1:5 dilution of tetanus toxoid injected intradermally, since almost all individuals will have been sensitized. Purified protein derivative (PPD), histoplasmin, mumps antigen, and extracts of *Candida* or *Trichophyton* also may be used.

T lymphocyte function may be estimated in vitro by the capacity of cells to proliferate in response to antigens to which the patient has been sensitized, to lymphocytes from an unrelated donor, to antibodies that cross-link the CD3/TCR complex, or to the T cell mitogens, such as phytohemagglutinin and concanavalin A. The response is usually quantified by measurement of incorporation of radioactive thymidine into newly synthesized DNA. The production of cytokines (or interleukins) by activated T cells, can be measured as can the ability of T cells activated in mixed lymphocyte culture to lyse target cells. Finally, assays exist for detection of defects in T cell surface receptors and specific elements of their signal transduction pathways.

CLASSIFICATION

Primary immunodeficiencies may be either congenital or manifested later in life and are currently classified according to mode of inheritance and whether the genetic defect affects T cells, B cells, or both (Table 308-2). The following discussion emphasizes three related concepts: (1) that immunodeficiencies are logically viewed as defects of cellular differentiation; (2) that these defects may involve either primary development of T or B cells or the antigen-dependent phase of their differentiation; and (3) that defects of B cell differentiation may in some instances reflect faulty T-B collaboration.

Secondary immunodeficiencies are those not caused by intrinsic abnormalities in development or function of T and B cells. The best known of these is AIDS, which may follow infection with the human immunodeficiency virus (Chap. 309). Other examples are immune deficiency associated with malnutrition, protein-losing enteropathy, and intestinal lymphangiectasia. Also considered secondary are immunodeficiencies resulting from hypercatabolic states such as occur in myotonic dystrophy, immunodeficiency associated with lymphoreticular malignancy, and immunodeficiency resulting from treatment with x-rays, antilymphocyte antibodies, or immunosuppressive drugs.

Incidence As a group, the primary immunodeficiencies are relatively common. The most frequent, isolated IgA deficiency, occurs in approximately 1 in 600 individuals in North America. Common variable immunodeficiency, a related disorder characterized by pan-hypogammaglobulinemia, is the next most common disorder. Both of these immunodeficiency states often become clinically evident in young adults.

The more severe forms of primary immunodeficiency are relatively rare, have their onset early in life, and all too frequently result in death during childhood. However, patients with congenital hypogammaglobulinemia may survive to middle age and beyond with replacement antibody therapy. In a referral center for patients with immunodeficiency diseases, approximately two-thirds of the immunodeficient patients will be adults.

Severe Combined Immunodeficiency TheSCIDsyndrome is characterized by gross functional impairment of both humoral and cell-mediated immunity and by susceptibility to devastating fungal, bacterial, and viral infections. It is usually congenital, may be inherited either as an X-linked or autosomal recessive defect, or may occur sporadically. Affected infants rarely survive beyond 1 year without treatment.

The syndrome has been associated with a diversity of defects in development of immunocompetent cells, which are caused by mutations in genes whose products are necessary for the normal differentiation of T, B, and, sometimes,NKcells.

In one autosomal recessive form ofSCID characterized by severe lymphopenia, the failure in T and B cell development is due to *mutations in the* RAG-*1* or RAG-*2 genes*, the combined activities of which are needed for V(D)J recombination. A function-loss *mutation in the DNA-dependent tyrosine kinase gene* in SCID mice may prove to be a cause for SCID in humans as well, since this is another essential enzyme in the V(D)J gene rearrangement process. About half of patients with autosomal recessive SCID are deficient in an enzyme involved in purine metabolism, adenosine deaminase (ADA), due to *mutations in the* ADA *gene*. The abortive lymphoid differentiation associated with ADA deficiency is due to intracellular accumulation of adenosine and deoxyadenosine nucleotides that interferes with critical metabolic functions, including DNA synthesis.

SCIDalso may occur with an X-linked inheritance pattern. Aborted thymocyte differentiation and an absence of peripheral T cells andNK cells is seen in *X-linked* SCID. B lymphocytes are present in normal numbers but are functionally defective. The defective gene encodes a commong chain of the receptors for IL-2, -4, -7, -9, and -15, thus disrupting the action of this important set of lymphokines.

The same T-NK-B+SCIDphenotype seen in X-linked SCID can be inherited as an autosomal recessive disease due to mutations in the gene for *JAK3 protein kinase deficiency*. This enzyme associates with the common g chain of the receptors for IL-2, -4, -7, -9, and -15 to serve as a key element in their signal transduction pathways.

TREATMENT

The cellular defects inSCIDpatients logically rest with the pluripotent hematopoietic stem cells or their lymphoid progenitor progeny. Accordingly, the immunological deficits in all of the different types of SCID patients have been repaired by transplantation of histocompatible bone marrow as a source of stem cells, thereby implying that the stromal microenvironments of these individuals are intact and capable of supporting T and B cell development. However, antibody deficiency requiring immunoglobulin replacement therapy may persist for years in thegc deficient and JAK3 deficient patients, unless the defective B cells are eliminated prior to bone marrow transplantation to allow their replacement with normal B cells of donor origin. InADA-deficient patients without histocompatible bone marrow donors, the administration of exogenous ADA (conjugated to polyethylene glycol to prolong its half-life) may improve immunological function and clinical status. ADA gene therapy has also been used with limited success in these patients. Treatment of SCID patients should be performed in centers with a strong research interest in this problem. It is crucial that these patients be recognized early and not be given live viral vaccines or blood transfusions, which may cause fatal graft-versus-host disease.

Primary T Cell Immunodeficiency Reflecting the diversity of T cell functions, abnormalities of T cell development may be responsible for a wide spectrum of immune deficiencies, including combined immunodeficiency, selective defects in cell-mediated immunity, and syndromes presenting as antibody deficiency. These defects may be acquired (Chap. 309) as well as congenital.

DiGeorge's syndrome This classic example of isolated T cell deficiency results from maldevelopment of thymic epithelial elements derived from the third and fourth pharyngeal pouches. The gene defect has been mapped to chromosome 22q11 in most patients with DiGeorge's syndrome, and to chromosome 10p in others. Defective development of organs dependent on cells of embryonic neural crest origin includes: congenital cardiac defects, particularly those involving the great vessels; hypocalcemic tetany, due to failure of parathyroid development; and absence of a normal thymus. Facial abnormalities may include abnormal ears, shortened philtrum, micrognathia, and hypertelorism. Serum immunoglobulin concentrations are frequently normal, but antibody responses, particularly of IgG and IgA isotypes, are usually impaired. T cell levels are reduced, whereas B cell levels are normal. Affected individuals usually have a small, histologically normal thymus located near the base of the tongue or in the neck, allowing most patients to develop functional T cells in numbers that may or may not be adequate for host defense.

The Nude Syndrome The human disease counterpart to the *nude* mouse is also caused by mutations of the *whn (winged-helix-nude)* gene that result in impairment of hair follicle and epithelial thymic development. The human *nude* phenotype is characterized by congenital baldness, nail dystrophy and severe T cell immunodeficiency.

T Cell Receptor Deficiency Since the expression and function of antigen-specific T cell receptors (TCR) is dependent on their companion CD3g,d, e, andz-h chains, defective genes for any of these receptor components can impair T cell development and function. Immunodeficiencies due to inherited CD3g and CD3e mutations have been identified. CD3g mutations result in a selective deficit in CD8 T cells, whereas CD3e mutations lead to a preferential reduction in CD4 T cells, thus implying differences in the signal transduction roles for each CD3 component.

Major histocompatibility complex (MHC) class II deficiency Because T cells are required for B cell responses to most antigens, any gene defect (or acquired disorder) that interferes with T cell development and cell-mediated immunity will also compromise antibody production and humoral immunity. MHC class II deficiency results in one such immunodeficiency in that theTCRabmust see protein antigens as peptide fragments held within thea helical grooves of class II and class I molecules encoded by the MHC. Antigen-presenting cells in individuals with this relatively rare disorder fail to express the class II molecules DP, DQ, and DR on their surface. Limited numbers of helper CD4 T cells are therefore generated in the thymus, and they fail to see antigen in the periphery. Affected individuals experience recurrent bronchopulmonary infections, chronic diarrhea, and severe viral infections that usually prove fatal before 4 years of age. The defect is caused by mutations in genes that encode essential transcriptional factors that bind to promoter elements for the MHC class II genes. The class II transactivator gene is mutated in one subgroup of MHC class II deficient patients, whereas mutations in RFX genes encoding additional transcriptional factors for MHC class II genes are responsible for the defective development and function of CD4 T cells in other families: RFXANK in subgroup B, RFX5 in subgroup C, and RFXAP in subgroup D.

ZAP70 Tyrosine Kinase Deficiency Recurrent and opportunistic infections begin within the first year of life in individuals with a deficiency in ZAP70 tyrosine kinase, a pivotal component in theTCR/CD3 signal transduction cascade. The rare inheritance of mutations in both alleles of the ZAP70 gene results in a selective deficiency of CD8 T cells and dysfunction of CD4 T cells, which are present in normal numbers. Severe immunodeficiency is the inevitable consequence.

Purine Nucleoside Phosphorylation Deficiency Function-loss mutations of the purine nucleoside phosphorylase (PNP) gene are associated with an often severe and selective deficiency of T lymphocyte function. This enzyme functions in the same purine salvage pathway asADA; toxic effects of the PNP deficiency may result from the intracellular accumulation of deoxyguanosine triphosphate.

Ataxia-Telangiectasia Ataxia-telangiectasia (AT) is an autosomal recessive genetic disorder characterized by cerebellar ataxia, oculocutaneous telangiectasia, and immunodeficiency. The mutant ATM gene has sequence similarity to the phosphatidyl-inositol-3 kinases that are involved in signal transduction. The ATM gene belongs to a conserved family of genes that monitor DNA repair and coordinate DNA synthesis with cell division. The deleterious effects of the ATM gene are widespread. Truncal ataxia may become evident when walking begins and is progressive. Telangiectasia, primarily represented by dilated blood vessels in the ocular sclera, a butterfly area of the face and on the ears, is an early diagnostic feature. Immunodeficiency may be clinically manifest by recurrent and chronic sinopulmonary infection leading to bronchiectasis, although not all patients have overt immunodeficiency. Ovarian agenesis is a frequent occurrence. Persistence of very high serum levels of oncofetal proteins, includinga fetoprotein and carcinoembryonic antigen, may be of diagnostic value. Frequent causes of death are chronic pulmonary disease and malignancy. Lymphomas are most common, although carcinomas also have occurred.

The immunologic abnormalities seem to be related to maldevelopment of the thymus. The markedly hypoplastic thymus is similar in appearance to an embryonic thymus. The peripheral T cell pool is reduced in size, especially in lymphoid tissue compartments. Cutaneous anergy and delayed rejection of skin grafts are common. Although B lymphocyte development is normal, most patients are deficient in serum IgE and IgA, and a smaller number have reduced serum levels of IgG, particularly of the IgG2, IgG4 subclasses.

The defect in DNA repair mechanisms in these patients renders their cells highly susceptible to radiation-induced chromosomal damage and resultant tumor development.AT is a rare disorder, one in 10,000 to 100,000 incidence, but 1% of the population is heterozygous for an AT mutation. This is important because the heterozygous state also predisposes to enhanced cellular radiosensitivity and cancer, especially breast cancer in females (Chap. 364).

TREATMENT

Therapeutic options other than symptomatic treatment are limited for this group of patients. Live vaccines and blood transfusions containing viable T cells should be assiduously avoided. Exposure to X-irradiation should also be avoided in patients withAT. Therapeutic intervention in the form of an epithelial thymic transplant should repair the T cell deficiency in patients with the *nude* syndrome and in the most severe cases of DiGeorge's syndrome where T cells are absent. Preventive therapy for *P. carinii* in the form of trimethoprim-sulfamethoxazole should be considered. Immunoglobulin infusions are also recommended for those T cell deficient individuals with severe antibody deficiency reflected by low serum levels of IgG.

Immunoglobulin Deficiency Syndromes

X-LINKED AGAMMAGLOBULINEMIA Males with this syndrome often begin to have recurrent bacterial infections late in the first year of life, when maternally derived immunoglobulins have disappeared. Although B cell progenitors are found in the bone marrow, affected individuals have very few immunoglobulin-bearing B lymphocytes in their circulation and lack primary and secondary lymphoid follicles. The developmental block is evident at the pre-B cell level (Fig. 308-1). Mutations of Bruton's tyrosine kinase (Btk) gene are responsible for X-linked agammaglobulinemia. B cells in heterozygous female carriers exclusively utilize the X chromosome with the normal Btk gene, while T cells and myeloid cells express either X chromosome. *X-linked agammaglobulinemia with growth hormone deficiency* is a rare variant disorder caused by another gene defect that maps to the same region of the X chromosome.

Agammaglobulinemia is a misnomer, since most of these patients synthesize some immunoglobulins. Within the same family, some affected males may have substantial levels of IgM, IgG, and IgA, while others are nearly agammaglobulinemic.Btk-deficient patients typically are very deficient in circulating B lymphocytes. The few B lymphocytes that escape the block in pre-B cell differentiation are impaired in their responsiveness to antigenic stimulation, making antibody replacement therapy essential in these patients.

Sinopulmonary bacterial infections constitute the most frequent clinical problem. Mycoplasma infections also cause arthritis in some of these patients. Chronic encephalitis of viral etiology, sometimes associated wtih dermatomyositis, can be a fatal complication. These complications are reduced by treatment with intravenous immunoglobulin.

Autosomal Recessive Agammaglobulinemia This syndrome can result from mutations in a variety of genes whose products are required for B lineage differentiation. For example, signals induced via pre-B receptors are essential for pre-B cell development.

Consequently, mutations in any of the genes coding pre-B receptor components -- u heavy chains, surrogate light chains (VpreB and l5/14.1), Iga and Igb -- can block B lineage differentiation. Congenital absence of B cells, agammaglobulinemia and recurrent bacterial infections have been seen in children with function-loss mutations in both alleles of the u heavy chain gene or the l5/14.1 surrogate light chain gene. Disruption of B cell development may also occur as a consequence of mutations in genes coding transcription factors for pre-B receptor genes or for key elements in the pre-B receptor signaling pathway.

Transient Hypogammaglobulinemia of Infancy This diagnosis is reserved for those rare instances in which normal physiologic hypogammaglobulinemia of infancy is unusually prolonged and severe. IgG levels normally drop to 3.0 to 4.0 g/L between 3 and 6 months of age as maternally derived IgG is catabolized. The IgG levels subsequently rise, reflecting the infants' increased synthetic capacity. Periodic immunologic assessment is needed to differentiate transient hypogammaglobulinemia from other forms of antibody deficiency. Antibody replacement therapy is recommended only in the face of severe or recurrent infections.

IgA Deficiency An inability to produce antibodies of the IgA1 and IgA2 subclasses occurs in approximately 1 in 600 individuals of European origin, a much higher incidence than is seen for other primary immunodeficiencies. IgA deficiency is much less common in people of Asian and African origin. In Japan, for example, the incidence is approximately 1 in 18,500. While the precise genetic basis for this difference in incidence is unknown, IgA deficiency is frequently associated with certainMHChaplotypes that are more common in Caucasians.

Individuals with isolated IgA deficiency may appear healthy or present with an increased number of respiratory infections of varying severity, and a few have progressive pulmonary disease leading to bronchiectasis. Chronic diarrheal diseases also occur. Reductions in the IgG2 and IgG4 subclasses are associated with the increased infections in some IgA-deficient individuals. The incidence of asthma and other atopic diseases among IgA-deficient patients is high. Conversely, the incidence of IgA deficiency among atopic children has been found to be more than 20 times that in the normal population. IgA deficiency is also significantly associated with arthritis (Chap. 312) and systemic lupus erythematosus (Chap. 311). IgA-deficient patients frequently produce autoantibodies. Some of them develop significant levels of antibodies to IgA, which may render them vulnerable to severe anaphylactic reactions when transfused with normal blood or blood products.

An accurate picture of the clinical consequences of IgA deficiency requires lifelong study of affected individuals. Among 204 healthy young adults whose IgA deficiency was identified when they served as blood donors, 80% were found to experience episodes of infections, drug allergy, autoimmune disorders, or atopic disease during the next 20 years of their life. They had an increased susceptibility to pneumonia, recurrent episodes of respiratory infections, and a higher incidence of autoimmune diseases, including vitiligo, autoimmune thyroiditis, and possibly rheumatoid arthritis.

IgA deficiency is often familial. It can also occur in association with congenital intrauterine infections, such as toxoplasmosis, rubella, and cytomegalovirus infection, or following treatment with phenytoin, penicillamine, or other medications in genetically susceptible individuals. The pathogenesis of IgA deficiency, whether genetic or triggered by environmental insult, involves a block in B cell differentiation that may reflect defective interaction between T and B cells.

Treatment of IgA deficiency is essentially symptomatic. IgA cannot be effectively replaced by exogenous immunoglobulin or plasma, and use of either can increase the risk of development of antibodies to IgA. IgA-deficient patients in need of transfusion should be screened for the presence of antibodies to IgA and ideally should be given blood only from IgA-deficient donors. Immunoglobulin infusions may benefit the exceptional IgA deficient person in whom IgG2 and IgG4 subclass deficiencies are associated with severe infections, but the risk of anaphylactic reactions to contaminating IgA must always be considered in treating these patients.

IgG Subclass Deficiencies Selective deficiencies in one or more of the four IgG subclasses are seen in some patients with repeated infections. The IgG subclass deficiency may easily go undetected when the total serum IgG level is measured, because IgG2, IgG3, and IgG4 together account for only 30 to 40% of the IgG antibodies. Even a deficiency in IgG1 may be masked by increases in the remaining IgG isotypes. However, the availability of subclass-specific monoclonal antibodies allows precise measurement of IgG subclass levels.

Homozygous deletions of genes encoding the constant region of the different g chains is the basis for the IgG subclass deficiency in some individuals. For example, deletion of the C_{a1}, C_{g2}, C_{g4}, and C_{e} genes in the heavy chain locus on both chromosomes 14 was responsible for one individual's inability to make IgA1, IgG2, IgG4, and IgE. Because other components of their immune system are intact, individuals with this and other patterns of C_{H}-gene deletions may not have unusual infections.

Most of the IgG subclass-deficient individuals with repeated infections appear to have regulatory defects that prevent normal B cell differentiation. The defect may extend to other isotypes. IgA deficiency may accompany IgG2 and IgG4 subclass deficiencies (see "IgA Deficiency" above); an inability to produce IgM antibodies to polysaccharide antigens often reflects a broader defect in antibody responsiveness. While patients with IgG subclass deficiency may benefit from administration of immunoglobulin, a thorough assessment of humoral immunity is needed to identify the relatively few who need this therapy.

Common Variable Immunodeficiency This diagnostic category includes a heterogeneous group of males and females, mostly adults, who have in common the clinical manifestations of deficient production of all the different classes of antibodies. The majority of these hypogammaglobulinemic patients have normal numbers of B lymphocytes that are clonally diverse but phenotypically immature. B lymphocytes in these patients are able to recognize antigens and can proliferate in response, but they largely fail to become mature plasma cells. This abortive differentiation pattern leads to the frequent occurrence of nodular B lymphocyte hyperplasia, resulting in splenomegaly and intestinal lymphoid hyperplasia, sometimes of massive proportion.

It is important to note that common variable immunodeficiency and IgA deficiency represent polar ends of a clinical spectrum due to the same underlying gene defect in a large subset of these patients. The two disorders feature similar B cell differentiation arrests, differing only in the numbers of immunoglobulin classes involved. Over a period of years, IgA deficient patients may progress to the pan-hypogammaglobulinemia phenotype characteristic of common variable immunodeficiency, and vice versa. Both disorders occur frequently within the same family, and the sameMHChaplotypes are associated with both immunodeficiency patterns. Family studies suggest an underlying susceptibility gene in the MHC class III region for both disorders.

It is important to consider the diagnosis of common variable immunodeficiency in adults with chronic pulmonary infections, some of whom will present with bronchiectasis. Intestinal diseases -- including chronic giardiasis, intestinal malabsorption, and atrophic gastritis with pernicious anemia -- are common in this group of patients. Patients with common variable immunodeficiency also may present with signs and symptoms highly suggestive of lymphoid malignancy, including fever, weight loss, anemia, thrombocytopenia, splenomegaly, generalized lymphadenopathy, and lymphocytosis. Histologic examination of lymphoid tissues usually reveals germinal center hyperplasia which may be difficult to distinguish from nodular lymphoma (Chap. 112). Demonstration of a normal distribution of immunoglobulin isotypes and light chain classes for circulating and tissue B lymphocytes can serve to distinguish these patients from those having a monoclonal B cell malignancy with secondary hypogammaglobulinemia. The administration of intravenous immunoglobulin in adequate doses (see below) is an essential part of the prevention and treatment of all these complications.

X-Linked Immunodeficiency with Hyper IgM In this syndrome, typically the IgG and IgA levels are very low, while IgM levels may be very high, normal, or even low. The development of B lymphocytes bearing IgM and IgD and the absence of IgG and IgA B lymphocytes indicate a defect in isotype switching. The defective gene in these patients encodes a transiently expressed molecule on activated T cells that is the ligand for the CD40 molecule on dendritic cells (DC) and B cells. Gene mutations that preclude normal CD40 ligand expression prevent normal T and B cell cooperation, germinal center formation, V-region diversification by somatic hypermutation, and isotype switching. T cell responses are also compromised in these CD40 ligand deficient patients because their T cells are deprived of an important feedback stimulus as a consequence of the defective T, DC, B cell interactions (Chap. 305). Consequently, these patients experience more severe infections than those occurring with other hypogammaglobulinemic states. In addition to recurrent bacterial infections, pneumonia may be caused by *P. carinii*, cytomegalovirus, *Aspergillus*, *Cryptosporidium*, and other unusual organisms. Enteritis due to cryptosporidial infection may extend into the biliary tract to result in a sclerosing cholangitis and hepatic cirrhosis. Neutropenia is frequent in affected males and increases their vulnerability to infections.

Immunodeficiency with hyper IgM is also seen in patients of both sexes who lack mutations in their CD40 ligand gene. While the phenotype in the non-X-linked form of immunodeficiency with hyper IgM is similar, the clinical course is usually milder than in CD40 ligand deficient patients. Candidate disease genes in this syndrome include the CD40 gene and genes coding signaling elements in the CD40 signaling pathway.

TREATMENT

Replacement therapy with human immunoglobulin is the therapeutic cornerstone for antibody-deficient patients who have recurrent infections and who are deficient in IgG. Maintenance of serum IgG levels above 5.0 g/L will prevent most systemic infections in the patients. These serum levels usually can be achieved by intravenous administration of immunoglobulin, 400 to 500 mg/kg, at 3- to 4-week intervals. In patients with mild to moderate IgG deficiency (3.0 to 5.0 g/L) or isolated IgG subclass deficiencies, the decision to treat should be based on evaluation of clinical symptoms and antibody responses to antigenic challenge. Since immunoglobulin preparations are comprised almost entirely of IgG antibodies, they are of no value for repairing deficiencies of immunoglobulins other than IgG. Infusions of immunoglobulin are also not benign. While HIV transmission has not been reported, previous epidemics of hepatitis C virus infections in hypogammaglobulinemic patients receiving contaminated immunoglobulin preparations have led to improved safety measures for current commercial preparations. Some antibody-deficient patients develop symptoms of diaphoresis, tachycardia, flank pain, and hypotension during immunoglobulin infusion. This reaction may be mediated by aggregates of IgG or other biologically active substances and often is resolved by slowing the rate of immunoglobulin infusion. More serious anaphylactic reactions may occur as a consequence of antibodies produced by the patient against donor immunoglobulins, particularly IgA (Chap. 114). The potential for severe adverse reactions merits administration of the initial immunoglobulin infusion under medical supervision in a hospital or clinic setting.

A heightened index of suspicion of infection is essential for antibody-deficient patients. Identification of infectious agents in order to select appropriate antibiotic, antiparasitic, or antiviral therapy is also very important. Immunoglobulin infusions usually do not suffice to eliminate chronic sinopulmonary infections with *H. influenzae* and other microorganisms, and a prolonged course of antibiotic therapy may be required to effectively treat these infections and prevent progression to pulmonary fibrosis and bronchiectasis. Maintenance of good pulmonary toilet with regular postural drainage can also be especially important in management of these patients. Infestation with *G. lamblia*, a common cause of chronic diarrhea in antibody deficient patients, usually responds to therapy with metronidazole.

Cryptosporidial infections in CD40 ligand deficient patients may respond to long-term treatment with amphotericin B and flucytosine. The neutropenia frequently associated with infections in these patients may or may not resolve with improvement of infections and antibody replacement therapy. Bone marrow transplantation following myeloablative pretransplantation therapy can be curative for boys with this devastating immunodeficiency. This treatment has a much greater chance of success when performed during childhood.

Miscellaneous Immunodeficiency Syndromes Infection with *Candida albicans* is the almost universal accompaniment of severe deficiencies in cell-mediated immunity. *Chronic mucocutaneous candidiasis* is different because superficial candidiasis is usually the only major manifestation of immunodeficiency in this syndrome. These patients rarely develop systemic infection with *Candida* or other fungal agents and are not unusually susceptible to virus or bacterial disease. No uniformity of immunologic defects has been identified in these patients, although defects of antibody formation have been detected occasionally. Humoral immunity, including ability to make specific anti-*Candida* antibodies, is usually normal. Many patients are anergic, some to a variety of antigens and some only to *Candida*. The syndrome is often congenital and may be associated with single or multiple endocrinopathies as well as iron deficiency. Treatment of associated conditions may lead to improvement or even cure of *Candida* infection. In other patients, intensive treatment with amphotericin B coupled with surgical removal of infected nails has led to sustained improvement. Oral antifungal agents, such as fluconazole and itraconazole, also may be effective.

Interferong Receptor Deficiency This immunodeficiency is characterized by serious infections caused by bacille Calmette-Guerin vaccine and environmental non-tuberculous mycobacteria. Associated salmonella infections occur in a minority of the cases. This syndrome can be caused by mutations in the interferon greceptor signal-transducing chain (IFNGR2). Two additional forms of this syndrome are caused by different types of mutations in the interferong receptor 1 (IFNGR1) gene that encodes the ligand binding chain of the interferon g receptor. Null mutations in both IFNGR1 alleles are responsible for a more severe autosomal recessive form. A less severe form, inherited in an autosomal dominant pattern, is caused by IFNGR1 mutations in a small deletional hotspot that result in a truncated receptor chain lacking the cytoplasmic tail. Accumulation of the truncated receptor on the surface of macrophages compromises their response to interferon g and the killing of ingested mycobacterium.

Interleukin 12 Receptor Deficiency Mutations in the gene coding theb$_1$subunit of the IL-12 receptor can cause this syndrome. Affected patients suffer from disseminated mycobacterial infections attributable to bacille Calmette-Guerin and nontuberculous mycobacteria, and in some cases non-typhi salmonella infections. Although the clinical manifestations are usually less severe than in patients with completeIFNGR1deficiency, IL-12 receptor deficiency may predispose individuals to clinical tuberculosis as well. Deficient interferon gproduction by the otherwise normalNK and T cells is seen in IL-12 receptor deficient patients, and therapeutic use of interferon g may cure their mycobacterial infection.

Immunodeficiency with Thymoma The association of hypogammaglobulinemia with spindle cell thymoma usually occurs relatively late in adult life. Bacterial infections and severe diarrhea often reflect the antibody deficiency, whereas fungal and viral infections are infrequent complications. T cell numbers and cell-mediated immunity are usually intact, but these patients are very deficient in circulating B lymphocytes and pre-B cells in the bone marrow. They also frequently have eosinopenia and may develop erythroid aplasia. Complete bone marrow failure sometimes occurs. The relationship between the thymoma and apparent abnormalities of hematopoietic stem cells remains conjectural, and treatment is limited to immunoglobulin administration and symptomatic therapy.

Wiskott-Aldrich Syndrome This X-linked disease characterized by eczema, thrombocytopenia, and repeated infections, is caused by mutations in the WASP gene. The WASP protein is expressed in cells of all hematopoietic lineages. It may serve a cytoskeletal organizing role for signaling elements that are particularly important in platelets and T cells. The platelets are small and have a shortened half-life. Affected male infants often present with bleeding, and most do not survive childhood, dying of complications of bleeding, infection, or lymphoreticular malignancy. The immunologic defects include low serum concentrations of IgM, while IgA and IgG are normal and IgE is frequently increased. The number and class distribution of B lymphocytes are usually normal. Functionally, these patients are unable to make antibodies to polysaccharide antigens normally; responses to protein antigens may also be impaired late in the course of the disease. Most patients eventually acquire T cell deficiencies. Affected boys frequently become anergic, and their T cells do not respond normally to antigenic challenge. This results in vulnerability to overwhelming infections with herpes simplex virus and other infectious agents.

Transplantation of histocompatible bone marrow from a sibling donor following myeloablative therapy can correct both the hematologic and immunologic abnormalities. In patients lacking a suitable donor, intravenous immunoglobulin infusions or splenectomy may improve platelet counts and reduce the risk of serious hemorrhage. Because of the increased risk of pneumococcal bacteremia, splenectomized patients should receive prophylactic penicillin.

X-Linked Lymphoproliferative Syndrome This disease involves a selective impairment in immune elimination of Epstein-Barr virus (EBV). A fulminant and fatal outcome is the consequence of EBV infection in approximately half of the affected males. Hypogammaglobulinemia is the outcome in 30%, and B cell malignancies are acquired in approximately 25% of EBV-infected patients. The disease may be manifested from early childhood onward, depending on the time of EBV infection. Carrier females handle EBV infections normally. Generation of cytotoxic T cells appears to be the primary mechanism of control of EBV infection in normal persons. In males with the X-linked lymphoproliferative syndrome, this process is impaired as a consequence of mutations in a gene coding for a T cell signaling element called SH2D1A or SAP. Intravenous immunoglobulins should be administered to affected males who develop hypogammaglobulinemia. Bone marrow transplantation from an HLA-matched donor may be curative, especially in younger children with this syndrome. However, myeloablative chemotherapy is a necessary prerequisite to successful bone marrow transplantation, thereby increasing the risk of this procedure.

Hyper-IgE Syndrome The hyper IgE syndrome (Chap. 64) is characterized by recurrent abscesses involving skin, lungs, and other organs and very high IgE levels. IgE levels may decline with time to reach normal levels in approximately 20% of affected adults. Staphylococcal infection is common to all patients, but most have infections with other pyogenic organisms as well. Abnormal neutrophil chemotaxis is an inconsistent finding, and diminished antibody responses to secondary immunization have been noted. Non-immunologic features include impaired shedding of the primary teeth, recurrent bone fractures, hyperextensible joints and scoliosis. Males and females are affected in an inheritance pattern suggesting an autosomal dominant defect with variable penetrance, but the gene defect has not been identified. Prophylaxis with penicillinase-resistant penicillins or cephalosporins is highly recommended to prevent staphylococcal infections. Pneumatoceles, a frequent complication of pneumonias, may require surgical excision.

Metabolic Abnormalities Associated with Immunodeficiency The relation of deficiencies of the purine salvage enzymes adenosine deaminase and purine nucleoside phosphorylase to immunodeficiency was discussed earlier. The syndrome of *acrodermatitis enteropathica* includes severe desquamating skin lesions, intractable diarrhea, bizarre neurologic symptoms, variable combined immunodeficiency, and an often fatal outcome. This disease is apparently caused by an inborn error of metabolism resulting in malabsorption of dietary zinc and can be treated effectively by parenteral or large oral doses of zinc. Zinc deficiency might in part account for the immunodeficiency that accompanies severe malnutrition. Inherited *deficiency of transcobalamin II*, the serum carrier molecule responsible for transport of vitamin B_{12} to tissues, is associated with failure of immunoglobulin production as well as megaloblastic anemia, leukopenia, thrombocytopenia, and severe malabsorption. All abnormalities of this rare disorder are reversed by administration of vitamin B_{12}.

CONCLUSION

Defective genes have been identified for most of the primary immunodeficiency diseases that are currently recognized (Table 308-3). It can be anticipated that many different gene mutations will be identified in other individuals with increased susceptibility to infection. Identification of the mutant genes is the first step toward a better understanding of the pathogenesis of immunodeficiency disease and improved therapeutic strategies. Successful gene repair is the ultimate goal for these individuals.

ACKNOWLEDGEMENT

This chapter represents a revised version of a chapter by Dr. Max D. Cooper and Dr. Alexander R. Lawton III that has appeared in previous editions of this textbook.

(Bibliography omitted in Palm version)

Back to Table of Contents

309. HUMAN IMMUNODEFICIENCY VIRUS (HIV) DISEASE: AIDS AND RELATED DISORDERS - *Anthony S. Fauci*, *H. Clifford Lane*

AIDS was first recognized in the United States in the summer of 1981, when the U.S. Centers for Disease Control and Prevention (CDC) reported the unexplained occurrence of *Pneumocystis carinii* pneumonia in five previously healthy homosexual men in Los Angeles and of Kaposi's sarcoma (KS) in 26 previously healthy homosexual men in New York and Los Angeles. Within months, the disease became recognized in male and female injection drug users (IDUs) and soon thereafter in recipients of blood transfusions and in hemophiliacs. As the epidemiologic pattern of the disease unfolded, it became clear that a microbe transmissible by sexual (homosexual and heterosexual) contact and blood or blood products was the most likely etiologic agent of the epidemic.

In 1983, HIV was isolated from a patient with lymphadenopathy, and by 1984 it was demonstrated clearly to be the causative agent of AIDS. In 1985, a sensitive enzyme-linked immunosorbent assay (ELISA) was developed, which led to an appreciation of the scope of HIV infection among cohorts of individuals in the United States who admitted to practicing high-risk behavior (see below) as well as among selected populations that had been screened, such as blood donors, military recruits and active-duty military personnel, Job Corps applicants, and patients in selected sentinel hospitals. In addition, seroprevalance studies revealed the enormity of the global pandemic, particularly in developing countries (see below).

The staggering worldwide growth of the HIV pandemic has been matched by an explosion of information in the areas of HIV virology, the pathogenesis (both immunologic and virologic) and treatment of HIV disease, the treatment and prophylaxis of the opportunistic diseases associated with HIV infection, and vaccine development. The information flow related to HIV disease is enormous, and it has become almost impossible for the health care generalist to stay abreast of the literature. The purpose of this chapter is to present the most current information available on the scope of the epidemic; on its pathogenesis, treatment, and prevention; and on prospects for vaccine development. Above all, the aim is to provide a solid scientific basis and practical guidelines for a state-of-the-art approach to the HIV-infected patient.

DEFINITION

With the identification of HIV in 1983 and its proof as the etiologic agent of AIDS in 1984, and with the availability of sensitive and specific diagnostic tests for HIV infection, the case definition of AIDS has undergone several revisions over the years. The latest revision took place in 1993; this revisedCDCclassification system for HIV-infected adolescents and adults categorizes persons on the basis of clinical conditions associated with HIV infection and CD4+ T lymphocyte counts. The system is based on three ranges of CD4+ T lymphocyte counts and three clinical categories and is represented by a matrix of nine mutually exclusive categories (Tables 309-1 and309-2). Using this system, any HIV-infected individual with a CD4+ T cell count of<200/uL has AIDS by definition, regardless of the presence of symptoms or opportunistic diseases (Table 309-1). The clinical conditions in clinical category C now include pulmonary tuberculosis (TB), recurrent pneumonia, and invasive cervical cancer (Table 309-2). Once individuals have had a clinical condition in category B, their disease cannot again be classified as category A, even if the condition resolves; the same holds true for category C in relation to category B.

While the definition of AIDS is complex and comprehensive, the clinician should not focus on whether AIDS is present but should view HIV disease as a spectrum ranging from primary infection, with or without the acute syndrome, to the asymptomatic stage, to advanced disease (see below). The definition of AIDS was established not for the practical care of patients but for surveillance purposes.

ETIOLOGIC AGENT

The etiologic agent of AIDS is HIV, which belongs to the family of human retroviruses (Retroviridae) and the subfamily of lentiviruses (Chap. 191). Nononcogenic lentiviruses cause disease in other animal species, including sheep, horses, goats, cattle, cats, and monkeys. The four recognized human retroviruses belong to two distinct groups: the human T lymphotropic viruses (HTLV) I and HTLV-II, which are transforming retroviruses; and the human immunodeficiency viruses, HIV-1 and HIV-2, which are cytopathic viruses (Chap. 191). The most common cause of HIV disease throughout the world, and certainly in the United States, is HIV-1. HIV-1 comprises several subtypes with different geographic distributions (see below). HIV-2 was first identified in 1986 in West African patients and was originally confined to West Africa. However, a number of cases that can be traced to West Africa or to sexual contacts with West Africans have been identified throughout the world. HIV-2 is more closely related phylogenetically to the simian immunodeficiency virus (SIV) found in sooty mangabeys than it is to HIV-1. In 1999, it was demonstrated that HIV-1 infection in humans was zoonotic and had originated from the *Pan troglodytes troglodytes* species of chimpanzees in whom the virus had co-evolved over centuries. The taxonomic relationship among primate lentiviruses is shown in Fig. 309-1.

MORPHOLOGY OF HIV

Electron microscopy shows that the HIV virion is an icosahedral structure (Fig. 309-2*A*) containing numerous external spikes formed by the two major envelope proteins, the external gp120 and the transmembrane gp41. The virion buds from the surface of the infected cell and incorporates a variety of host proteins, including major histocompatibility complex (MHC) class I and II antigens (Chap. 306), into its lipid bilayer. The structure of HIV-1 is schematically diagrammed inFig. 309-2*B* (Chap. 191).

REPLICATION CYCLE OF HIV

HIV is an RNA virus whose hallmark is the reverse transcription of its genomic RNA to DNA by the enzyme *reverse transcriptase*. The replication cycle of HIV begins with the high-affinity binding of the gp120 protein via a portion of its V1 region near the N terminus to its receptor on the host cell surface, the CD4 molecule (Fig. 309-3). The CD4 molecule is a 55-kDa protein found predominantly on a subset of T lymphocytes that are responsible for helper or inducer function in the immune system (Chap. 305). It is also expressed on the surface of monocytes/macrophages and dendritic/Langerhans cells. In order for HIV-1 to fuse to and enter its target cell, it must also bind to one of a group of co-receptors. The two major co-receptors for HIV-1 are CCR5 and CXCR4.

Both receptors belong to the family of seven-transmembrane-domain G protein-coupled cellular receptors, and the use of one or the other or both receptors by the virus for entry into the cell is an important determinant of the cellular tropism of the virus (see below for details). Following binding, the conformation of the viral envelope changes dramatically, and fusion with the host cell membrane occurs in a coiled-spring fashion via the newly exposed gp41 molecule (Fig. 309-4); the HIV genomic RNA is uncoated and internalized into the target cell (Fig. 309-3). The reverse transcriptase enzyme, which is contained in the infecting virion, then catalyzes the reverse transcription of the genomic RNA into double-stranded DNA. The DNA translocates to the nucleus, where it is integrated randomly into the host cell chromosomes through the action of another virally encoded enzyme, *integrase*. This provirus may remain transcriptionally inactive (latent), or it may manifest varying levels of gene expression, up to active production of virus.

Cellular activation plays an important role in the life cycle of HIV and is critical to the pathogenesis of HIV disease (see below). Following initial binding and internalization of virions into the target cell, incompletely reverse-transcribed DNA intermediates are labile in quiescent cells and will not integrate efficiently into the host cell genome unless cellular activation occurs shortly after infection. Furthermore, some degree of activation of the host cell is required for the initiation of transcription of the integrated proviral DNA into either genomic RNA or mRNA. In this regard, activation of HIV expression from the latent state depends on the interaction of a number of cellular and viral factors. Following transcription, HIV mRNA is translated into proteins that undergo modification through glycosylation, myristylation, phosphorylation, and cleavage. The viral particle is formed by the assembly of HIV proteins, enzymes, and genomic RNA at the plasma membrane of the cells. Budding of the progeny virion occurs through the host cell membrane, where the core acquires its external envelope (Chap. 191). The virally encoded protease then catalyzes the cleavage of the gag-pol precursor to yield the mature virion. Each point in the life cycle of HIV is a real or potential target for therapeutic intervention (see below). Thus far, the reverse transcriptase and protease enzymes have proven to be susceptible to pharmacologic disruption (see below).

HIV GENOME

Figure 309-5 illustrates the arrangement of the HIV genome schematically. Like other retroviruses, HIV-1 has genes that encode the structural proteins of the virus: *gag* encodes the proteins that form the core of the virion (including p24 antigen); *pol* encodes the enzymes responsible for reverse transcription and integration; and *env* encodes the envelope glycoproteins. However, HIV-1 is more complex than other retroviruses, particularly those of the nonprimate group, in that it also contains at least six other genes (*tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu*), which code for proteins involved in the regulation of gene expression (Chap. 191). Several of these proteins are felt to play a role in the pathogenesis of HIV disease. For example, Tat, Nef, and Vpu have all been shown to downregulateMHCclass I expression; this may be a strategy that the virus employs to evade immune-mediated elimination by CD8+ cytolytic T cells. Nef also downregulates cell surface expression of CD4 by inducing endocytosis and lysosomal degradation. Supernatants from Nef-expressing macrophages have been shown to induce chemotaxis and activation of resting T lymphocytes leading to productive HIV infection. In addition to its primary function as a transcriptional enhancer in infected cells, Tat may also be secreted and directly activate potential target cells. In addition, in its secreted form Tat has been shown to be immunosuppressive directly as well as indirectly by inducing secretion of interferon (IFN)a from monocytes/macrophages. Flanking these genes are the long terminal repeats (LTRs), which contain regulatory elements involved in gene expression (see below) such as the polyadenylation signal sequence; the TATA promotor sequence; the NF-kB and Sp1 enhancer binding sites; the transactivating response (TAR) sequences, where the Tat protein binds; and the negative regulatory element (NRE), whose deletion increases the level of gene expression (Fig. 309-5). The major difference between the genomes of HIV-1 and HIV-2 is the fact that HIV-2 lacks the *vpu* gene and has a *vpx* gene not contained in HIV-1.

MOLECULAR HETEROGENEITY OF HIV-1

Molecular analyses of various HIV isolates reveal sequence variations over many parts of the viral genome. For example, in different isolates, the degree of difference in the coding sequences of the viral envelope protein ranges from a few percent (very close) to 50%. These changes tend to cluster in hypervariable regions. One such region, called V3, is a target for neutralizing antibodies and contains recognition sites for T cell responses (see below). Variability in this region is likely due to selective pressure from the host immune system. The extraordinary variability of HIV-1 is in marked contrast to the relative genetic stability ofHTLV-I and -II.

There are two groups of HIV-1: group M (major), which is responsible for most of the infections in the world; group O (outlier), a relatively rare viral form found originally in Cameroon, Gabon, and France; and a third group (group N) first identified in a Cameroonian woman with AIDS. The M group comprises eight subtypes, or *clades*, designated A, B, C, D, F, G, H, and J, as well as four major circulating recombinant forms (CRFs). These four CRFs are the AE virus, prevalent in southeast Asia and often referred to simply as E, despite the fact that the parental E virus has never been found; AG from west and central Africa; AGI from Cyprus and Greece; and AB from Russia. These 8 subtypes and 4 CRFs create the major branches in the phylogenetic tree that represents the lineage of the M group of HIV-1 (Fig. 309-6; also seehttp://hiv-web.lanl.gov/ALIGN_CURRENT/SUBTYPE-REF/Table1.html).

The global patterns of HIV-1 variation likely result from accidents of viral trafficking. Subtype B viruses, which now differ by up to 17% in their *env* coding sequences, are the overwhelmingly predominant viruses seen in the United States, Canada, certain countries in South America, western Europe, and Australia. Other subtypes are also present in these countries to varying degrees. It is thought that, purely by chance, subtype B was seeded into the United States in the late 1970s, thereby establishing an overwhelming founder effect. Subtype C viruses (of the M group) are the most common form worldwide; many countries have cocirculating viral subtypes that are giving rise toCRFs.Figure 309-7 schematically diagrams the worldwide distribution of HIV-1 subtypes by region. The predominant subtype in Europe and the Americas is subtype B. In Africa, >75% of strains recovered to date have been of subtypes A, C, and D, with C being the most common. In Asia, HIV-1 isolates of subtypes E, C, and B predominate. Subtype E accounts for most infections in Southeast Asia, while subtype C is prevalent in India (see "HIV Infections and AIDS Worldwide," below). Sequence analyses of HIV-1 isolates from infected individuals indicate that recombination among viruses of different clades likely occurs as a result of infection of an individual with viruses of more than one clade, particularly in geographic areas where clades overlap.

TRANSMISSION

HIV is transmitted by both homosexual and heterosexual contact; by blood and blood products; and by infected mothers to infants either intrapartum, perinatally, or via breast milk. After approximately 20 years of scrutiny, there is no evidence that HIV is transmitted by casual contact or that the virus can be spread by insects, such as by a mosquito bite.

SEXUAL TRANSMISSION

HIV infection is predominantly a sexually transmitted disease (STD) worldwide. Although approximately 42% of new HIV infections in the United States are among men who have sex with men, heterosexual transmission is clearly the most common mode of infection worldwide, particularly in developing countries. Furthermore, the yearly incidence of new cases of AIDS attributed to heterosexual transmission of HIV is steadily increasing in the United States, mainly among minorities, particularly women in minority groups (Fig. 309-8).

HIV has been demonstrated in seminal fluid both within infected mononuclear cells and in the cell-free state. The virus appears to concentrate in the seminal fluid, particularly in situations where there are increased numbers of lymphocytes and monocytes in the fluid, as in genital inflammatory states such as urethritis and epididymitis, conditions closely associated with otherSTDs(see below). The virus has also been demonstrated in cervical smears and vaginal fluid. There is a strong association of transmission of HIV with receptive anal intercourse, probably because only a thin, fragile rectal mucosal membrane separates the deposited semen from potentially susceptible cells in and beneath the mucosa and trauma may be associated with anal intercourse. Anal douching and sexual practices such as insertion of hard objects or a clenched fist into the rectum ("fisting") traumatize the rectal mucosa, thereby increasing the likelihood of infection during receptive anal intercourse. It is likely that anal intercourse provides at least two modalities of infection: (1) direct inoculation into blood in cases of traumatic tears in the mucosa; and (2) infection of susceptible target cells, such as Langerhans cells, in the mucosal layer in the absence of trauma (see below). Although the vaginal mucosa is several layers thicker than the rectal mucosa and less likely to be traumatized during intercourse, it is clear that the virus can be transmitted to either partner through vaginal intercourse. In a 10-year prospective study in the United States of heterosexual transmission of HIV, male-to-female transmission was approximately eight times more efficient than female-to-male transmission. This difference may be due in part to the prolonged exposure to infected seminal fluid of the vaginal and cervical mucosa, as well as the endometrium (when semen enters through the cervical os). By comparison, the penis and urethral orifice are exposed relatively briefly to infected vaginal fluid. Among various cofactors examined in this study, a history of STDs (see below) was most strongly associated with HIV transmission. In this regard, there is a close association between genital ulcerations and transmission, from the standpoints of both susceptibility to infection and infectivity. Infections with microorganisms such as *Treponema pallidum* (Chap. 172), *Haemophilus ducreyi* (Chap. 149), and herpes simplex virus (HSV;Chap. 182) are important causes of genital ulcerations linked to transmission of HIV. In addition, pathogens responsible for nonulcerative inflammatory STDs such as those caused by *Chlamydia trachomatis* (Chap. 179), *Neisseria gonorrhoeae* (Chap. 147), and *Trichomonas vaginalis* (Chap. 218) are also associated with an increased risk of transmission of HIV infection. Bacterial vaginosis, an infection related to sexual behavior, but not strictly an STD, may also be linked to an incresed risk of transmission of HIV infection. Several studies suggest that treating other STDs and genital tract syndromes may help prevent transmission of HIV. In two studies in Africa aimed at decreasing the incidence of HIV infection by empirical treatment of village inhabitants for other STDs, there were divergent results. In Mwanza, Tanzania, empirical treatment for STDs resulted in a decrease in STDs, including HIV infection. In contrast, in the Rakai district of Uganda, empirical treatment of STDs resulted in a decrease in these diseases but not a decrease in HIV infections. The prevalance of HIV infection in Uganda was considerably greater than that in Tanzania at the time of the studies, and, according to a model of the dynamics of sexual spread of HIV, treatment of other STDs would be expected to have less of an effect on decreasing the transmission of HIV in a population with a higher prevalence than in a population with a lower prevalence of HIV infection. Subsequent studies in Uganda indicated that the chief predictor of heterosexual transmission of HIV was the level of plasma viremia. In some studies the use of oral contraceptives was associated with an increase in incidence of HIV infection over and above that which might be expected by not using a condom for birth control. Finally, lack of circumcision has been strongly associated with a higher risk of HIV infection in certain cohorts. This difference may be due to increased susceptibility of uncircumcised men to ulcerative STDs, as well as other factors such as microtrauma. In addition, the moist environment under the foreskin may promote the presence or persistence of microbial flora which, via inflammatory changes, may lead to higher concentrations of target cells for HIV in the foreskin. Some studies suggest that only circumcision performed before age 20 is associated with a reduced risk of HIV infection. Thus, in certain cases, these phenomena can also be considered as cofactors for HIV transmission.

Oral sex is a much less efficient mode of transmission of HIV than is receptive anal intercourse. However, there is a misperception by some persons that oral sex, particularly among homosexual men, can be proposed as a form of "safe sex" and a substitute for receptive anal intercourse. This is a dangerous approach, as there have been several reports of documented HIV transmission resulting solely from receptive fellatio and insertive cunnilingus. For example, one study reported that in 12 subjects where the precise date of seroconversion could be identified, 4 individuals reported oral-genital contact as their sole risk factor. There are probably many more cases that go unreported because of the frequent practice of both oral sex and receptive anal intercourse by the same person. The association of alcohol consumption and illicit drug use with unsafe sexual behavior, both homosexual and heterosexual, leads to an increased risk of sexual transmission of HIV.

TRANSMISSION BY BLOOD AND BLOOD PRODUCTS

HIV can be transmitted to individuals who receive HIV-tainted blood transfusions, blood products, or transplanted tissue, as well as toIDUs who are exposed to HIV while sharing injection paraphernalia such as needles, syringes, the water in which drugs are mixed, or the cotton through which drugs are filtered. Parenteral transmission of HIV during injection drug use does not require intravenous puncture; subcutaneous ("skin popping") or intramuscular ("muscling") injections can transit HIV as well, even though these behaviors are sometimes erroneously perceived as low-risk. Among IDUs, the risk of HIV infection increases with the duration of injection drug use; the frequency of needle sharing; the number of partners with whom paraphernalia are shared, particularly in the setting of "shooting galleries" where drugs are sold and large numbers of IDUs may share a limited number of "works"; comorbid psychiatric conditions such as antisocial personality disorder; the use of cocaine in injectable form or smoked as "crack"; and the use of injection drugs in a geographic location with a high prevalence of HIV infection, such as certain inner-city areas in the United States.

From the late 1970s until the spring of 1985, when mandatory testing of donated blood for HIV-1 was initiated, it has been estimated that over 10,000 individuals in the United States were infected through transfusions of blood or blood products (Chap. 114). Approximately 8900 individuals in the United States who survived the illness for which they received HIV-contaminated blood transfusions, blood components, or transplanted tissue have developed AIDS. It is estimated that 90 to 100% of individuals who were exposed to such HIV-contaminated products became infected. Transfusions of whole blood, packed red blood cells, platelets, leukocytes, and plasma are all capable of transmitting HIV infection. In contrast, hyperimmune gamma globulin, hepatitis B immune globulin, plasma-derived hepatitis B vaccine, and Rh$_o$immune globulin have not been associated with transmission of HIV infection. The procedures involved in processing these products either inactivate or remove the virus.

In addition to the above, several thousand individuals in the United States with hemophilia or other clotting disorders were infected with HIV by receipt of HIV-contaminated fresh frozen plasma or concentrates of clotting factors; approximately 5310 of these individuals have developed AIDS. Currently, in the United States and in most developed countries, the following measures have made the risk of transmission of HIV infection by transfused blood or blood products extremely small: (1) the screening of all blood for p24 antigen and for HIV antibody byELISA, with a confirmatory western blot where applicable; (2) the self-deferral of donors on the basis of risk behavior; (3) the screening out of HIV-negative individuals with positive surrogate laboratory parameters of HIV infection, such as hepatitis B and C; and (4) serologic testing for syphilis. It is currently estimated that the risk of infection with HIV in the United States via transfused screened blood is approximately 1 in 676,000 donations. Therefore, among the 12 million donations collected in the United States each year, an estimated 18 infectious donations are available for transfusion. The addition of nucleic acid testing to the blood screening protocol to capture some of these rare HIV antibody-negative units should decrease even further the chances of transmission by transfused blood or blood products. There have been several reports of sporadic breakdowns in routinely available screening procedures in certain countries, where contaminated blood was allowed to be transfused, resulting in small clusters of patients becoming infected. There have been no reported cases of transmission of HIV-2 in the United States via donated blood, and, currently, donated blood is screened for both HIV-1 and HIV-2 antibodies. The chance of infection of a hemophiliac via clotting factor concentrates has essentially been eliminated because of the added layer of safety resulting from heat treatment of the concentrates.

Prior to the screening of donors, a small number of cases of transmission of HIV via semen used in artificial insemination and tissues used in organ transplantation were well documented. At present, donors of such tissues are prescreened for HIV infection.

OCCUPATIONAL TRANSMISSION OF HIV: HEALTH CARE WORKERS AND LABORATORY WORKERS

There is a small, but definite, occupational risk of HIV transmission in health care workers and laboratory personnel and potentially in others who work with HIV-infected specimens, particularly when sharp objects are used. An estimated 600,000 to 800,000 health care workers are stuck with needles or other sharp medical instruments in the United States each year. Large, multi-institutional studies have indicated that the risk of HIV transmission following skin puncture from a needle or a sharp object that was contaminated with blood from a person with documented HIV infection is approximately 0.3% (see "HIV and the Health Care Worker," p. 1909). The risk of hepatitis B infection following a similar type of exposure is 6 to 30% in nonimmune individuals; if a susceptible worker is exposed to HBV, postexposure prophylaxis with hepatitis B immune globulin and initiation of HBV vaccine is more than 90% effective in preventing HBV infection. The risk of HCV infection following percutaneous injury is approximately 1.8% (Chap. 295). An increased risk for HIV infection following percutaneous exposures to HIV-infected blood is associated with exposures involving a relatively large quantity of blood, as in the case of a device visibly contaminated with the patient's blood, a procedure that involves a needle placed directly in a vein or artery, or a deep injury. In addition, the risk increases for exposures to blood from patients with advanced-stage disease, probably owing to the higher titer of HIV in the blood as well as to other factors, such as the presence of more virulent strains of virus (see "HIV and the Health Care Worker," p. 1909).

There have been reports of health care workers who became infected through the exposure of mucous membranes or abraded skin to HIV-infected material; however, the risk associated with mucocutaneous exposure has been difficult to quantify, because transmission by this route is extremely rare. Factors that might be associated with mucocutaneous transmission of HIV include exposure to an unusually large volume of blood, prolonged contact, and a potential portal of entry. A prospective study has indicated that the use of antiretroviral drugs as postexposure prophylaxis decreases the risk of infection compared to historic controls in occupationally exposed health care workers. Transmission of HIV through intact skin has not been documented (see "HIV and the Health Care Worker," p. 1909).

Since the beginning of the HIV epidemic, there have been at least three reported instances in which transmission of infection from a health care worker to patients seemed highly probable. The first involved a dentist in Florida who apparently infected six of his patients, most likely through contaminated instruments. Another case involved an orthopedic surgeon in France who apparently infected a patient during placement of a total hip prosthesis. A third case involved the apparent transmission of HIV from a nurse to a surgical patient in France. An additional situation involved the apparent infection of four patients by an HIV-negative general surgeon in Australia during routine outpatient surgery. The cause of the transmission was felt to be a failure on the part of the surgeon to sterilize instruments properly between procedures following prior surgery on an infected patient. Despite these few cases, the risk of transmission from an infected health care worker to patients is extremely low; in fact, too low to be measured accurately. Indeed several epidemiologic studies have been performed tracing thousands of patients of HIV-infected dentists, physicians, surgeons, obstetricians, and gynecologists and no other cases of HIV infection that could be linked to the health care providers were identified. The very occurrence of transmission of HIV as well as hepatitis B and C to and from health care workers in the workplace underscores the importance of the use of universal precautions when caring for all patients (see below andChap. 134).

MATERNAL-FETAL/INFANT TRANSMISSION

HIV infection can be transmitted from an infected mother to her fetus during pregnancy or to her infant during delivery. This is an extremely important form of transmission of HIV infection in developing countries, where the proportion of infected women to infected men is approximately 1:1. Virologic analysis of aborted fetuses indicate that HIV can be transmitted to the fetus as early as the first and second trimester of pregnancy. However, maternal transmission to the fetus occurs most commonly in the perinatal period. This conclusion is based on a number of considerations, including the time frame of identification of infection by the sequential appearance of classes of antibodies to HIV (i.e., the appearance of HIV-specific IgA antibody within 3 to 6 months after birth); a positive viral culture; the appearance of p24 antigenemia weeks to months after delivery, but not at the time of delivery; a polymerase chain reaction (PCR) assay of infant blood following delivery that is negative at birth and positive several months later; the demonstration that the firstborn twin of an infected mother is more commonly infected than is the second twin; and the evidence that cesarean section results in decreased transmission to the infant.

In the absence of prophylactic antiretroviral therapy to the mother during pregnancy, labor, and delivery, and to the fetus following birth (see below), the probability of transmission of HIV from mother to infant/fetus ranges from 15 to 25% in industrialized countries and from 25 to 35% in developing countries. These differences may relate to the adequacy of prenatal care as well as to the stage of HIV disease and the general health of the mother during pregnancy. Higher rates of transmission have been associated with many factors, including high maternal levels of plasma viremia, low maternal CD4+ T cell counts and HIV p24 antibody levels, maternal vitamin A deficiency, a prolonged interval between membrane rupture and delivery, presence of chorioamnionitis at delivery,STDsduring pregnancy, cigarette smoking and hard drug use during pregnancy, preterm labor, obstetric procedures such as amniocentesis and amnioscopy, and other factors that may increase the exposure of the infant to the mother's blood. With regard to levels of viremia, several studies indicate that the risk of transmission increases with the maternal plasma HIV RNA level. In one series of 552 singleton pregnancies in the United States, the rate of mother-to-baby transmission was 0% among women with <1000 copies of HIV RNA per milliliter of blood, 16.6% among women with 1000 to 10,000/mL, 21.3% among women with 10,001 to 50,000/mL, 30.9% among women with 50,001 to 100,000/mL, and 40.6% among women with >100,000/mL. However, there may be no lower "threshold" below which transmission never occurs, since other studies have reported transmission by women with viral RNA levels below the level of detectability of 50 copies per milliliter. Finally, it has been speculated that if the mother experiences acute primary infection during pregnancy, there is a higher rate of transmission to the fetus, owing to the high levels of viremia that occur during primary infection (see below). In the United States and other industrialized countries, zidovudine treatment of HIV-infected pregnant women from the beginning of the second trimester through delivery and of the infant for 6 weeks following birth has dramatically decreased the rate of intrapartum and perinatal transmission of HIV infection from 22.6% in the untreated group to <5%. It is expected that the rate of transmission will decrease even further as more potent combinations of drugs are used in HIV-infected pregnant women (see below).

In developed countries, current recommendations to reduce perinatal transmission of HIV include universal voluntary HIV testing and counseling of pregnant women, zidovudine prophylaxis, obstetric management that attempts to minimize exposure of the infant to maternal blood and genital secretions, and avoidance of breast feeding. It is also recommended that the choice of antiretroviral therapy for pregnant women should be based on the same considerations used for women who are not pregnant, with discussion of the recognized and unknown risks and benefits of such therapy during pregnancy. The cost and logistics of the above protocol are not feasible for developing countries, particularly those in sub-Saharan Africa where the per capita health care delivery allocation is often only a few dollars per year. Studies have demonstrated that truncated regimens of zidovudine alone or in combination with lamivudine given to the mother during the last few weeks of pregnancy or even only during labor and delivery, and to the infant for a week or less, reduced transmission to the infant by 50% compared to placebo. One important study in Uganda demonstrated that a single dose of nevirapine given to the mother at the onset of labor followed by a single dose to the newborn within 72 h of birth decreased transmission to 13% compared to 25% transmission at age 14 to 16 weeks when the mother received multiple doses of zidovudine throughout labor and delivery and the infant received zidovudine daily for a full week following birth. The cost of the nevirapine for the mother and infant was a mere $4.00, which would make this regimen affordable for many developing countries. Approximately 1800 babies are born infected each day throughout the world, and 90% of these are in sub-Saharan Africa; thus, implementation of such a regimen could potentially save 1000 babies per day from becoming infected.

Although most transmission of HIV occurs during pregnancy and at birth, breast feeding may account for 5 to 15% of infants becoming infected after delivery. This is an important modality of transmission of HIV infection in developing countries, particularly where mothers continue to breast feed for prolonged periods. The risk factors for mother-to-child transmission of HIV via breast feeding are not fully understood; factors that increase the likelihood of transmission include detectable levels of HIV in breast milk, the presence of mastitis, low maternal CD4+ T cell counts, and maternal vitamin A deficiency. The risk of HIV infection via breast feeding is highest in the early months of breast feeding. In addition, exclusive breast feeding has been reported to carry a lower risk of HIV transmission than mixed feeding. Certainly, in developed countries breast feeding by an infected mother should be avoided. However, there is disagreement regarding recommendations for breast feeding in certain developing countries, where breast milk is the only source of adequate nutrition as well as immunity against potentially serious infections for the infant. Studies are being conducted to determine whether intermittent administration of nevirapine, which has a relatively long half-life, to uninfected babies born of infected mothers decreases the incidence of infection via breast feeding.

TRANSMISSION BY OTHER BODY FLUIDS

There is no convincing evidence that saliva can transmit HIV infection, either through kissing or through other exposures, such as occupationally to health care workers. HIV can be isolated from saliva of only a small proportion of infected individuals, typically in titers that are low compared to those in blood and genital secretions. In addition, saliva contains endogenous antiviral factors; among these factors, HIV-specific immunoglobulins of IgA, IgG, and IgM isotypes are detected readily in salivary secretions of infected individuals. It has been suggested that large glycoproteins such as mucins and thrombospondin-1 sequester HIV into aggregates for clearance by the host. In addition, a number of soluble salivary factors inhibit HIV to various degrees in vitro, probably by targeting host cell receptors rather than the virus itself. Perhaps the best-studied of these, secretory leukocyte protease inhibitor (SLPI), blocks HIV infection in several cell culture systems, and it is found in saliva at levels that approximate those required for inhibition of HIV in vitro. It has also been suggested that submandibular saliva reduces HIV infectivity by stripping gp120 from the surface of virions, and that saliva-mediated disruption and lysis of HIV-infected cells occurs because of the hypotonicity of oral secretions. There have been outlier cases of suspected transmission by saliva, but these have probably been blood-to-blood transmissions. One case was reported of a 91-year-old man who was bitten during a robbery attempt by an HIV-infected person. He seroconverted, and there was no question that the source of the infection was the human bite. However, the individual who bit him had bleeding gums, and it was thought that the infection was actually transmitted via blood. In addition, a most unusual form of HIV transmission from infected children to mothers in the former Soviet Union has been identified. In those cases, the children (infected through transfusion) were said to have bleeding sores in the mouth, and the mothers were said to have lacerations and abrasions on and around the nipples of the breast resulting from trauma from the children's teeth. Breast feeding had been continued until the children were older than is usual in other developed countries.

Although virus can be identified, if not isolated, from virtually any body fluid, there is no evidence that HIV transmission can occur as a result of exposure to tears, sweat, and urine. However, there have been isolated cases of transmission of HIV infection by body fluids that may or may not have been contaminated with blood. Most of these situations occurred in the setting of a close relative providing intensive nursing care for an HIV-infected person without observing universal precautions. These cases underscore the importance of observing universal precautions in the handling of body fluids and wastes from HIV-infected individuals (see below).

EPIDEMIOLOGY

HIV INFECTION AND AIDS WORLDWIDE

HIV infection/AIDS is a global pandemic, with cases reported from virtually every country. The current estimate of the number of cases of HIV infection among adults worldwide is approximately 33 million, two-thirds of whom are in sub-Saharan Africa; 47% of cases are women. In addition, an estimated 1.3 million children under 15 are living with HIV/AIDS. The global distribution of these cases is illustrated in Fig. 309-9. According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), in 1999 alone there were an estimated 5.4 million new cases of infection worldwide (more than 15,000 new infections each day) and 2.8 million death from AIDS, making it the fourth leading cause of mortality worldwide. The estimated number of AIDS-related deaths worldwide through the year 2000 is illustrated inFig. 309-10. The HIV epidemic has occurred in "waves" in different regions of the world, each wave having somewhat different characteristics depending on the demographics of the country and region in question and the timing of the introduction of HIV into the population. As noted above, different subtypes, or clades, of HIV-1 are prevalent in different regions of the world (see above andFig. 309-7), increasing the difficulty in the development of vaccines and perhaps accounting for different degrees of virulence. It is unlikely that a single vaccine will be applicable to all regions of the world. In this regard, in addition to HIV-1 subtype B, the predominant subtype in the United States, HIV-1 subtypes A, AE, AG, C, D, and O have been detected in individuals in the United States, as might be expected given the degree of international travel that occurs.

Table 309-3 provides the statistics and demographic features of HIV/AIDS in different regions of the world. Although the epidemic was first recognized in the United States and shortly thereafter in western Europe, it very likely began in sub-Saharan Africa (see above), which has been particularly devastated by the epidemic, with the prevalance of infection in many cities in the double digits. According to the United Nations Population Division, by the year 2015 life expectancy in the nine countries in Africa with the highest HIV prevalence rates will fall, on average, 17 years. In certain sub-Saharan African countries such as Zimbabwe and Botswana, available seroprevalence data indicate >25% of the adult population aged 15 to 49 is HIV-infected. In addition, among high-risk individuals (e.g., commercial sex workers, patients attendingSTDclinics) who live in urban areas of sub-Saharan Africa, seroprevalence is now>50% in many countries. The epidemic in Asian countries, particularly India and Thailand, has lagged temporally behind that in Africa; however, the number of new cases in this region is accelerating rapidly, and the magnitude of the epidemic is projected to exceed that of sub-Saharan Africa in the early part of the twenty-first century. The estimated number of cases in China is still relatively small; however, the potential exists for a major expansion of the epidemic in that nation of over 1 billion people.

The major mode of transmission of HIV worldwide is unquestionably heterosexual sex; this is particularly true and has been so since the begining of the epidemic in developing countries, where the numbers of infected men and women are approximately equal. The epidemic in most developed countries was first introduced among homosexual men and, to a greater or lesser degree (depending on the individual country), amongIDUs. In this regard, the total numbers of AIDS cases in those countries still reflect a high proportion of cases among these high-risk groups. However, in most developed countries, including the United States (see below), there has been a gradual shift such that among new cases of AIDS, there is a greater prevalence among heterosexuals and IDUs than among homosexual men.

AIDS IN THE UNITED STATES

AIDS has had and will continue to have an extraordinary public health impact in the United States. As of January 1, 2000, >724,600 cumulative cases of AIDS had been reported in adults and adolescents in the United States (Table 309-4) and approximately 425,000 AIDS-related deaths had been reported. It is the fifth leading cause of death among Americans aged 25 to 44 (Fig. 309-11), having dropped from first within the past few years. The death rate from AIDS declined 42% from 1996 to 1997 and 18% from 1997 to 1998. This trend is due to several factors including the improved prophylaxis and treatment of opportunistic infections, the growing experience among health professions in caring for HIV-infected individuals, improved access to health care, and the decrease in infections due to saturational effects and prevention efforts. However, the most influential factor clearly has been the increased use of potent antiretroviral drugs, generally administered in a combination of three or four agents, usually including a protease inhibitor (see below). When one looks at the totality of data collected from the beginning of the epidemic, approximately one-half of cases are among men who have had sex with men. However, over the past few years, the numbers of newly reported cases of AIDS among other groups, includingIDUs and heterosexuals, have surpassed the numbers of newly reported cases among men who have had sex with men. The proportion of new cases of AIDS per year attributed to heterosexual contact has increased dramatically over the past 15 years in the United States (Fig. 309-8). Women are increasingly affected; the proportion of AIDS cases in the United States reported among adult and adolescent females has increased from <5% to 24% from 1985 to 1998 (Fig. 309-8). Most cases of transmission by injection drug use and heterosexual contact are reported from the northeast and southeast regions of the country, particularly among minorities. HIV infection and AIDS have disproportionately affected minority populations in the United States. The rates of AIDS cases per 100,000 population reported among adults and adolescents in 1999 were 84.2 for African Americans, 34.6 for Hispanics, 9.0 for whites, 11.3 for American Indians/Alaska Natives, and 4.3 for Asian/Pacific Islanders (Fig. 309-12).

As of January 1, 2000, 8718 cases of AIDS in children <13 years old had been reported, and approximately 60% of these children have died (Table 309-5). Approximately 90% of these children were born to mothers who were HIV-infected or who were at risk for HIV infection and, in approximately 60% of those cases, the mother was either anIDU or the heterosexual partner of an IDU. About 42% of women with AIDS have become infected through injection drug use, compared to 22% of men with AIDS; 40% of women have become infected by heterosexual contact, compared to 4% of men with AIDS. Only 1% of AIDS cases are among hemophiliacs, and 1% are among recipients of blood transfusions, blood products, or transplanted tissue. The relative contribution of the latter groups will gradually decrease, even though individuals infected previously through this mode of transmission will continue to develop AIDS. The risk of additional infections via this mode of transmission in the United States is extremely small (see above). In recent years, the incidence of AIDS has decreased considerably, with ~46,000 new cases in 1999 compared to ~60,000 in 1996. This trend likely reflects both reduced infection rates since the mid-1980s; more widespread use of prophylactic therapies, which delay the onset of AIDS; and the use of highly effective antiretroviral therapy early in the course of HIV infection (see below). Also, the demography of newly infected individuals has changed considerably since the mid-1980s (see below).

HIV PREVALENCE AND INCIDENCE IN THE UNITED STATES

It is estimated that between 650,000 and 900,000 adults and adolescents in the United States are living with HIV infection, including 120,000 to 160,000 women. This estimate results in an overall nationwide prevalence of HIV infection of approximately 0.3%. Prevalence is highest among young adults in their late twenties and thirties and among minorities. An estimated 3% of black men and 1% of black women in their thirties are living with HIV infection. The number of new infections per year is estimated to be approximately 40,000, and this number has remained stable for at least 9 years. The estimated proportion of HIV infections has declined among white males, especially those >30, while the proportion of new HIV infections appears to have increased among women and minorities. Among newly infected persons in the United States, ~70% are men and ~30% are women (Fig. 309-13). Of these newly infected individuals, half are <25 years. Of new infections among men, the CDC estimates that ~60% were infected through homosexual sex, 25% through injection drug use, and 15% through heterosexual sex. Of new infections among women, ~75% were infected through heterosexual sex and 25% through injection drug use.

HIV infection and AIDS are widespread in the United States; although the epidemic on the whole is plateauing, it is spreading rapidly among certain populations, stabilizing in others, and decreasing in others. Similar to otherSTDs, HIV infection will not spread homogeneously throughout the population of the United States. However, it is clear that anyone who practices high-risk behavior is at risk for HIV infection. In addition, the alarming increase in infections and AIDS cases among heterosexuals (particularly sexual partners ofIDUs, women, and adolescents) as well as the spread in certain inner city areas (particularly among underserved minority populations with inadequate access to health care) testifies to the fact that the epidemic of HIV infection in the United States is a public health problem of major proportions.

PATHOPHYSIOLOGY AND PATHOGENESIS

The hallmark of HIV disease is a profound immunodeficiency resulting primarily from a progressive quantitative and qualitative deficiency of the subset of T lymphocytes referred to as *helper T cells*, or *inducer T cells*. This subset of T cells is defined phenotypically by the presence on its surface of the CD4 molecule (Chap. 305), which serves as the primary cellular receptor for HIV. A co-receptor must also be present together with CD4 for efficient fusion and entry of HIV-1 into its target cells (Figs. 309-3 and309-4). HIV uses two major co-receptors for fusion and entry; these co-receptors are also the primary receptors for certain chemoattractive cytokines termed *chemokines* and belong to the seven-transmembrane-domain G protein-coupled family of receptors. CCR5 and CXCR4 are the major co-receptors used by HIV (see above and below). Although a number of mechanisms responsible for cytopathicity and immune dysfunction of CD4+ T cells have been demonstrated in vitro, particularly direct infection and destruction of these cells by HIV (see below), it remains unclear which mechanisms or combination of mechanisms are primarily responsible for their progressive depletion and functional impairment in vivo. When the number of CD4+ T cells declines below a certain level (see below), the patient is at high risk of developing a variety of opportunistic diseases, particularly the infections and neoplasms that are AIDS-defining illnesses. Some features of AIDS, such asKS and neurologic abnormalities (see below), cannot be explained completely by the immunosuppressive effects of HIV, since these complications may occur prior to the development of severe immunologic impairment.

The combination of viral pathogenic and immunopathogenic events that occurs during the course of HIV disease from the moment of initial (primary) infection through the development of advanced-stage disease is complex and varied. It is important to appreciate that the pathogenic mechanisms of HIV disease are multifactorial and multiphasic and are different at different stages of the disease. Therefore, it is essential to consider the typical clinical course of an untreated HIV-infected individual in order to more fully appreciate these pathogenic events (Fig. 309-14).

PRIMARY HIV INFECTION, INITIAL VIREMIA, AND DISSEMINATION OF VIRUS

The events associated with primary HIV infection are likely critical determinants of the subsequent course of HIV disease. In particular, the dissemination of virus to lymphoid organs is a major factor in the establishment of a chronic and persisent infection (see below). The initial infection of susceptible cells may vary somewhat with the route of infection. Virus that enters directly into the bloodstream via infected blood or blood products (i.e., transfusions, use of contaminated needles for injecting drugs, sharp-object injuries, maternal-to-fetal transmission either intrapartum or perinatally, or sexual intercourse where there is enough trauma to cause bleeding) is likely cleared from the circulation to the spleen and other lymphoid organs, where it replicates to a critical level and then leads to a burst of viremia that disseminates virus throughout the body. It is uncertain which cell in the blood or lymphoid tissue is the first to actually become infected; however, studies in animal models suggest that dendritic lineage cells may be the initial cells infected. Depending on their stage of maturation, dendritic cells can either be directly infected with virus and pass virus on to CD4+ T cells or physically bring the virus into contact with CD4+ T cells without themselves becoming infected. Studies in the monkey model of mucosal exposure toSIVstrongly suggest that the initial cell to become infected at the site of exposure is the Langerhans cell, which is a dendritic lineage cell, and that this cell passes the virus on to CD4+ T cells in the draining lymph nodes. This mechanism likely operates in humans when HIV enters "locally" (as opposed to directly into the blood), via the vagina, rectum, or urethra during intercourse or via the upper gastrointestinal tract from swallowed infected semen, vaginal fluid, or breast milk. Certainly, CD4+ T cells and to a lesser extent cells of monocyte lineage are the major ultimate targets of HIV infection. In primary HIV infection, virus replication in CD4+ T cells intensifies prior to the initiation of an HIV-specific immune response (see below), leading to a burst of viremia (Fig. 309-14) and then to a rapid dissemination of virus to other lymphoid organs, the brain, and other tissues. Individuals who experience the "acute HIV syndrome," which occurs to varying degrees in approximately 50% of individuals with primary infection, have high levels of viremia that last for several weeks (see below). The acute mononucleosis-like symptoms are well correlated with the presence of viremia. Virtually all patients appear to develop some degree of viremia during primary infection, which contributes to virus dissemination, even though they remain asymptomatic or do not recall experiencing symptoms. Careful examination of lymph nodes from more than one site in patients with established HIV infection who did not report symptoms of a primary infection strongly indicate that wide dissemination to lymphoid tissue occurs in most patients. A more detailed description of the role of lymphoid tissue in the immunopathogenesis of HIV disease is given below. It appears that the initial level of plasma viremia in primary HIV infection does not necessarily determine the rate of disease progression; however, the set point of the level of steady-state plasma viremia after approximately 1 year does seem to correlate with the rapidity of disease progression (see below).

ESTABLISHMENT OF CHRONIC AND PERSISTENT INFECTION

Persistent Virus Replication HIV infection is relatively unique among human viral infections. Despite the robust cellular and humoral immune responses that are mounted following primary infection (see below), once infection has been established the virus is virtually never cleared completely from the body. Rather, a chronic infection develops that persists with varying degrees of virus replication for a median of approximately 10 years before the patient becomes clinically ill (see below). It is this establishment of a chronic, persistent infection that is the hallmark of HIV disease. Throughout the often protracted course of chronic infection, virus replication can almost invariably be detected in untreated patients, both by highly sensitive assays for plasma viremia as well as by demonstration of virus replication in lymphoid tissue. In human viral infections, with very few exceptions, if the host survives, the virus is completely cleared from the body and a state of immunity against subsequent infection develops. HIV infection very rarely kills the host during primary infection. Certain viruses, such asHSV(Chap. 182), are not completely cleared from the body after infection, but instead enter a latent state; in these cases, clinical latency is accompanied by microbiologic latency. This is not the case with HIV infection, in which some degree of virus replication invariably occurs during the period of clinical latency (see below). Chronicity associated with persistent virus replication can also be seen in certain cases of hepatitis B and C infections (Chap. 297); however, in these infections the immune system is not a target of the virus. As mentioned above, HIV usually does not abruptly kill the host; rather it generally succeeds in escaping from a rather vigorous immune response and establishing a state of chronic infection with varying degrees of persistently active virus replication.

Evasion of Immune System Control Clearly, HIV successfully evades elimination by the immune system in order to establish chronicity. The mechanisms whereby this occurs are not completely clear; however, several have been proposed as playing a role in this phenomenon. HIV has an extraordinary ability to mutate, but this mechanism probably acts mainly after the establishment of chronic infection and contributes to the maintenance of chronicity. Since the transmitted virus and the virus that initially becomes established as a chronic infection are relatively homogeneous, the initial escape from immune system control likely involves mechanisms other than viral mutation. Molecular analysis of clonotypes has demonstrated that clones of CD8+ cytolytic T lymphocytes (CTLs) that expand greatly during primary HIV infection and likely represent the high-affinity clones that would be expected to be most efficient in eliminating virus-infected cells are no longer detectable after their initial burst of expansion. The marked diminution of frequency or disappearance of these HIV-specific cells cannot be explained by mutations in the viral epitope to which they are directed, since virus-sequencing studies indicate that the initial viral epitope is still present when the clones are no longer detected. Furthermore, other, less expanded clones of CD8+ T cells that recognize the same viral epitope persist and likely account for the partial control of virus replication. It is thought that the initially expanded clones may have been deleted owing to the overwhelming exposure to viral antigens during the initial burst of viremia, similar to the exhaustion of CD8+ CTLs that has been reported in the murine model of lymphocytic choriomeningitis virus (LCMV) infection. To compound this phenomenon, virus replication and thus saturation of antigen-presenting cells with viral antigen take place in the lymphoid tissue (see below), which is also the site of generation of HIV-specific CTLs.

Another potential mechanism of HIV escape is related to the fact that, during primary HIV infection and the transition to established chronic infection, both activated HIV-specificCTLs and CTL precursors are preferentially and paradoxically segregated in the peripheral blood, where very little active virus replication takes place, rather than in the lymphoid tissue, which is the main site of virus replication and spread, and the major source of plasma viremia. Finally, the escape of HIV from elimination during primary infection allows the formation of a large pool of latently infected cells that cannot be eliminated by virus-specific CTLs (see below). Thus, despite a potent immune response and the marked downregulation of virus replication following primary HIV infection, HIV succeeds in establishing a state of chronic infection with a variable degree of persistent virus replication. In most cases, during this period the patient makes the clinical transition from acute primary infection to a relatively prolonged state of clinical latency (see below).

Reservoir of Latency Infected Cells It has been clearly demonstrated that there exists in virtually all HIV-infected individuals a pool of latently infected, resting CD4+ T cells, and that this pool of cells likely serves as at least one component of the persistent reservoir of virus. Such cells manifest postintegration latency in that the HIV provirus integrates into the genome of the cell and can remain in this state until an activation signal drives the expression of HIV transcripts and ultimately replication-competent virus. This form of latency is to be distinguished from preintegration latency, in which HIV enters a resting CD4+ T cell and, in the absence of an activation signal, only a limited degree of reverse transcription of the HIV genome occurs. This period of preintegration latency may last hours to days, and if no activation signal is delivered to the cell, the proviral DNA loses its capacity to initiate a productive infection. If these cells do become activated, reverse transcription proceeds to completion and the virus continues along its replication cycle (see above and Fig. 309-15). The pool of cells that are in the postintegration state of latency are established early during the course of primary HIV infection. Despite the suppression of plasma viremia to below detectable levels (<50 copies of HIV RNA per milliliter) by potent combinations of several antiretroviral drugs for as long as 3 years, this pool of latently infected cells persists and can give rise to replication-competent virus. This persistent pool of latently infected cells is a major obstacle to any goal of eradication of virus from infected individuals.

Viral Dynamics It was originally thought that very little virus replication occurred during clinical latency. However, studies of lymphoid tissue usingPCRanalysis for HIV RNA and in situ hybridization for individual virus-expressing cells clearly demonstrated that HIV replication occurs throughout the course of HIV infection, even during clinical latency when it is very difficult to culture virus from unfractionated peripheral blood mononuclear cells. The availability of sensitive PCR techniques led to the demonstration that some degree of plasma viremia is present in virtually all untreated patients at all stages of HIV disease. Subsequently, the dynamics of viral production and turnover were quantified using mathematical modeling in the setting of the administration of reverse transcriptase and protease inhibitors to HIV-infected individuals in clinical studies. Treatment with these drugs resulted in a precipitous decline in the level of plasma viremia, which typically fell by 99% within 2 weeks. The number of CD4+ T cells in the blood increased concurrently, which implies that the killing of CD4+ T cells is linked directly to the levels of replicating virus. However, it is generally agreed that a significant component of the early rise in CD4+ T cell numbers following the initiation of therapy is due to the redistribution of cells into the peripheral blood from other body compartments. It was determined on the basis of the emergence of resistant mutants during therapy that 93 to 99% of the circulating virus originated from recently infected, rapidly turning over CD4+ T cells and that approximately 1 to 7% of circulating virus originated from longer-lived cells, likely monocyte/macrophages. A negligible amount of circulating virus originated from the pool of latently infected cells (see above) (Fig. 309-16). It was also determined that the half-life of a circulating virion was approximately 30 min and that of productively infected cells was 1 day. Given the relatively steady level of plasma viremia and of infected cells, it appears that extremely large amounts of virus (approximately 10_{10} to 10_{11}virions) are produced and cleared from the circulation each day. In addition, data suggest that the minimum duration of the HIV-1 replication cycle in vivo averages 1.5 days. Other studies have demonstrated that the decrease in plasma viremia that results from antiretroviral therapy correlates closely with a decrease in virus replication in lymph nodes, further confirming that lymphoid tissue is the main site of HIV replication and the main source of plasma viremia. Using a mathematical formula that assumed a two-phase decay of virus-infected cells, it was originally estimated that virus could be eradicated within 2.3 to 3.1 years from an HIV-infected individual who was receiving antiretroviral therapy that successfully suppressed all virus replication. However, recent data taking into account the pool of latently infected cells (see above) indicate that there is a third, much longer phase of decay that results in a projected time to viral eradication ranging from 10 to 60 years. Concomitant with this finding was the realization that even the most potent combinations of antiretroviral drugs did not completely suppress virus replication, as indicated by the detection of variable degress of cell-associated HIV RNA by sensitive PCR assays in most patients despite the absence of detectable plasma virema. Therefore, it is highly unlikely that virus will be eradicated from HIV-infected individuals with the currently available antiretoviral drugs despite the favorable clinical outcomes that have resulted from such therapy (see below).

The level of steady-state viremia, called the viral *set point*, at approximately 1 year has important prognostic implications for the progression of HIV disease. It has been demonstrated that HIV-infected individuals who have a low set point at 6 months to 1 year progress to AIDS much more slowly than individuals whose set point is very high at that time (Fig. 309-17). Levels of viremia generally increase as disease progresses. Measurement of the level of viremia is playing an increasingly important role in guiding therapeutic decisions in HIV-infected individuals (see below).

Immunopathogenic Events during Clinical Latency With few exceptions, the level of CD4+ T cells in the blood decreases gradually and progressively in HIV-infected individuals. The slope of this decline, together with the level of plasma viremia (see above), predict well the pattern of the clinical course and the development of advanced disease. Most patients are entirely asymptomatic while this progressive decline is taking place (see below) and are often described as being in a state of *clinical latency*. However, clinical latency does not mean disease latency, since progression is generally relentless during this period. Furthermore, clinical latency should not be confused with microbiologic latency. Although there are cells present in an infected individual that are latently infected and do not express detectable viral RNA, there is virtually always some degree of ongoing virus replication, even during the early stages of HIV disease.

ADVANCED HIV DISEASE

In untreated patients or in patients in whom therapy has not adequately controlled virus replication (see below), after a variable period, usually measured in years, the CD4+ T cell count falls below a critical level (<200 cells per microliter), and the patient becomes highly susceptible to opportunistic disease (Fig. 309-14). For this reason, theCDC case definition of AIDS was modified to include all HIV-infected individuals with CD4+ T cell counts below this level (Table 309-1). Patients may experience constitutional signs and symptoms or may develop an opportunistic disease abruptly without any prior symptoms, although the latter scenario is unusual. The depletion of CD4+ T cells continues to be progressive and unrelenting in this phase. It is not uncommon for CD4+ T cell counts to drop as low as 10/uL or even to zero, yet the patients may survive for months or even for >1 year. This situation has become increasingly common as patients are treated more aggressively and are given prophylaxis against the common life-threatening opportunistic infections (see below). In addition, control of plasma viremia by antiretroviral therapy, even in individuals with extremely low CD4+ T cell counts, has increased survival in these patients despite the fact that their CD4+ T cell counts may not significantly increase as a result of therapy. Ultimately, patients who progress to this severest form of immunodeficiency usually succumb to opportunistic infections or neoplasms (see below).

LONG-TERM SURVIVORS AND LONG-TERM NONPROGRESSORS

The median time from primary HIV infection to the development of AIDS in untreated individuals is approximately 10 years. Treatment with effective combinations of antiretroviral drugs has clearly extended this period; the full extent of this benefit has yet to be realized. The definitions of *long-term survivor* and *long-term nonprogressor* continue to evolve as more data are collected from prospective cohort studies. Predictions from one study that antedated the availability of effective antiretroviral therapy estimated that approximately 13% of homosexual/bisexual men who were infected at an early age may remain free of clinical AIDS for >20 years. Currently, individuals are considered to be long-term survivors if they remain alive for 10 to 15 years after initial infection. In most such individuals the disease has progressed, in that they have significant immunodeficiency, and many have experienced opportunistic diseases. Some of these individuals have CD4+ T cell counts that have decreased to £200/uL but have remained stable at that level for years. The mechanisms of this stabilization are not entirely clear but may relate to the beneficial effects of antiretroviral therapy and prophylaxis against opportunistic infections. In addition, a number of viral and/or host determinants likely contribute to the long-term survival of these individuals. In some individuals, the virus may either have been less virulent initially or may have mutated to a less virulent form under the influence of antiretroviral therapy. Quantitative and qualitative aspects of the HIV-specific immune response, as well as recognized and unrecognized genetic factors (see below), may also contribute to the long-term survival of these individuals.

Fewer than 5% of HIV-infected individuals are characterized as long-term nonprogressors. All long-term nonprogressors are long-term survivors; however, the reverse is not true. Individuals who have been infected with HIV for a long period (10 years), whose CD4+ T cell counts are in the normal range and have remained stable over years, and who have not received antiretroviral therapy are considered to be long-term nonprogressors. These patients are characterized by a low viral burden (low number of HIV-infected cells), low levels of plasma viremia, generally normal immune function according to commonly measured parameters (skin tests, in vitro lymphocyte responses to various mitogens and antigens), and normal-appearing lymphoid tissue architecture as determined on lymph node biopsy. In general, long-term nonprogressors manifest robust HIV-specific immune responses, both humoral (neutralizing antibodies) and cell-mediated (HIV-specificCTLs). However, this may also be true of some individuals early in the course of disease who ultimately progress to advanced disease. Although viremia is consistently very low in long-term nonprogressors, many have persistent viremia as determined by sensitivePCRassays. No qualitative abnormalities in the virus have been detected in most of these patients. However, a small subset of patients do have defective virus; in particular, in one cohort of five long-term nonprogressors, the virus had a defect in the *nef* gene. In another report, a blood donor in Australia who was HIV-infected and a group of seven individuals who were infected by blood or blood products from that donor remained free of HIV-related disease and maintained normal and stable CD4+ T cell counts for several years after infection. Sequence analysis of viruses isolated from the donor and recipients revealed similar deletions in the *nef* gene and the region of overlap of *nef* and the U3 region of the HIV long terminal repeat (Fig. 309-5). However, several of these individuals have now begun to show indications of progressive immunodeficiency, and thus they can no longer be considered nonprogressors. The precise role of host factors in long-term nonprogression remains unclear. There is no obvious and consistent genetic determinant for nonprogression. However, several genetic mutations have been demonstrated to result in a delay in the progression of HIV disease. These include heterozygosity for the *CCR5*-D32 deletion, heterozygosity for the *CCR2*-64I mutation, homozygosity for the *SDF1*-3¢A mutation, and heterozygosity for the *RANTES-28G* mutation (see "Genetic Factors in HIV Pathogenesis," below). Since CCR5 is the major co-receptor for R5 or macrophage-tropic strains of HIV and since individuals who are homozygous for the *CCR5*-D32 deletion are, with rare exceptions, protected against HIV infection, the potential mechanism for slow progression in heterozygotes is clear. In addition, certain single nucleotide polymorphisms in the *CCR5* promoter have been shown to be associated with slower progression of disease. The reason for the slowing of progression of HIV disease in individuals who are heterozygous for the *CCR2*-64I mutation is less clear; however, it has been demonstrated that CXCR4 can dimerize with the CCR2V64I mutant but not with wild-type CCR2. This dimerization may reduce the amount of CXCR4 on the cell surface and as a result inhibit infection with X4 viruses. Homozygosity for the *SDF1*-3¢A mutation may upregulate the *SDF1* gene enabling SDF-1, which is the natural ligand for CXCR4, to compete more effectively with X4 or T cell tropic virus for the CXCR4 coreceptor. The *RANTES-28G* mutation increases RANTES expression, which is the natural ligand for CCR5 and may thus inhibit infection with R5 viruses. Finally, maximal HLA heterozygosity of class I loci (A, B, and C) has been shown to be associated with delayed progression of HIV disease. Although long-term nonprogressors have robust HIV-specific immune responses as well as competent CD8+ T cell suppressors of HIV replication, it is unclear whether these factors are directly responsible for the state of nonprogression. A substantial proportion of HIV-infected individuals manifest comparable immune responses early in the course of their disease and still experience disease progression. Long-term nonprogressors likely represent a heterogeneous group. The lack of disease progression may be explained in some by a defect in the virus; in others by any of a variety of host factors, including recognized and as yet unrecognized genetic factors; and in others by a combination of both.

ROLE OF LYMPHOID ORGANS IN HIV PATHOGENESIS

Lymphoid tissues are the major anatomic sites for the establishment and propagation of HIV infection (see above). For practical reasons, most studies on the pathogenesis of HIV infection have focused on peripheral blood mononuclear cells. However, lymphocytes in the peripheral blood represent only approximately 2% of the total body lymphocyte pool and so may not always accurately reflect the status of the entire immune system; most of the body's lymphocytes reside in lymphoid organs, such as the lymph nodes, spleen, and gut-associated lymphoid tissue. Furthermore, virus replication occurs mainly in lymphoid tissue and not in blood; the level of plasma viremia reflects virus production in lymphoid tissue. Finally, since HIV disease is an infectious disease of the immune system, it is critical to appreciate the pathogenic events that occur in the lymphoid tissue in HIV infection.

Some patients experience progressive generalized lymphadenopathy (see below) early in the course of the infection; others experience varying degrees of transient lymphadenopathy. Lymphadenopathy reflects the cellular activation and immune response to the virus in the lymphoid tissue, which is generally characterized by follicular or germinal-center hyperplasia. Lymph node involvement is a common denominator of virtually all patients with HIV infection, even those without easily detectable lymphadenopathy.

Simultaneous examination of lymph node and peripheral blood in the same patients during various stages of HIV disease, including the early asymptomatic stage (when CD4+ T cell counts generally are >500/uL), the intermediate stage (when counts are usually 200 to 500/uL) and the advanced stage (when counts are<200/uL) has led to substantial insight into the pathogenesis of HIV disease. Using a combination ofPCRtechniques for HIV DNA and RNA in tissue and RNA in plasma, in situ hybridization for HIV RNA, and light and electron microscopy, the following picture has emerged. In most untreated patients, early in the course of infection when the viral set point has been reached and prior to significant immunodeficiency (CD4+ T cell counts >500/uL), levels of plasma viremia are variable but generally low; the viral burden (number of infected cells) in the peripheral blood is usually extremely low, and expression of HIV in these cells is minimal or undetectable. Remarkably, at this time copious amounts of extracellular virions are trapped on the processes of the follicular dendritic cells (FDCs) in the germinal centers of the lymph nodes (Fig. 309-18*A*). In situ hybridization reveals expression of virus in individual cells of the paracortical area and, to a lesser extent, the germinal center (Fig. 309-18*B*). The number of cells expressing virus is low early in the course of disease and increases as disease progresses. Examination of lymph nodes during primary HIV infection in humans (see above) and

SIVinfection in macaques indicates that during the transition from primary infection to established chronic infection, germinal centers form and virus is trapped. This trapping, together with the generation of a vigorous HIV-specific immune response, likely contributes to the rapid decrease in plasma viremia seen in most patients following the initial burst of viremia associated with primary infection. A considerable amount of virus can be trapped during the period of high viremia associated with primary infection. The persistence of trapped virus after chronic infection likely reflects a steady state whereby trapped virus turns over and is replaced by fresh virions, which are produced persistently, albeit usually at low levels during the early, clinically latent stage of disease.

During early-stage HIV disease, the architecture of the germinal centers is generally preserved and may even be hyperplastic owing to in situ proliferation of cells (mostly B lymphocytes) and recruitment to the lymph nodes of a number of cell types (B cells, CD4+ and CD8+ T cells). Electron microscopy demonstrates a fine network ofFDCswith many long, finger-like processes that envelop virtually every lymphocyte in the germinal center (Fig. 309-18*C*). Extracellular virions can be seen attached to the processes, yet the FDCs appear to be relatively healthy. The trapping of antigen is a physiologically normal function for the FDCs, which present antigen to B cells and contribute to the generation of B cell memory. However, in the case of HIV, the trapped virions serve as a persistent source of cellular activation, resulting in the secretion of proinflammatory cytokines such as interleukin (IL) 1b, tumor necrosis factor (TNF)a, and IL-6, which can upregulate virus replication in infected cells (see below). Furthermore, although trapped virus is coated by neutralizing antibodies, it has been demonstrated that these virions remain infectious for CD4+ T cells while attached to the processes of the FDCs. CD4+ T cells that migrate into the germinal center to provide help to B cells in the generation of an HIV-specific immune response thus are susceptible to infection by these trapped virions. Thus, in HIV infection, a normal physiologic function of the immune system, which contributes to the clearance of virus as well as to the generation of a specific immune response, can also have deleterious consequences. It is difficult to demonstrate infection of the FDCs at this point, or even in advanced disease; however, rare examples of virus budding off FDCs have been reported.

As the disease progresses, the architecture of the germinal centers begins to show disruption, and the trapping efficiency of the lymph node diminishes. Electron microscopy reveals swollen organelles, and theFDCsbegin to undergo cell death. The mechanisms of FDC death remain unclear; there is no indication by electron microscopy of copious virus replication or budding of virions off the cell in great quantities. At this stage, the level of plasma viremia generally increases. In addition, both the relative number of infected cells in the blood and the expression of virus from these cells increases, approaching the levels in the lymph nodes. As the disease progresses to an advanced stage, there is complete disruption of the architecture of the germinal centers, accompanied by dissolution of the FDC network and massive dropout of FDCs (Fig. 309-18*D*). The trapping function of the lymph nodes is completely lost, and virus freely spills out into the circulation. SimultaneousPCRanalysis of lymph node and peripheral blood mononuclear cells indicates that the relative number of infected cells in the blood and their expression of virus begin to equal the levels in the lymph nodes at this stage. Advanced disease is accompanied by high levels of plasma viremia, which represent a true increase in virus replication, due in part to a further diminution of immune control of virus replication (see below) as well as to the loss of the mechanical trapping function of the lymph nodes. At this point, the lymph nodes are "burnt out." This destruction of lymphoid tissue compounds the immunodeficiency of HIV disease and contributes to the inability to mount adequate immune responses against opportunistic pathogens. The events from primary infection to the ultimate destruction of the immune system are illustrated inFig. 309-19.

ROLE OF CELLULAR ACTIVATION IN HIV PATHOGENESIS

The immune system is normally in a state of homeostasis, awaiting perturbation by foreign antigenic stimuli. Activation of the immune system is an essential component of an appropriate immune response to a foreign antigen. Once the immune response deals with and clears the antigen, the system returns to relative quiescence (Chap. 305). In HIV infection, however, the immune system is chronically activated owing to the chronicity of infection and the persistence of virus replication (see above). This activated state is reflected by hyperactivation of B cells leading to hypergammaglobulinemia; spontaneous lymphocyte proliferation; activation of monocytes; expression of activation markers on CD4+ and CD8+ T cells; lymph node hyperplasia, particularly early in the course of disease (see above); increased secretion of proinflammatory cytokines (see below); elevated levels of neopterin,b_2-microglobulin, acid-labile interferon, and solubleIL-2 receptors; and autoimmune phenomena (see below). Even in the absence of direct infection of a target cell, HIV envelope proteins can interact with cellular receptors (CD4 molecules and chemokine receptors) to deliver potent activation signals resulting in calcium flux, the phosphorylation of certain proteins involved in signal transduction, co-localization of cytoplasmic proteins including those involved in cell trafficking, secretion of certain cytokines, immune dysfunction, and under certain circumstances, apoptosis (see below).

Persistent immune activation may have several deleterious consequences. From a virologic standpoint, although quiescent CD4+ T cells can be infected with HIV, reverse transcription, integration, and virus spread are much more efficient in activated cells. Furthermore, cellular activation induces expression of virus in cells latently infected with HIV (see above). From an immunologic standpoint, chronic exposure of the immune system to a particular antigen over an extended period may ultimately lead to an inability to sustain an adequate immune response to the antigen. Furthermore, the ability of the immune system to respond to a broad spectrum of antigens may be compromised if immune-competent cells are maintained in a state of chronic activation. In addition, activation of the immune system may favor the elimination of cells via programmed cell death (apoptosis) (see below) as well as the secretion of certain cytokines that can induce HIV expression (see below).

Role of Apoptosis *Apoptosis* is a form of programmed cell death that is a normal mechanism for the elimination of effete cells in organogenesis as well as in the cellular proliferation that occurs during a normal immune response (Chap. 305). Apoptosis is strictly dependent on cellular activation. It has been hypothesized that, in HIV infection, sequential activation signals delivered to CD4+ T cells induce apoptosis. Cross-linking of the CD4 molecule by gp120 or gp120/anti-gp120 complexes delivers the first of two signals required for apoptosis. The second signal supposedly leading to cell death is delivered via the T cell receptor by antigen. According to this hypothesis, direct infection of CD4+ T cells is not required for apoptosis to occur, although it has been demonstrated that alterations in tyrosine kinase activity of HIV-infected cells may induce the cell to undergo apoptosis. HIV can trigger both Fas-dependent and Fas-independent pathways of apoptosis. Mechanisms involved in this process include upregulation of Fas and Fas ligand, upregulation of caspase-1 and caspase-8, downregulation of the anti-apoptotic Bcl-2 protein, and activation of cyclin-dependent kinases. Certain viral gene products have been associated with enhanced susceptibility to apoptosis including envelope, Nef, and Vpu. A number of studies, including those examining lymphoid tissue, have demonstrated that the rate of apoptosis is elevated in HIV infection and that apoptosis is seen in "bystander" cells such as CD8+ T cells and B cells as well as in CD4+ T cells. Macrophages have been shown to mediate apoptosis of CD8+ T cells by a mechanism involving gp120-induced upregulation of Fas ligand expression on macrophages and enhanced secretion of macrophage-derivedTNF-a. The intensity of apoptosis correlates with the general state of activation of the immune system and not with the stage of disease or with viral burden. The potential role of apoptosis in the pathogenesis of HIV disease is underscored by results from animal studies that show an increased frequency of apoptosis in CD4+ T cells in primates infected with pathogenic strains ofSIV but not in primates infected with nonpathogenic strains of SIV. It is likely that apoptosis of immune-competent cells contributes to the immune abnormalities in HIV disease; however, this is probably a nonspecific mechanism that merely reflects the aberrant state of immune activation.

Autoimmune Phenomena The autoimmune phenomena that are common in HIV-infected individuals reflect, at least in part, chronic immune system activation as well as molecular mimickry by viral components. Although these phenomena usually occur in the absence of autoimmune disease, a wide spectrum of clinical manifestations that may be associated with autoimmunity have been described (see below). Autoimmune phenomena include antibodies to lymphocytes and, less commonly, to platelets and neutrophils. Antiplatelet antibodies have some clinical relevance, in that they may contribute to the thrombocytopenia of HIV disease (see below). Antibodies to nuclear and cytoplasmic components of cells have been reported, as have antibodies to cardiolipin; CD4 molecules; CD43 molecules, C1q-A; variable regions of the T cell receptor a,b, and g chains; Fas; denatured collagen; andIL-2. In addition, autoantibodies to a range of serum proteins, including albumin, immunoglobulin, and thyroglobulin, have been reported. There is antigenic cross-reactivity between HIV viral proteins (gp120 and gp41) andMHCclass II determinants, and anti-MHC class II antibodies have been reported in HIV infection. These antibodies could potentially lead to the elimination of MHC class II-bearing cells via antibody-dependent cellular cytotoxicity (ADCC) (Chap. 305). In addition, regions of homology exist between HIV envelope glycoproteins and IL-2 as well as MHC class I molecules.

Cofactors Contributing to HIV Pathogenesis Both endogenous and exogenous factors can contribute to HIV pathogenesis by a number of mechanisms; paramount among these is the upregulation of virus expression, a process intimately connected with cellular activation. The main endogenous factors that regulate HIV expression are cytokines (see below). Among exogenous factors, other microbes likely have important effects on HIV replication and HIV pathogenesis. They can thus be considered real or potential *cofactors* in the pathogenesis of HIV disease. Co-infection or simultaneous cotransfection of cells with HIV and other viruses or viral genes has demonstrated that certain viruses, such asHSV type 1, cytomegalovirus (CMV), human herpesvirus (HHV) 6, Epstein-Barr virus (EBV), hepatitis B virus (HBV), adenovirus, pseudorabies virus, andHTLV-I can upregulate HIV expression. Other microbes, such as *Mycoplasma* have been reported to contribute to the induction of HIV expression. *Mycobacterium tuberculosis* is a common opportunistic infection in HIV-infected individuals (see below and Chap. 169). In addition to the fact that HIV-infected individuals are more likely to develop activeTB after exposure, it has been demonstrated that active TB can accelerate the course of HIV infection. It has also been shown that levels of plasma viremia are greatly elevated in HIV-infected individuals with active TB, compared to pre-TB levels and levels of viremia after successful treatment of the active TB. In vitro studies demonstrated that virus replication was markedly enhanced in lymphocytes of HIV-infected individuals who were skin test-positive for purified protein derivative (PPD) when PPD antigen was added to culture, resulting in cellular activation. Confirmatory evidence that antigen-induced activation was a major contributor to the accelerated viremia in HIV-infected individuals with active TB was provided by studies in which HIV-infected individuals were immunized with common recall antigens such as tetanus toxoid, influenza, or pneumococcal polysaccharide. Under these circumstances, a transient elevation of plasma viremia accompanied the cellular activation induced by the immunization. A greater degree of induction of virus was seen in those individuals with early stage as opposed to advanced stage HIV disease, and the degree of virus induction correlated with the level of immune system activation.

THE CYTOKINE NETWORK IN HIV PATHOGENESIS

Cytokine Regulation of HIV Expression The immune system is homeostatically regulated by a complex network of immunoregulatory cytokines, which are pleiotropic and redundant and operate in an autocrine and paracrine manner. They are expressed continuously, even during periods of apparent quiescence of the immune system. On perturbation of the immune system by antigenic challenge, the expression of cytokines increases to varying degrees (Chap. 305). Cytokines that are important components of this immunoregulatory network have been demonstrated to play a major role in the regulation of HIV expression in vitro. A number of in vitro model systems of chronically infected monocyte or T cell lines, primary cultures of peripheral blood or lymph node mononuclear cells from HIV-infected individuals, and acutely infected primary cell cultures have been used to demonstrate the role of cytokines in the regulation of HIV expression. Potent modulation of HIV expression has been demonstrated either by manipulating endogenous cytokines or by adding exogenous cytokines to culture. Cytokines that induce HIV expression in one or more of these systems includeIL-1, IL-2, IL-3, IL-6, IL-12,TNF-a, and TNF-b, macrophage colony stimulating factor (M-CSF), and granulocyte-macrophage colony stimulating factor (GM-CSF). Among these cytokines, the most consistent and potent inducers of HIV expression are the *proinflammatory cytokines* TNF-a, IL-1b, and IL-6.IFN-aand -b suppress HIV replication, whereas transforming growth factor (TGF) b, IL-4, IL-10, and IFN-g can either induce or suppress HIV expression, depending on the system involved. The *CC-chemokines* RANTES, macrophage inflammatory protein (MIP) 1a, and MIP-1b(Chap. 305) inhibit infection by and spread of R5 (macrophage-tropic) HIV-1 strains, while *stromal cell-derived factor* (SDF) 1 inhibits infection by and spread of X4 (T cell-tropic) strains (see below). Several of these cytokines act synergistically in regulating HIV infection and replication, and others function in an autocrine and paracrine manner, similar to their physiologic function in the regulation of the immune system. Blocking of endogenous HIV-inducing cytokines or addition of inhibitors of HIV suppressor cytokines in cultures of peripheral blood and lymph node mononuclear cells from HIV-infected individuals has demonstrated that HIV replication is controlled tightly by endogenous cytokines acting in an autocrine and paracrine manner. Indeed, the net level of virus replication in an HIV-infected individual at least in part reflects a balance between inductive and suppressive host factors, mediated mainly by cytokines. An example of this endogenous regulation is the case of IL-10, which inhibits HIV replication in acutely infected monocyte/macrophages by blocking the secretion of the HIV-inducing cytokines TNF-a and IL-6. In addition, IL-4, IL-13, and TGF-b inhibit HIV expression in chronically infected monocytic cell lines stimulated by lipopolysaccharide and GM-CSF by increasing the ratio of expression of endogenous IL-1 receptor antagonist to IL-1b.

The molecular mechanisms of HIV regulation are best understood forTNF-a, which activates NF-kB proteins that function as transcriptional activators of HIV expression. The HIV-inducing effect ofIL-1bis thought to occur at the level of viral transcription in an NF-kB-independent manner. IL-6,GM-CSF, andIFN-gregulate HIV expression mainly by posttranscriptional mechanisms. Elevated levels of TNF-a and IL-6 have been demonstrated in plasma and cerebrospinal fluid (CSF), and increased expression of TNF-a, IL-1b, IFN-g, and IL-6 has been demonstrated in the lymph nodes of HIV-infected individuals. The mechanisms whereby the CC-chemokines RANTES,MIP-1a, and MIP-1b inhibit infection of R5 strains of HIV very likely involve blocking of the binding of the virus to its co-receptor, the CC-chemokine receptor CCR5 (see above and below). Of note is the fact that CC-chemokines that inhibit infection by R5 strains of virus actually enhance infection by X4 strains of virus by inducing intracellular signal transduction through the CCR5 and CD4 molecules. In addition, products of bacterial pathogens as well as of certain viruses including HIV-1 itself can induce the expression of CXCR4 and thus potentially favor infection with X4 strains of virus that utilize this co-receptor.

Dysregulation of Cytokines HIV-infected individuals show an imbalance in the T cell limbs of the immune response, which are defined by the patterns of cytokine secretion. T helper (TH)1 cells are characterized by secretion ofIL-2 andIFN-gand favor cell-mediated immune responses, whereas TH2 cells are characterized by secretion of IL-4, IL-5, and IL-10 and favor humoral immune responses (Chap. 305). Since several cell types in addition to CD4+ T cells secrete these cytokines, it is more accurate to refer to immune responses that reflect one or the other cytokine pattern as *TH1 or TH2 type responses*. HIV-infected individuals show a decrease in TH1 type responses relative to TH2 type cytokine patterns. They manifest a progressive loss in expression of the IL-2 receptor and in the ability to produce the immunoregulatory cytokines IL-2 and IL-12; these cytokines are critical for effective cell-mediated immune responses in that they stimulate proliferation and lytic activity ofCTLs and natural killer (NK) cells. Furthermore, IL-12 is important for the stimulation of TH1 type cytokines such as IL-2 and IFN-g that favor the development of cell-mediated immune responses. TH1 type cytokines such as IL-2, IL-12,TNF-a, and IFN-g upregulate CCR5 expression, while TH2 type cytokines such as IL-4 and IL-10 upregulate CXCR4 expression and downregulate CCR5 expression. It has also been demonstrated that in vitro apoptosis can be inhibited in T cells from HIV-infected donors by antibodies to IL-4 and IL-10 and enhanced by antibodies to IL-12. Although it has been proposed that a clear-cut switch from a TH1 type to a $T_{H}2$ type of cytokine pattern is a critical step in the pathogenesis of HIV disease, no sharp dichotomy between these two types of cytokine patterns that is directly related to progression of disease has been corroborated. Cytokine dysregulation in HIV infection is complex and cannot be neatly classified in terms of the polarity of $T_{H}1$ and $T_{H}2$ responses.

CELLULAR TARGETS OF HIV

Although the CD4+ T lymphocytes and CD4+ cells of monocyte lineage are the principal targets of HIV, virtually any cell that expresses the CD4 molecule together with co-receptor molecules (see above and below) can potentially be infected with HIV. Circulating dendritic cells have been reported to express low levels of CD4, and depending on their stage of maturation, these cells can be infected with HIV (see below). Epidermal Langerhans cells express CD4 and have been infected by HIV in vivo. In vitro, HIV has been reported also to infect a wide range of cells and cell lines that express low levels of CD4, no detectable CD4, or only CD4 mRNA; among these areFDCs; megakaryocytes; eosinophils; astrocytes; oligodendrocytes; microglial cells; CD8+ T cells; B cells;NKcells; renal epithelial cells; cervical cells; rectal and bowel mucosal cells such as enterochromaffin, goblet, and columnar epithelial cells; trophoblastic cells; and cells from a variety of organs, such as liver, lung, heart, salivary gland, eye, prostate, testis, and adrenal gland. Since the only cells that have been shown unequivocally to be infected with HIV and to support replication of the virus are CD4+ T lymphocytes and cells of monocyte/macrophage lineage, the relevance of the in vitro infection of these other cell types is questionable.

Of potentially important clinical relevance is the demonstration that thymic precursor cells, which were assumed to be negative for CD3, CD4, and CD8 molecules, actually do express low levels of CD4 and can be infected with HIV in vitro. In addition, human thymic epithelial cells transplanted into an immunodeficient mouse can be infected with HIV by direct inoculation of virus into the thymus. Since these cells may play a role in the normal regeneration of CD4+ T cells, it is possible that their infection and depletion contribute, at least in part, to the impaired ability of the CD4+ T cell pool to completely reconstitute itself in certain infected individuals in whom antiretroviral therapy has suppressed viral replication to below detectable levels (<50 copies of HIV RNA per milliliter; see below). In addition, CD34+ monocyte precursor cells have been shown to be infected in vivo in patients with advanced HIV disease. It is likely that these cells express low levels of CD4, and therefore it is not essential to invoke CD4-independent mechanisms to explain the infection.

ROLE OF CO-RECEPTORS IN CELL TROPISM OF HIV

Different strains of HIV-1 utilize two major co-receptors along with CD4 to bind to, fuse with, and enter target cells; these co-receptors are CCR5 and CXCR4, which are receptors for certain chemokines and belong to the seven-transmembrane-domain G protein-coupled family of receptors (see above). Strains of HIV that utilize CCR5 as a co-receptor are referred to as *R5 viruses*. These viruses were formerly classified as *macrophage tropic viruses* since they readily infect macrophages but not T cell lines. Strains of HIV that utilize CXCR4 are referred to as *X4 viruses*. These viruses are also referred to as *T cell-tropic viruses* since they readily infect T cell lines but not macrophages. In actuality, X4 viruses enter macrophages but do not proceed efficiently along the replication cycle unless an appropriate signal is delivered to the cell. Many virus strains are *dual tropic* in that they utilize both CCR5 and CXCR4; these are referred to as *R5X4 viruses*. Other terminology that has been associated with R5 versus X4 viruses is *non-syncytium-inducing viruses* versus *syncytium-inducing viruses*, respectively, based on the observation that R5 viruses generally do not form syncytia in culture with certain T cell lines, whereas X4 viruses readily form syncytia. In reality, under certain conditions both R5 and X4 viruses are capable of forming syncytia in culture.

The natural chemokine ligands for the major HIV co-receptors can readily block entry of HIV. For example, the CC-chemokines RANTES,MIP-1a, and MIP-1b, which are the natural ligands for CCR5, block entry of R5 viruses, whereasSDF-1, the natural ligand for CXCR4, blocks entry of X4 viruses. The mechanism of inhibition of viral entry is a steric inhibition of binding that is not dependent on signal transduction (Fig. 309-20).

The transmitting virus is almost invariably an R5 virus that predominates during the early stages of HIV disease. In approximately 40% of HIV-infected individuals, there is a transition to a predominantly X4 virus that is associated with a relatively rapid progression of disease. However, at least 60% of infected individuals progress in their disease while maintaining predominance of an R5 virus. Other chemokine receptor family members may function as coreceptors for HIV andSIVentry, but to a much lesser extent than do CCR5 and CXCR4; these include CCR3, BOB/GPR15, Bonzo/STRL33/TYMSTR, CCR2, CCR8, CX_3CR1(V28), and GPR1.

The basis for the tropism of different envelope glycoproteins for either CCR5 or CXCR4 relates to the ability of the HIV envelope, particularly the third variable region (V3 loop) of gp120, to interact with these co-receptors. In this regard, binding of gp120 to CD4 induces a conformational change in gp120 that increases its affinity for CCR5. It appears that the interaction of gp120 with CXCR4 is less dependent on the conformational change induced in gp120 by CD4. In fact, there are X4 strains of HIV that bind to CXCR4 in the absence of surface-bound or soluble CD4. Finally, R5 viruses are more efficient in infecting monocyte/macrophages and microglial cells of the brain (see "Neuropathogenesis," below).

ABNORMALITIES OF MONONUCLEAR CELLS

CD4+ T Cells The range of T cell abnormalities in advanced HIV infection is broad. The defects are both quantitative and qualitative and involve virtually every limb of the immune system (see below), indicating the critical dependence of the integrity of the immune system on the inducer/helper function of CD4+ T cells. Virtually all of the immune defects in advanced HIV disease can ultimately be explained by the quantitative depletion of CD4+ T cells. However, T cell dysfunction (see below) can be demonstrated in patients early in the course of infection, even when the CD4+ T cell count is in the low-normal range. The degree and spectrum of dysfunctions increase as the disease progresses. One of the first abnormalities to be detected is a defect in response to remote recall antigens, such as tetanus toxoid and influenza, at a time when mononuclear cells can still respond normally to mitogenic stimulation. Defects in responses to soluble antigens are followed in time by the loss of T cell proliferative responses to alloantigens, and subsequently to mitogens. Essentially every T cell function has been reported to be abnormal at some stage of HIV infection. These abnormalities include defective T cell cloning and colony-forming efficiencies, impaired expression ofIL-2 receptors, defective IL-2 production, and decreasedIFN-gproduction in response to antigens. The proportion of CD4+ T cells that express CD28, which is a major co-stimulatory molecule necessary for the normal activation of T cells, is reduced during HIV infection. Cells lacking expression of CD28 do not respond to activation signals and may express markers of terminal activation including HLA-DR, CD38, and CD45RO. CD4+ T cells from HIV-infected individuals express abnormally low levels of CD40 ligand, which may explain the dysregulation of B cell function observed in HIV disease.

It is difficult to explain completely the profound immunodeficiency noted in HIV-infected individuals solely on the basis of direct infection and quantitative depletion of CD4+ T cells. This is particularly apparent during the early stages of HIV disease, when CD4+ T cell numbers may be only marginally decreased. Certainly, at the stage of advanced disease when the CD4+ T cell count is in the range of 0 to 50/uL, quantitative depletion alone can explain the immune defects. However, it is likely that CD4+ T cell dysfunction results from a combination of depletion of cells due to direct infection and a number of virus-related but indirect effects on the cell (Table 309-6).

Single-cell killing and the formation of syncytia between infected and uninfected cells have been demonstrated clearly in vitro, although the precise mechanisms of cell death in vivo have not been determined. Cytopathicity in an infected cell in vitro may result from a number of mechanisms, including copious budding of virions from the cell surface with resulting disruption of the integrity of the cell membrane; interference with cellular RNA processing or the accumulation of high levels of heterodisperse RNA molecules; disruption of cellular protein synthesis owing to high levels of viral RNA; accumulation of high levels of unintegrated viral DNA in the cell cytoplasm; induction of aberrant patterns of protein tyrosine phosphorylation; and the interaction between HIV gp120 and CD4 intracellularly. Strain differences in single-cell killing are determined largely by gp120 sequences, which supports the importance of the viral envelope in this process. *Syncytia formation* involves fusion of the cell membrane of an infected cell with the cell membranes of variable numbers of uninfected CD4+ cells. Although cell fusion has not been shown to be an important pathogenic process in vivo, a direct relationship between the presence of syncytia and the degree of cytopathic effect has been demonstrated in vitro, and a correlation has been reported between the presence of virus isolates that readily induce syncytia in vitro and a more aggressive clinical course in the patient. Efficient syncytia formation depends on the leukocyte adhesion molecule LFA-1 (Chap. 305) on human CD4+ T cells acutely infected with HIV in vitro.

Humoral and cellular immune responses to HIV may contribute to protective immunity by eliminating virus and virus-infected cells (see below). However, since the main targets of HIV infection are immune-competent cells, these responses may contribute to immune-cell depletion and immunologic dysfunction by eliminating both infected cells and "innocent bystander" cells. Soluble viral proteins, particularly gp120, can bind with high affinity to the CD4 molecules on uninfected T cells and monocytes; in addition, virus and/or viral proteins can bind to dendritic cells orFDCs. HIV-specific antibody can recognize these bound molecules and potentially collaborate in the elimination of the cells byADCC.

Nonpolymorphic determinants ofMHCclass I products share a degree of homology with gp120 and gp41 proteins of HIV. Such similarities may lead to the generation of autoantibodies to self-MHC determinants. In fact, anti-HLA-DR antibodies have been demonstrated in the sera of HIV-infected individuals (see "Autoimmune Phenomena," above). These antibodies could contribute to the elimination of HLA-DR-expressing cells byADCC; in addition, it has been suggested that these antibodies may inhibit certain T cell functions that involve HLA-DR molecules.

HIV envelope glycoproteins gp120 and gp160 manifest high-affinity binding to CD4 as well as to various chemokine receptors (see above). Intracellular signals transduced by gp120 have been associated with a number of immunopathogenic processes including anergy, apoptosis, and abnormalities of cell trafficking. The molecular mechanisms responsible for these abnormalities include dysregulation of the T cell receptor-phosphoinositide pathway, p56lck activation, phosphorylation of focal adhesion kinase, activation of the MAP kinase and ras signaling pathways, and downregulation of the co-stimulatory molecules CD40 ligand and CD80.

Finally, the inexorable decline in CD4+ T cell counts that occurs in most HIV-infected individuals may result in part from the inability of the immune system to regenerate the CD4+ T cell pool rapidly enough to compensate for both HIV-mediated destruction of cells and natural attrition of cells. At least two major mechanisms may contribute to the failure of the CD4+ T cell pool to reconstitute itself adequately over the course of HIV infection. The first is the destruction of lymphoid precursor cells, including thymic and bone marrow progenitor cells (see above); the other is the gradual disruption of the lymphoid tissue microenvironment, which is essential for efficient regeneration of immune-competent cells (see above).

CD8+ T Cells The level of CD8+ T cells varies throughout the course of disease. Following the resolution of acute primary infection, CD8+ T cells generally rebound to higher than normal levels and may remain that way throughout the clinically latent stage of disease. This CD8+ T lymphocytosis may in part reflect the expansion of clones of HIV-specific CD8+CTLs. During the late stages of HIV infection, there may be a significant reduction in the numbers of CD8+ T cells. HIV-specific CD8+ CTLs have been demonstrated in HIV-infected individuals early in the course of disease (see below). As the disease progresses, this functional capability decreases and may be lost entirely. The cause of this loss of cytolytic activity is unclear. However, it has been demonstrated that, as disease progresses, CD8+ T cells assume an abnormal phenotype characterized by expression of activation markers such as HLA-DR with an absence of expression of theIL-2 receptor (CD25) and a loss of clonogenic potential. It has been reported that the phenotype of CD8+ T cells in HIV-infected individuals may be of prognostic significance. Those individuals whose CD8+ T cells developed a phenotype of HLA-DR+/CD38- following seroconversion had stabilization of their CD4+ T cell counts, whereas those whose CD8+ T cells developed a phenotype of HLA-DR+/CD38+ had a more aggressive course and a poorer prognosis. In addition to the defects in HIV-specific CTLs, functional defects in otherMHC-restricted CTLs, such as those directed against influenza andCMV, have been demonstrated. Since the integrity of CD8+ T cell function depends in part on adequate inductive signals from

CD4+ T cells, the defect in CD8+ CTLs is likely compounded by the quantitative loss of CD4+ T cells.

B Cells B cells from HIV-infected individuals manifest abnormal activation, which is reflected by spontaneous proliferation and immunoglobulin secretion and by increased spontaneous secretion ofTNF-aandIL-6. The enhanced spontaneous in vitro transformation of B cells withEBV is probably due to defective T cell immune surveillance and has as its in vivo counterpart an increase in the incidence of EBV-related B cell lymphomas. Untransformed B cells cannot be infected with HIV. However, HIV or its products can activate B cells directly; portions of the HIV gp41 envelope protein have been reported to induce polyclonal B cell activation. In addition, it has been reported that products of the VH_3genes on the surface of B cells can serve as a receptor for HIV. It is likely that in vivo activation of B cells by virus products accounts at least in part for the spontaneous activation of these cells noted ex vivo. B cells from HIV-infected individuals express abnormally low levels of HLA-DR on their surface and fail to upregulate CD70 normally following stimulation with activated T cells; this latter defect is associated with impaired CD70-dependent immunoglobulin synthesis. In advanced HIV disease, B cells fail to proliferate and differentiate in response to ligation of the B cell antigen receptor and CD40, suggesting a defect in signal transduction. In vivo, this activated state manifests itself by hypergammaglobulinemia and by the presence of circulating immune complexes and autoantibodies (see above). HIV-infected individuals respond poorly to primary and secondary immunizations with protein and polysaccharide antigens. These B cell defects are likely responsible in part for the increase in certain bacterial infections seen in advanced HIV disease in adults, as well as for the important role of bacterial infections in the morbidity and mortality of HIV-infected children, who cannot mount an adequate humoral response to common bacterial pathogens. The absolute number of circulating B cells may be depressed in primary HIV infection; however, this phenomenon is usually transient and likely reflects in part a redistribution of cells out of the circulation and into the lymphoid tissue. In certain patients, the number of circulating B cells decreases in advanced-stage disease.

Monocyte/Macrophages Circulating monocytes are generally normal in number in HIV-infected individuals. Monocytes express the CD4 molecule and several co-receptors for HIV on their surface, including CCR5, CXCR4, and CCR3, and thus are targets of HIV infection. Of note is the fact that the degree of cytopathicity of HIV for cells of the monocyte lineage is low, and HIV can replicate extensively in cells of the monocyte lineage with little cytopathic effect. Hence, monocyte-lineage cells may play a role in the dissemination of HIV in the body and can serve as reservoirs of HIV infection, thus representing an obstacle to the eradication of HIV by antiretroviral drugs. In vivo infection of circulating monocytes is difficult to demonstrate; however, infection of tissue macrophages and macrophage-lineage cells in the brain (infiltrating macrophages or resident microglial cells) and lung (pulmonary alveolar macrophages) can be demonstrated easily. Infection of monocyte precursors in the bone marrow may directly or indirectly be responsible for certain of the hematologic abnormalities in HIV-infected individuals. A number of abnormalities of circulating monocytes have been reported in HIV-infected individuals, including decreased secretion ofIL-1 and IL-12; increased secretion of IL-10; defects in antigen presentation and induction of T cell responses due to decreasedMHCclass II expression; and abnormalities of Fc receptor function, C3 receptor-mediated clearance, oxidative burst responses, and certain cytotoxic functions such asADCC, possibly related to low levels of expression of Fc and complement receptors. The mechanisms of the monocyte defects are uncertain but almost certainly cannot be even partly explained by direct infection with HIV. Exposure of monocytes to viral proteins such as gp120 and Tat, as well as to certain cytokines, can cause abnormal activation, and this may play a role in cellular dysfunction (see above).

Dendritic and Langerhans Cells There has been considerable disagreement regarding the HIV infectibility and hence the depletion as well as the dysfunction of circulating dendritic cells. Depending on their state of maturation, dendritic cells express varying levels of CD4 as well as several chemokine receptors. In this regard, it appears that the ability of a dendritic cell to become infected depends in part on its state of maturation. Mature dendritic cells have been demonstrated to be infectable by both R5 and X4 isolates of HIV-1. Immature tissue dendritic cells have been less well studied in their native state. Certain groups have reported infection and dysfunction of dendritic cells from HIV-infected individuals, particularly a decreased ability to present antigen to T cells, and other groups have found little if any HIV infection or functional abnormalities. In this regard, there is general agreement regarding the ability of skin and mucous membrane Langerhans cells to be infected (see above). These latter cells likely play an important role in the initiation and propagation of HIV infection (see above). Even in those dendritic cells in which infection occurs, the efficiency of infection and level of productivity of infection is quite low compared to CD4+ T cells.

Natural Killer Cells The role ofNK cells is to provide immunosurveillance against virus-infected cells, certain tumor cells, and allogeneic cells (Chap. 305). Functional abnormalities in NK cells have been observed throughout the course of HIV disease, and the severity of these abnormalities increases as disease progresses. Most studies report that NK cells are normal in number and phenotype in HIV-infected individuals; however, a numerical decrease in the CD16+/CD56+ subpopulation of NK cells has been reported together with an increase in activation markers. The abnormality in NK cell function is thought to result from a defect in postbinding lysis. However, the lytic machinery does not appear to be impaired, since NK cells from HIV-infected individuals mediateADCCnormally. The addition of eitherIL-2, IL-12, IL-15, orIFN-ato cultures improves the defective in vitro NK cell function of HIV-infected individuals. Enhanced expression of cytolytic inhibitory receptors in HIV-infected individuals may contribute to the abnormalities in NK function. Furthermore, selective HIV-mediated downregulation of HLA-A and -B, but not HLA-C and -D molecules may inhibit NK-mediated killing of HIV-infected target cells. Finally, NK cells serve as important sources of HIV-inhibitory CC-chemokines. NK cells isolated from HIV-infected individuals constitutively produce high levels ofMIP-1a, MIP-1b, and RANTES. In addition, high levels of these chemokines are seen when NK cells are stimulated with IL-2 or IL-15 or when CD16 is cross-linked or during the process of lytic killing of target cells.

GENETIC FACTORS IN HIV PATHOGENESIS

Several reports have describedMHCalleles and other host factors that may influence the pathogenesis and course of HIV disease. These include associations with certain HIV-related manifestations, such asKS and diffuse lymphadenopathy, or with the type of clinical course, such as long-term survival or rapid progression (Table 309-7). A number of mechanisms have been proposed whereby MHC-encoded molecules might predispose an individual either to rapid progression or to nonprogression to AIDS. These proposed mechanisms include the ability to present certain immunodominant HIV T helper orCTLepitopes, leading to a relatively protective immune response against HIV and hence to slow progression of disease. In contrast, certain MHC class I or class II alleles might predispose an individual to an immunopathogenic response against viral epitopes in certain tissues, such as the central nervous system (CNS) or lungs, or against certain HIV-infected cell types, such as macrophages or dendritic cells/Langerhans cells. In addition, certain rare MHC class I and class II alleles might facilitate rapid recognition of HIV-infected cells from the infecting partner in primary HIV infection and promote rejection of these cells by alloreactive responses. Similarly, common MHC alleles could lead to less effective removal of HIV-infected allogeneic cells. It has been clearly demonstrated that maximal *HLA* heterozygosity for class I loci (A, B, and C) is associated with a delayed onset of AIDS among HIV-infected individuals, whereas homozygosity for these loci was associated with a more rapid progression to AIDS and death. This observation is likely due to the fact that individuals heterozygous at *HLA* loci are able to present a greater variety of antigenic peptides to cytotoxic T lymphocytes than are homozygotes resulting in a more effective immune response against a number of pathogens including HIV. Of particular note is the fact that the HLA class I alleles B*35 and Cw*04 were consistently associated with rapid development of AIDS. Other data have indicated that transporter associated with antigen-presenting (TAP) genes play a role in determining the outcome of HIV infection. HLA profiles that reflect certain combinations of MHC-encoded TAP and class I and class II genes are strongly associated with different rates of progression to AIDS.

Rare individuals have been reported who had had repetitive sexual exposure to HIV in high-risk situations but remained uninfected. The peripheral blood mononuclear cells of two such individuals were found to be highly resistant to infection in vitro with R5 strains of HIV-1, but they were readily infected with X4 strains. Genetic analysis revealed that these two individuals inherited a homozygous defect in the gene that codes for CCR5, the cellular co-receptor for R5 strains of HIV-1. The defective *CCR5* allele contained a 32-bp deletion corresponding to the second extracellular loop of the receptor. The encoded protein was severely truncated, and the receptor was nonfunctional, explaining the refractoriness to infection with R5 strains of HIV-1. Population studies revealed that approximately 1% of the Caucasian population of western European ancestry possessed the homozygous defect. Up to 20% of this group had the heterozygous defect. Of note, cohort studies of hundreds of DNA samples originating from western and central Africa and Japan did not reveal a single mutant allele, suggesting that the allele is either absent or extremely rare in Africa and Japan. In a cohort of 1400 HIV-1-infected Caucasian individuals, no subject homozygous for the mutation was found, strongly supporting the concept that the homozygous defect confers protection against infection. This finding is particularly compelling in light of the fact that transmitting viruses are strongly biased towards R5 strains of HIV-1 (see above). Furthermore, there was a higher frequency of individuals heterozygous for the genetic defect among HIV-infected patients who were long-term nonprogressors compared to HIV-infected individuals who progressed more rapidly (see above). Of note, several individuals have been identified who were homozygous for the *CCR5* D32 defect who in fact did become infected with HIV. These individuals were found to have an X4 strain of HIV that was associated in some cases with an accelerated course of disease. Slow progression of HIV disease is also seen in individuals who are heterozygous for the

CCR2V64I mutation; this is felt to be due to dimerization of CXCR4 with the mutated CCR2V64I resulting in a decreased expression of CXCR4 on the cell surface. Individuals who are homozygous for the *SDF1*-3¢A mutation manifest slow progression, likely due to the upregulation ofSDF-1 and resulting inhibition of binding of X4 viruses to mononuclear cells. Delayed progression of disease is also seen in those individuals who have any of a number of single nucleotide polymorphisms in the *CCR5* promoter. In addition, individuals who carry a certain allele (IL-10-5¢592A) of theIL-10 promoter are at increased risk of infection and, once infected, progress more rapidly than homozygotes for the alternative genotype. The mechanism of this effect is felt to be a downregulation of the inhibitory cytokine IL-10 resulting in facilitation of HIV replication. Finally, individuals with a mutation of the *RANTES* gene (*RANTES-28G*) manifest a delay in disease progression due likely to the increased expression of RANTES and resulting inhibition of infection with R5 viruses (Table 309-7).

NEUROPATHOGENESIS

HIV-infected individuals can experience a variety of neurologic abnormalities due either to opportunistic infections and neoplasms (see below) or to direct effects of HIV or its products. With regard to the latter, HIV has been demonstrated in the brain andCSF of infected individuals with and without neuropsychiatric abnormalities. The main cell types that are infected in the brain in vivo are those of the monocyte/macrophage lineage, including monocytes that have migrated to the brain from the peripheral blood as well as resident microglial cells. HIV entry into brain is felt to be due, at least in part, to the ability of virus-infected and immune-activated macrophages to induce adhesion molecules such as E-selectin and vascular cell adhesion molecule-1 (VCAM-1) on brain endothelium. Other studies have demonstrated that HIV gp120 enhances the expression of intercellular adhesion molecule-1 (ICAM-1) in glial cells; this effect may facilitate entry of HIV-infected cells into theCNS and may promote syncytia formation. Virus isolates from the brain are preferentially R5 strains as opposed to X4 strains (see above); in this regard, HIV-infected individuals who are heterozygous for *CCR5D32* appear to be relatively protected against the development of HIV encephalopathy compared to wild-type individuals. Distinct HIV envelope sequences are associated with the clinical expression of the AIDS dementia complex (see below). Although there have been reports of infrequent HIV infection of neuronal cells and astrocytes, there is no convincing evidence that brain cells other than those of monocyte/macrophage lineage can be productively infected in vivo. Nonetheless, it has been demonstrated that galactosyl ceramide may be an essential component of the HIV gp120 receptor on neural cells, and antibodies to galactosyl ceramide inhibit entry of HIV into neural cell lines in vitro.

HIV-infected individuals may manifest white matter lesions as well as neuronal loss. Given the relative absence of evidence of HIV infection of neurons either in vivo or in vitro, it is unlikely that direct infection of these cells accounts for their loss. Rather, the HIV-mediated effects on brain tissue are thought to be due to a combination of direct effects, either toxic or function-inhibitory, of gp120 on neuronal cells and effects of a variety of neurotoxins released from infiltrating monocytes, resident microglial cells, and astrocytes. In this regard, it has been demonstrated that both HIV-1 Nef and Tat can induce chemotaxis of leukocytes, including monocytes, into theCNS. Neurotoxins can be released from monocytes as a consequence of infection and/or immune activation.

Monocyte-derived neurotoxic factors have been reported to kill neurons via the *N*-methyl-D-aspartate (NMDA) receptor. In addition, HIV gp120 shed by virus-infected monocytes could cause neurotoxicity by antagonizing the function of vasoactive intestinal peptide (VIP), by elevating intracellular calcium levels, and by decreasing nerve growth factor levels in the cerebral cortex. A variety of monocyte-derived cytokines can contribute directly or indirectly to the neurotoxic effects in HIV infection; these includeTNF-a,IL-1, IL-6,TGF-b,IFN-g, platelet-activating factor, and endothelin. Certain studies have correlated levels of CC-chemokinesMIP-1a, MIP-1b, and RANTES inCSF with HIV-related encephalopathy. In addition, infection and/or activation of monocyte-lineage cells can result in increased production of eicosanoids, nitric oxide, and quinolinic acid, which may contribute to neurotoxicity. Astrocytes may play diverse roles in HIV neuropathogenesis. Reactive gliosis or astrocytosis has been demonstrated in the brains of HIV-infected individuals, and TNF-a and IL-6 have been shown to induce astrocyte proliferation. In addition, astrocyte-derived IL-6 can induce HIV expression in infected cells in vitro. Furthermore, it has been suggested that astrocytes may downregulate macrophage-produced neurotoxins. It has been reported that HIV-infected individuals with the E4 allele for apolipoprotein E (apo E) are at increased risk for AIDS encephalopathy and peripheral neuropathy. The likelihood that HIV or its products are involved in neuropathogenesis is supported by the observation that neuropsychiatric abnormalities may undergo remarkable and rapid improvement upon the initiation of antiretroviral therapy, particularly in HIV-infected children.

PATHOGENESIS OF KAPOSI'S SARCOMA

KSis an opportunistic disease in HIV-infected individuals. Unlike opportunistic infections, its occurrence is not strictly related to the level of depression of CD4+ T cell counts (see below). There are at least four distinct epidemiologic forms of Kaposi's sarcoma: (1) the classic form that occurs in older men of predominantly Mediterranean or eastern European Jewish backgrounds with no recognized contributing factors; (2) the equatorial African form that occurs in all ages, also without any recognized precipitating factors; (3) the form associated with organ transplantation and its attendent iatrogenic immunosuppressed state; and (4) the form associated with HIV-1 infection. The pathogenesis of KS is complex and has not been fully delineated. KS does not result from a neoplastic transformation of cells in the classic sense and so is not truly a sarcoma. It is a manifestation of excessive proliferation of spindle cells that are believed to be of vascular origin and have features in common with endothelial and smooth-muscle cells. In HIV disease the development of KS is dependent on the interplay of a variety of factors including HIV-1 itself,HHV-8, immune activation, and cytokine secretion. A number of epidemiologic and virologic studies have clearly linked HHV-8, which is also referred to as *Kaposi's sarcoma-associated herpesvirus* (KSHV), to KS not only in HIV-infected individuals but also in individuals with the other forms of KS. KSHV is a g-herpesvirus related toEBV and herpesvirus saimiri. It encodes a homologue to human IL-6 and in addition to KS has been implicated in the pathogenesis of body cavity lymphoma, multiple myeloma, and monoclonal gammopathy of undetermined significance. Sequences of HHV-8 are found universally in the lesions of KS, and patients with KS are virtually all seropositive for HHV-8. HHV-8 DNA sequences can be found in the B cells of 30 to 50% of patients with KS and 7% of patients with AIDS without clinically apparent KS.

Between 1 and 2% of eligible blood donors are positive for antibodies toHHV-8, while the prevalence of HHV-8 seropositivity in HIV-infected men is 30 to 35%. The prevalence in HIV-infected women is approximately 4%. This finding is reflective of the lower incidence ofKS in women. It has been debated whether HHV-8 is actually the transforming agent in KS; the bulk of the cells in the tumor lesions of KS are not neoplastic cells. However, a recent study has demonstrated that endothelial cells can be transformed in vitro by HHV-8. Despite the complexity of the pathogenic events associated with the development of KS in HIV-infected individuals, it is generally felt that HHV-8 is indeed the etiologic agent of this disease. The initiation and/or propagation of KS requires an activated state and is mediated, at least in part, by cytokines. A number of factors, includingTNF-a,IL-1b, IL-6,GM-CSF, basic fibroblast growth factor, and oncostatin M, function in an autocrine and paracrine manner to sustain the growth and chemotaxis of the KS spindle cells. It has been suggested that the HIV Tat protein plays a major role in the pathogenesis of KS. In this regard, it has been demonstrated thatIFN-gcan induce endothelial cells to proliferate and to invade the extracellular matrix in response to HIV Tat. This occurs as a result of the upregulation by IFN-g of the expression and activity of the receptors for Tat, which are the integrins$a_5b_1$anda_vb_3. In addition, the HIV-1 Tat protein has been shown to act synergistically with basic fibroblast growth factor in the induction of lesions resembling KS lesions in mice. Glucocorticoids have been shown to have a stimulatory effect, and human chorionic gonadotrophin an inhibitory effect, on KS spindle cells, suggesting that modulation of the balance of autocrine factors may have therapeutic potential in KS.

IMMUNE RESPONSE TO HIV

As detailed above and below, following the initial burst of viremia during primary infection, HIV-infected individuals mount a robust immune response that usually substantially curtails the levels of plasma viremia and likely contributes to delaying the ultimate development of clinically apparent disease for a median of 10 years. This immune response contains elements of both humoral and cell-mediated immunity (Table 309-8;Fig. 309-21). It is directed against multiple antigenic determinants of the HIV virion as well as against viral proteins expressed on the surface of infected cells. Ironically, those CD4+ T cells with T cell receptors specific for HIV are theoretically those CD4+ T cells most likely to bind to infected cells and themselves be infected and destroyed. Thus, an early consequence of HIV infection may be interference with the generation of an effective immune response through the elimination of HIV-specific CD4+ T lymphocytes.

Although a great deal of investigation has been directed toward delineating and better understanding the components of this immune response, it remains unclear which of these phenomena are most important in delaying progression of infection and which, if any, play a role in the pathogenesis of HIV disease.

HUMORAL IMMUNE RESPONSE

Antibodies to HIV usually appear within 6 weeks and almost invariably within 12 weeks of primary infection (Fig. 309-22); rare exceptions are individuals who have defects in the ability to produce HIV-specific antibodies. Detection of these antibodies forms the basis of most diagnostic screening tests for HIV infection. The appearance of

HIV-binding antibodies detected byELISAand western blot assays occurs prior to the appearance of neutralizing antibodies; the latter generally appear following the initial decreases in plasma viremia, which is more closely related to the appearance of HIV-specific CD8+ T lymphocytes. The first antibodies detected are those directed against the structural or gag proteins of HIV, p24 and p17, and the gag precursor p55. The development of antibodies to p24 is associated with a decrease in the serum levels of free p24 antigen. Antibodies to the gag proteins are followed by the appearance of antibodies to the envelope proteins (gp160, gp120, p88, and gp41) and to the products of the *pol* gene (p31, p51, and p66). In addition, one may see antibodies to the low-molecular-weight regulatory proteins encoded by the HIV genes *vpr*, *vpu*, *vif*, *rev*, *tat*, and *nef*.

While antibodies to multiple antigens of HIV are produced, the precise functional significance of these different antibodies is unclear. The best studied have been the antibodies directed towards the envelope proteins of the virus. As noted above, the envelope of HIV consists of an outer envelope glycoprotein with a molecular mass of 120 kDa and a transmembrane glycoprotein with a molecular mass of 41 kDa. These are initially synthesized as a 160-kDa precursor that is cleaved by cellular proteases. Most of the antienvelope antibodies are directed either toward an epitope in the gp41 region comprising amino acids 579 to 613 or toward a hypervariable region in the gp120 molecule, known as the *V3 loop region*, comprising amino acids 303 through 338. This V3 region is a major site for the development of mutations that lead to variants of HIV that are not well recognized by the immune system.

Antibodies directed toward the envelope proteins of HIV have been characterized both as being protective and as possibly contributing to the pathogenesis of HIV disease. Among the protective antibodies are those that function to neutralize HIV directly and prevent the spread of infection to additional cells, as well as those that participate inADCC. *Neutralizing antibodies* may be a component of primary HIV infection, and some long-term nonprogressors have been reported to have increased titers of neutralizing antibodies. Neutralizing antibodies appear to be of two forms, type-specific and group-specific. *Type-specific neutralizing antibodies* are generally directed to the V3 loop region. These antibodies neutralize only viruses of a given strain and are present in low titer in most infected individuals. *Group-specific neutralizing antibodies* are capable of neutralizing a wide variety of HIV isolates. At least two forms of group specific antibodies have been identified: those binding to amino acids 423 to 437 of gp120 and those binding to amino acids 728 to 745 of gp41. The other major class of protective antibodies are those that participate in ADCC, which is actually a form of cell-mediated immunity (Chap. 305) in whichNK cells that bear Fc receptors are armed with specific anti-HIV antibodies that bind to the NK cells via their Fc portion. These armed NK cells then bind to and destroy cells expressing HIV antigens. Antibodies to both gp120 and gp41 have been shown to participate in ADCC-mediated killing of HIV-infected cells. The levels of antienvelope antibodies capable of mediating ADCC are highest in the earlier stages of HIV infection. In vitro, IL-2 can augment ADCC-mediated killing.

In addition to playing a role in host defense, HIV-specific antibodies have also been implicated in disease pathogenesis. Antibodies directed to gp41, when present in low titer, have been shown in vitro to be capable of facilitating infection of cells through an Fc receptor-mediated mechanism known as *antibody enhancement*. Thus, the same regions of the envelope protein of HIV that give rise to antibodies capable of mediatingADCCalso elicit the production of antibodies that can facilitate infection of cells in vitro. In addition, it has been postulated that anti-gp120 antibodies that participate in the ADCC killing of HIV-infected cells might also kill uninfected CD4+ T cells if the uninfected cells had bound free gp120, a phenomenon referred to as *bystander killing*.

CELLULAR IMMUNE RESPONSE

Given the fact that T cell-mediated immunity is known to play a major role in host defense against most viral infections (Chap. 305), it is generally thought to be an important component of the host immune response to HIV. T cell immunity can be divided into two major categories, mediated respectively by the *helper/inducer CD4+ T cells* and the *cytotoxic/immunoregulatory CD8+ T cells*.

It has been difficult to demonstrate the presence of HIV-specific CD4+ T cells in HIV-infected patients directly, particularly in those with advanced disease. This difficulty may be related to the fact that these cells, with their high affinity for binding to HIV-infected cells, may be among the first to be infected and destroyed during HIV infection. CD4+ T lymphocytes with reactivity to the p24 antigen of HIV have been reported to be present in a subset of long-term nonprogressors and in a subset of patients in whom therapy was initiated shortly following infection. While a reverse correlation exists between the presence of these cells and levels of plasma HIV viremia, it is unclear whether or not there is a causal relationship between these parameters. Through the use of computer modeling, several regions of the HIV-1 envelope molecule have been identified that are structurally analogous to other known T cell epitopes by virtue of having structures known as *amphipathic helices*. Peptides from these envelope regions have been used to identify the presence of CD4+ T cells specific for these regions in the peripheral blood of HIV-infected individuals. Other studies have demonstrated that peripheral blood T cells of some healthy, HIV-negative individuals also react to the envelope proteins of HIV. It is unclear whether or not this represents the presence of a degree of protective immunity in these individuals.

MHCclass I-restricted, HIV-specific CD8+ T cells have been identified in the peripheral blood of patients with HIV-1 infection. These cells include cytotoxic T cells (CTLs) and T cells that can be induced by HIV antigens to express cytokines such asIFN-g. CTLs have been identified in the peripheral blood of patients within weeks of HIV infection. These CD8+ T lymphocytes, through their HIV-specific antigen receptors, bind to and cause the lytic destruction of target cells bearing identical MHC class I molecules associated with HIV antigens. Two types of CTL activity can be demonstrated in the peripheral blood or lymph node mononuclear cells of HIV-infected individuals. The first type directly lyses appropriate target cells in culture without prior in vitro stimulation (*spontaneous CTL activity*). The other type of CTL activity reflects the *precursor frequency of CTLs* (CTLp); this type of CTL activity can be demonstrated by stimulation of CD8+ T cells in vitro with a mitogen such as phytohemagglutinin or anti-CD3 antibody. Following primary HIV infection, the qualitative nature of the HIV-specific CTL response is an important predictor of eventual clinical outcome. Patients who mount a broad CD8+ CTL response generally have a more favorable clinical course than do patients who mount a more restricted CTL response. These data are consistent with studies in theSIVmodel where deletion of CD8+ T cells leads to a more accelerated clinical course.

In addition toCTLs, CD8+ T cells capable of being induced by HIV antigens to express cytokines such asIFN-galso appear in the setting of HIV-1 infection. It is not clear whether these are the same or different effector pools compared to those cells mediating cytotoxicity; in addition, the relative roles of each in host defense against HIV are not fully understood. It does appear that these CD8+ T cells are driven to in vivo expansion by HIV antigen. There is a direct correlation between levels of CD8+ T cells capable of producing IFN-g in response to HIV antigens and plasma levels of HIV-1 RNA. Thus, while these cells are clearly induced by HIV-1 infection, their overall ability to control infection remains unclear. Multiple HIV antigens, including Gag, Env, Pol, Tat, Rev, and Nef, can elicit CD8+ T cell responses.

At least three other forms of cell-mediated immunity to HIV have been described: CD8+ T cell-mediated suppression of HIV replication,ADCC, and NKcell activity. *CD8+ T cell-mediated suppression of HIV replication* refers to the ability of CD8+ T cells from an HIV-infected patients to inhibit the replication of HIV in tissue culture in a noncytolytic manner. There is no requirement for HLA compatibility between the CD8+ T cells and the HIV-infected cells. This effector mechanism is thus nonspecific and appears to be mediated by soluble factor(s) including the CC-chemokines RANTES,MIP-1aand MIP-1b (see above). These chemokines are potent suppressors of HIV replication and operate at least in part via blockade of the co-receptor (CCR5) on peripheral blood mononuclear cells for R5 or macrophage-tropic strains of HIV (see above). *ADCC*, as described above in relation to humoral immunity, involves the killing of HIV-expressing cells by NK cells armed with specific antibodies directed against HIV antigens. Finally, *NK cells* alone have been shown to be capable of killing HIV-infected target cells in tissue culture. This primitive cytotoxic mechanism of host defense is directed toward nonspecific surveillance for neoplastic transformation and viral infection through recognition of altered class IMHCmolecules.

DIAGNOSIS AND LABORATORY MONITORING OF HIV INFECTION

The establishment of HIV as the causative agent of AIDS and related syndromes early in 1984 was followed by the rapid development of sensitive screening tests for HIV infection. By March, 1985, blood donors in the United States were routinely screened for antibodies to HIV. In June 1996, blood banks in the United States added the p24 antigen capture assay to the screening process to help identify the rare infected individuals who were donating blood in the time (up to 3 months) between infection and the development of antibodies. The development of sensitive assays for monitoring levels of plasma viremia ushered in a new era of being able to monitor the progression of HIV disease more closely. Utilization of these tests, coupled with the measurement of levels of CD4+ T lymphocytes in peripheral blood, is essential in the management of patients with HIV infection.

DIAGNOSIS OF HIV INFECTION

The diagnosis of HIV infection depends upon the demonstration of antibodies to HIV and/or the direct detection of HIV or one of its components. As noted above, antibodies to HIV generally appear in the circulation 2 to 12 weeks following infection.

The standard screening test for HIV infection is theELISA, also referred to as an enzyme immunoassay (EIA). This solid-phase assay is an extremely good screening test with a sensitivity of>99.5%. Most diagnostic laboratories use a commercial EIA kit that contains antigens from both HIV-1 and HIV-2 and thus are able to detect either. These kits use both natural and recombinant antigens and are continuously updated to increase their sensitivity to newly discovered species, such as group O viruses (Fig. 309-6). EIA tests are generally scored as positive (highly reactive), negative (nonreactive), or indeterminate (partially reactive). While the EIA is an extremely sensitive test, it is not optimal with regard to specificity. This is particularly true in studies of low-risk individuals, such as volunteer blood donors. In this latter population, only 10% of EIA-positive individuals are subsequently confirmed to have HIV infection. Among the factors associated with false-positive EIA tests are antibodies to class II antigens, autoantibodies, hepatic disease, recent influenza vaccination, and acute viral infections. For these reasons, anyone suspected of having HIV infection based upon a positive or inconclusive EIA result must have the result confirmed with a more specific assay.

The most commonly used confirmatory test is the western blot (Fig. 309-23). This assay takes advantage of the fact that multiple HIV antigens of different, well-characterized molecular weights elicit the production of specific antibodies. These antigens can be separated on the basis of molecular weight, and antibodies to each component can be detected as discrete bands on the western blot. A negative western blot is one in which no bands are present at molecular weights corresponding to HIV gene products. In a patient with a positive or indeterminateEIA and a negative western blot, one can conclude with certainty that the EIA reactivity was a false positive. On the other hand, a western blot demonstrating antibodies to products of all three of the major genes of HIV (*gag*, *pol*, and *env*) is conclusive evidence of infection with HIV. Criteria established by the U.S. Food & Drug Administration (FDA) in 1993 for a positive western blot state that a result is considered positive if antibodies exist to two of the three HIV proteins: p24, gp41, and gp120/160. Using these criteria, approximately 10% of all blood donors deemed positive for HIV-1 infection lacked an antibody band to the *pol* gene product p31. Some 50% of these blood donors were subsequently found to be false positives. Thus, the absence of the p31 band should increase the suspicion that one may be dealing with a false-positive test result. In this setting it is prudent to obtain additional confirmation with an RNA-based test and/or a follow-up western blot. By definition, western blot patterns of reactivity that do not fall into the positive or negative categories are considered "indeterminate." There are two possible explanations for an indeterminate western blot result. The most likely explanation in a low-risk individual is that the patient being tested has antibodies that cross-react with one of the proteins of HIV. The most common patterns of cross-reactivity are antibodies that react with p24 and/or p55. The least likely explanation in this setting is that the individual is infected with HIV and is in the process of mounting a classic antibody response. In either instance, the western blot should be repeated in 1 month to determine whether or not the indeterminate pattern is a pattern in evolution. In addition, one may attempt to confirm a diagnosis of HIV infection with the p24 antigen capture assay or one of the tests for HIV RNA (discussed below). While the western blot is an excellent confirmatory test for HIV infection in patients with a positive or indeterminate EIA, it is a poor screening test. Among individuals with a negative EIA andPCR for HIV, 20 to 30% may show one or more bands on western blot. While these bands are usually faint and represent cross-reactivity, their presence creates a situation in which other diagnostic modalities [such as DNA PCR, RNA PCR, the (b)DNA assay, or p24 antigen capture] must be employed to ensure that the bands do not indicate early HIV infection.

A guideline for the use of these serologic tests in attempting to make a diagnosis of HIV infection is depicted in Fig. 309-24. In patients in whom HIV infection is suspected, the appropriate initial test is theEIA. If the result is negative, unless there is strong reason to suspect early HIV infection (as in a patient exposed within the previous 3 months), the diagnosis is ruled out and retesting should be performed only as clinically indicated. If the EIA is indeterminate or positive, the test should be repeated. If the repeat is negative on two occasions, one can assume that the initial positive reading was due to a technical error in the performance of the assay and that the patient is negative. If the repeat is indeterminate or positive, one should proceed to the HIV-1 western blot. If the western blot is positive, the diagnosis is HIV-1 infection. If the western blot is negative, the EIA can be assumed to have been a false positive for HIV-1 and the diagnosis of HIV-1 infection is ruled out. It would be prudent at this point to perform specific serologic testing for HIV-2 following the same type of algorithm. If the western blot for HIV-1 is indeterminate, it should be repeated in 4-6 weeks; in addition, one may proceed to a p24 antigen capture assay, HIV-1 RNA assay, or HIV-1 DNAPCR and specific serologic testing for HIV-2. If the p24 and HIV RNA assays are negative and there is no progression in the western blot, a diagnosis of HIV-1 is ruled out. If either the p24 or HIV-1 RNA assay is positive and/or the HIV-1 western blot shows progression, a tentative diagnosis of HIV-1 infection can be made and later confirmed with a follow-up western blot demonstrating a positive pattern.

As mentioned above, a variety of laboratory tests are available for the direct detection of HIV or its components (Table 309-9;Fig. 309-25). These tests may be of considerable help in making a diagnosis of HIV infection when the western blot results are indeterminate. In addition, the tests detecting levels of HIV RNA can be used to determine prognosis and to assess the response to antiretroviral therapies. The simplest of the direct detection tests is the *p24 antigen capture assay*. This is anEIA-type assay in which the solid phase consists of antibodies to the p24 antigen of HIV. It detects the viral protein p24 in the blood of HIV-infected individuals where it exists either as free antigen or complexed to anti-p24 antibodies. Overall, approximately 30% of individuals with untreated HIV infection have detectable levels of free p24 antigen. This increases to about 50% when samples are treated with a weak acid to dissociate antigen-antibody complexes. Throughout the course of HIV infection, an equilibrium exists between p24 antigen and anti-p24 antibodies. During the first few weeks of infection, before an immune response develops, there is a brisk rise in p24 antigen levels (Fig. 309-22). After the development of anti-p24 antibodies, these levels decline. Late in the course of infection, when circulating levels of virus are high, p24 antigen levels also increase, particularly when detected by techniques involving dissociation of antigen-antibody complexes. This assay has its greatest use as a screening test for HIV infection in patients suspected of having the acute HIV syndrome, as high levels of p24 antigen are present prior to the development of antibodies. In addition, it is currently routinely used along with the HIV EIA assay to screen blood donors in the United States for evidence of HIV infection. Its utility as an assay is decreasing with the increased use of the reverse transcriptase PCR (RT-PCR) and bDNA technique for direct detection of HIV RNA.

The ability to measure and monitor levels of HIV RNA in the plasma of patients with HIV infection has been of extraordinary value in furthering our understanding of the pathogenesis of HIV infection and in providing a diagnostic tool in settings where measurements of anti-HIV antibodies may be misleading, such as in acute infection and neonatal infection. Two assays are predominantly used for this purpose. They are the*RT-PCR*(Amplicor) and the *bDNA* (Quantiplex). It should be pointed out that the only test approved by theFDA at this time for the measurement of HIV RNA levels is the RT-PCR test. While this approval is limited to the use of the test for determining prognosis, it is the general consensus that this test as well as the bDNA test are also of value for monitoring the effects of therapy and in making a diagnosis of HIV infection. In addition to these two commercially available tests, the *DNA PCR* is also employed by research laboratories for making a diagnosis of HIV infection by amplifying HIV proviral DNA from peripheral blood mononuclear cells. The commercially available RNA detection tests have a sensitivity of 40 to 50 copies of HIV RNA per milliliter of plasma, while the DNA PCR tests can detect proviral DNA at a frequency of one copy per 10,000 to 100,000 cells. Thus, these tests are extremely sensitive. One frequent consequence of a high degree of sensitivity is some loss of specificity, and false-positive results have been reported with each of these techniques. For this reason, a positiveEIA with a confirmatory western blot remains the "gold standard" for a diagnosis of HIV infection, and the interpretation of other test results must be done with this in mind.

In theRT-PCRtechnique, following DNAase treatment, a cDNA copy is made of all RNA species present in plasma. Insofar as HIV is an RNA virus, this will result in the production of DNA copies of the HIV genome in amounts proportional to the amount of HIV RNA present in plasma. This proviral DNA is then amplified and characterized using standardPCRtechniques, employing primer pairs that can distinguish genomic cDNA from messenger cDNA. The bDNA assay involves the use of a solid-phase nucleic acid capture system and signal amplification through successive nucleic acid hybridizations to detect small quantities of HIV RNA. Both tests can achieve a tenfold increase in sensitivity to 40 to 50 copies of HIV RNA per milliliter with a preconcentration step in which plasma undergoes ultracentrifugation to pellet the viral particles. In addition to being a diagnostic and prognostic tool, RT-PCR is also useful for amplifying defined areas of the HIV genome for sequence analysis and has become an important technique for studies of sequence diversity and microbial resistance to antiretroviral agents. In patients with a positive or indeterminateEIA test and an indeterminate western blot, and in patients in whom serologic testing may be unreliable (such as patients with hypogammaglobulinemia or advanced HIV disease), these tests provide valuable tools for making a diagnosis of HIV infection. They should only be used for diagnosis when standard serologic testing has failed to provide a definitive result.

LABORATORY MONITORING OF PATIENTS WITH HIV INFECTION

The epidemic of HIV infection and AIDS has provided the clinician with new challenges for integrating clinical and laboratory data to effect optimal patient management. The close relationship between clinical manifestations of HIV infection and CD4+ T cell count has made measurement of the latter a routine part of the evaluation of HIV-infected individuals. Determinations of CD4+ T cell counts and measurements of the levels of

HIV RNA in serum or plasma provide a powerful set of tools for determining prognosis and monitoring response to therapy. While the CD4+ T cell count provides information on the current immunologic status of the patient, the HIV RNA level predicts what will happen to the CD4+ T cell count in the near future, and hence provides an important piece of prognostic information.

CD4+ T Cell Counts The CD4+ T cell count is the laboratory test generally accepted as the best indicator of the immediate state of immunologic competence of the patient with HIV infection. This measurement, which is the product of the percent of CD4+ T cells (determined by flow cytometry) and the total lymphocyte count [determined by the white blood cell count (WBC) and the differential count] has been shown to correlate very well with the level of immunologic competence. Patients with CD4+ T cell counts <200/uL are at high risk of infection with *P. carinii*, while patients with CD4+ T cell counts<50/uL are at high risk of infection withCMV and mycobacteria of the *M. avium complex* (MAC) (Fig. 309-26). Patients with HIV infection should have CD4+ T cell measurements performed at the time of diagnosis and every 3 to 6 months thereafter. More frequent measurements should be made if a declining trend is noted. According to most guidelines, a CD4 T cell count<500/uL is an indication for consideration of initiating antiretroviral therapy, and a decline in CD4+ T cell count of>25% is an indication for considering a change in therapy. Once the CD4+ T cell count is <200/uL, patients should be placed on a regimen for *P. carinii* prophylaxis, and once the count is <50/uL, primary prophylaxis for MAC infection is indicated. As with any laboratory measurement, one may wish to obtain two determinations prior to any significant changes in patient management based upon CD4+ T cell count alone.

HIV RNA Determinations Facilitated by highly sensitive techniques for the precise quantitation of small amounts of nucleic acids, the measurement of serum or plasma levels of HIV RNA has become an essential component in the monitoring of patients with HIV infection. As discussed under diagnosis of HIV infection, the two most commonly used techniques are theRT-PCRassay and the bDNA assay. Both assays generate data in the form of number of copies of HIV RNA per milliliter of serum or plasma and, by employing a 1:10 concentration step with ultracentrifugation, can detect as few as 40 to 50 copies of HIV RNA per milliliter of plasma. Although earlier versions of the bDNA assay generated values that were approximately 50% of those of the RT-PCR assay, the more recent versions (version 3 or higher) provide numbers essentially identical to those of the RT-PCR test (Fig. 309-25). While it is common practice to describe levels of HIV RNA below these cut-offs as "undetectable," this is a term that should be avoided as it is imprecise and leaves the false impression that the level of virus is 0. By utilizing more sensitive, nestedPCRtechniques and by studying tissue levels of virus as well as plasma levels, HIV RNA can be detected in virtually every patient with HIV infection. Measurements of changes in HIV RNA levels over time have been of great value in delineating the relationship between levels of virus and rates of disease progression (Fig. 309-17), the rates of viral turnover, the relationship between immune system activation and viral replication, and the time to development of drug resistance. HIV RNA measurements are greatly influenced by the state of activation of the immune system and may fluctuate greatly in the setting of secondary infections or immunization. For these reasons, decisions based upon HIV RNA levels should never be made on a single determination. Measurements of plasma HIV RNA levels should be made at the time of HIV diagnosis and every 3 to 4 months thereafter in the untreated patient. In general, most guidelines suggest that therapy be initiated in patients with>20,000 copies of HIV RNA per milliliter (see below). Following the initiation of therapy or any change in therapy, plasma HIV RNA levels should be monitored approximately every 4 weeks until the effectiveness of the therapeutic regimen is determined by the development of a new steady-state level of HIV RNA. In most instances of effective therapy this will be <50 copies per milliliter. This level of virus is generally achieved within 6 months of the initiation of effective treatment. During therapy, levels of HIV RNA should be monitored every 3 to 4 months to evaluate the continuing effectiveness of therapy.

HIV Resistance Testing The availability of multiple antiretroviral drugs as treatment options has generated a great deal of interest in the potential for measuring the sensitivity of an individual's HIV virus(es) to different antiretroviral agents. HIV resistance testing can be done through either genotypic or phenotypic measurements. In the genotypic assays, sequence analyses of the HIV genomes obtained from patients are compared to sequences of viruses with known antiretroviral resistance profiles. In the phenotypic assays, the in vivo growth of viral isolates obtained from the patient are compared to the growth of reference strains of the virus in the presence or absence of different antiretroviral drugs. A modification of this phenotypic approach utilizes a comparison of the enzymatic activities of the reverse transcriptase or protease genes obtained by molecular cloning of patients' isolates to the enzymatic activities of genes obtained from reference strains of HIV in the presence or absence of different drugs targeted to these genes. These tests are quite good in identifying those antiretroviral agents that have been utilized in the past in a given patient. Their clinical value in identifying which antiretroviral regimen is best for an individual patient is still under investigation.

Other Tests A variety of other laboratory tests have been studied as potential markers of HIV disease activity. Among these are quantitative culture of replication-competent HIV from plasma, peripheral blood mononuclear cells, or resting CD4+ T cells; circulating levels ofb$_2$-microglobulin, solubleIL-2 receptor, IgA, acid-labile endogenous interferon, orTNF-a; and the presence or absence of activation markers such as CD38 or HLA-DR on CD8+ T cells. While these measurements have value as markers of disease activity and help to increase our understanding of the pathogenesis of HIV disease, they do not currently play a major role in the monitoring of patients with HIV infection.

CLINICAL MANIFESTATIONS

The clinical consequences of HIV infection encompass a spectrum ranging from an acute syndrome associated with primary infection to a prolonged asymptomatic state to advanced disease. It is best to regard HIV disease as beginning at the time of primary infection and progressing through various stages. As mentioned above, active virus replication and progressive immunologic impairment occur throughout the course of HIV infection in most patients. With the exception of long-term nonprogressors (see above), HIV disease in untreated patients inexorably progresses even during the clinically latent stage.

THE ACUTE HIV SYNDROME

It is estimated that 50 to 70% of individuals with HIV infection experience an acute clinical syndrome approximately 3 to 6 weeks after primary infection (Fig. 309-27). Varying degrees of clinical severity have been reported, and although it has been suggested that symptomatic seroconversion leading to the seeking of medical attention indicates an increased risk for an accelerated course of disease, this has not been shown definitively. In fact, there does not appear to be a correlation between the level of the initial burst of viremia in acute HIV infection and the subsequent course of disease. The typical clinical findings are listed inTable 309-10; they occur along with a burst of plasma viremia. The syndrome is typical of an acute viral syndrome and has been likened to acute infectious mononucleosis. Symptoms usually persist for 1 to several weeks and gradually subside as an immune response to HIV develops and the levels of plasma viremia decrease. Opportunistic infections have been reported during this stage of infection, reflecting the immunodeficiency that results from reduced numbers of CD4+ T cells and likely also from the dysfunction of CD4+ T cells owing to cross-linking of the CD4 molecule on the cell surface by viral envelope proteins (see "Mechanisms of CD4+ T Lymphocyte Depletion and Dysfunction," above) associated with the extremely high levels of plasma viremia. A number of immunologic abnormalities accompany the acute HIV syndrome, including multiphasic perturbations of the numbers of circulating lymphocyte subsets. The number of total lymphocytes and T cell subsets (CD4+ and CD8+) are initially reduced. An inversion of the CD4+/CD8+ T cell ratio occurs later because of a rise in the number of CD8+ T cells. In fact, there may be a selective and transient expansion of CD8+ T cell subsets, as determined by T cell receptor analysis (see above). The total circulating CD8+ T cell count may remain elevated or return to normal; however, CD4+ T cell levels usually remain somewhat depressed, although there may be a slight rebound towards normal. Lymphadenopathy occurs in approximately 70% of individuals with primary HIV infection. Most patients recover spontaneously from this syndrome and many are left with only a mildly depressed CD4+ T cell count that remains stable for a variable period before beginning its progressive decline (see below); in some individuals, the CD4+ T cell count returns to the normal range. Approximately 10% of patients manifest a fulminant course of immunologic and clinical deterioration after primary infection, even after the disappearance of symptoms. In most patients, primary infection with or without the acute syndrome is followed by a prolonged period of clinical latency.

THE ASYMPTOMATIC STAGE -- CLINICAL LATENCY

Although the length of time from initial infection to the development of clinical disease varies greatly, the median time for untreated patients is approximately 10 years. As emphasized above, HIV disease with active virus replication is ongoing and progressive during this asymptomatic period. The rate of disease progression is directly correlated with HIV RNA levels. Patients with high levels of HIV RNA in plasma progress to symptomatic disease faster than do patients with low levels of HIV RNA (Fig. 309-17). Some patients referred to as long-term nonprogressors show little if any decline in CD4+ T cell counts over extended periods of time. These patients generally have extremely low levels of HIV RNA. Certain other patients remain entirely asymptomatic despite the fact that their CD4+ T cell counts show a steady progressive decline to extremely low levels. In these patients, the appearance of an opportunistic disease may be the first manifestation of HIV infection. During the asymptomatic period of HIV infection, the average rate of CD4+ T cell decline is approximately 50/uL per year. When the CD4+ T cell count falls to <200/uL, the resulting state of immunodeficiency is severe enough to place the patient at high risk for opportunistic infection and neoplasms, and hence for clinically apparent disease.

SYMPTOMATIC DISEASE

Symptoms of HIV disease can appear at any time during the course of HIV infection. Generally speaking, the spectrum of illness that one observes changes as the CD4+ T cell count declines. The more severe and life-threatening complications of HIV infection occur in patients with CD4+ T cells counts <200/uL. A diagnosis of AIDS is made in anyone with HIV infection and a CD4+ T cell count<200/uL and in anyone with HIV infection who develops one of the HIV-associated diseases considered to be indicative of a severe defect in cell-mediated immunity (category C,Table 309-2). While the causative agents of the secondary infections are characteristically opportunistic organisms such as *P. carinii*, atypical mycobacteria,CMV, and other organisms that do not ordinarily cause disease in the absence of a compromised immune system, they also include common bacterial and mycobacterial pathogens. Approximately 80% of deaths among AIDS patients are as a direct result of an infection other than HIV, with bacterial infections heading the list. Following the widespread use of combination antiretroviral therapy and implementation of guidelines for the prevention of opportunistic infections (Table 309-11), the incidence of secondary infections has decreased dramatically (Fig. 309-28). Overall, the clinical spectrum of HIV disease is constantly changing as patients live longer and new and better approaches to treatment and prophylaxis are developed. In general, it should be stressed that a key element of treatment of symptomatic complications of HIV disease, whether they are primary or secondary, is achieving good control of HIV replication through the use of combination antiretroviral therapy and instituting primary and secondary prophylaxis as indicated.

Disease of the Respiratory System Acute bronchitis and sinusitis are prevalent during all stages of HIV infection. The most severe cases tend to occur in patients with lower CD4+ T cell counts. Sinusitis presents as fever, nasal congestion, and headache. The diagnosis is made by computed tomography (CT) or magnetic resonance imaging (MRI). The maxillary sinuses are most commonly involved; however, disease is also frequently seen in the ethmoid, sphenoid, and frontal sinuses. While some patients may improve without antibiotic therapy, radiographic improvement is quicker and more pronounced in patients who have received antimicrobial therapy. It is postulated that this high incidence of sinusitis results from an increased frequency of infection with encapsulated organisms such as *H. influenzae* and *Streptococcus pneumoniae*. In patients with low CD4+ T cell counts one may see mucormycosis infections of the sinuses. In contrast to the course of this infection in other patient populations, mucormycosis of the sinuses in patients with HIV infection may progress more slowly. In this setting aggressive, frequent local debridement in addition to local and systemic amphotericin B may be needed for effective treatment.

Pulmonary disease is one of the most frequent complications of HIV infection. The most common manifestation of pulmonary disease is pneumonia. The two most common causes of pneumonia are bacteria infections and *P. carinii* infection. Other major causes of pulmonary infiltrates include mycobacterial infections, fungal infections, nonspecific interstitial pneumonitis, KS, and lymphoma.

Pneumonia is seen with an increased frequency in patients with HIV infection. Patients with HIV infection appear to be particularly prone to infections with encapsulated organisms. *S. pneumoniae* (Chap. 138) and *H. influenzae* (Chap. 149) are responsible for most cases of bacterial pneumonia in patients with AIDS. This may be a consequence of altered B cell function and/or defects in neutrophil function that may be secondary to HIV disease (see above). Pneumococcal infection may be the earliest serious infection to occur in patients with HIV disease. This can present as pneumonia, sinusitis, and/or bacteremia. Patients with HIV infection have a sixfold increase in the incidence of pneumococcal pneumonia and a 100-fold increase in the incidence of pneumococcal bacteremia. Pneumococcal disease may be seen in patients with relatively intact immune systems. In one study, the baseline CD4+ T cell count at the time of a first episode of pneumococcal pneumonia was ~300/uL. Of interest is the fact that the inflammatory response to pneumococcal infection appears proportional to the CD4+ T cell count. Due to this high risk of pneumococcal disease, immunization with pneumococcal polysaccharide is one of the generally recommended prophylactic measures for patients with HIV infection and CD4+ T cell counts >200/uL. It is less clear if this intervention is of benefit in patients with more advanced disease and high viral loads.

P. carinii pneumonia (PCP), once the hallmark of AIDS, has dramatically declined in incidence following the development of effective prophylactic regimens and the widespread use of combination antiretroviral therapy. The risk of PCP is most common among those who have experienced a previous bout of PCP and those who have CD4+ T cell counts of <200/uL. Overall, 79% of patients with PCP have CD4+ T cell counts <100/uL and 95% of patients have CD4+ T cell counts <200/uL. Recurrent fever, night sweats, thrush, and unexplained weight loss are also associated with an increased incidence of PCP. For these reasons, it is strongly recommended that all patients with CD4+ T cell counts<200/uL (or a CD4 percentage<15) receive some form of PCP prophylaxis. At present the incidence of PCP is approaching zero in patients with known HIV infection receiving appropriate antiretroviral therapy and prophylaxis. Primary PCP is now occurring at a median CD4+ T cell count of 36/uL, while secondary PCP is occurring at a median CD4+ T cell count of 10/uL. Patients with PCP generally present with fever and a cough that is usually nonproductive or productive of only scant amounts of white sputum. They may complain of a characteristic retrosternal chest pain that is worse on inspiration and is described as sharp or burning. HIV-associated PCP may have an indolent course characterized by weeks of vague symptoms and should be included in the differential diagnosis of fever, pulmonary complaints, or unexplained weight loss in any patient with HIV infection and <200 CD4+ T cells/uL. The most common finding on chest x-ray is either a normal film, if the disease is suspected early, or a faint bilateral interstitial infiltrate. The classic finding of a dense perihilar infiltrate is unusual in patients with AIDS. In patients with PCP who have been receiving aerosolized pentamidine for prophylaxis, one may see an x-ray picture of upper lobe cavitary disease, reminiscent ofTB. Other less common findings on chest x-ray include lobar infiltrates and pleural effusions. Routine laboratory evaluation is usually of little help in the differential diagnosis of PCP. A mild leukocytosis is common, although this may not be obvious in patients with prior neutropenia. Arterial blood gases may indicate hypoxemia with a decline in Pa_{O2} and an increase in the arterial-alveolar (a - A) gradient. Arterial blood gas measurements not only aid in making the diagnosis of PCP but also provide important information for staging the severity of the disease and directing treatment. A definitive diagnosis of PCP requires demonstration of the trophozoite or cyst form of the organism in samples obtained from induced sputum, bronchoalveolar lavage, transbronchial biopsy, or open lung biopsy.PCR has been used to detect specific DNA sequences for *P. carinii* in clinical specimens where histologic examinations have failed to make a diagnosis.

In addition to pneumonia, a number of other clinical problems have been reported in HIV-infected patients as a result of infection with *P. carinii*. Otic involvement may be seen as a primary infection, presenting as a polypoid mass involving the external auditory canal. In patients receiving aerosolized penamidine for prophylaxis againstPCP one may see a variety of extrapulmonary manifestations of *P. carinii*. These include ophthalmic lesions of the choroid, a necrotizing vasculitis that resembles Burger's disease, bone marrow hypoplasia, and intestinal obstruction. Other organs that have been involved include lymph nodes, spleen, liver, kidney, pancreas, pericardium, heart, thyroid, and adrenals. Organ infection may be associated with cystic lesions that may appear calcified onCT or ultrasound. The standard treatment for PCP or disseminated pneumocystosis is trimethoprim/sulfamethoxazole (TMP/SMZ). A high incidence of side effects, particularly skin rash and bone marrow suppression, is seen with TMP/SMZ in patients with HIV infection. Alternative treatments for mild to moderate PCP include dapsone/trimethoprim and clindamycin/primaquine. Intravenous pentamidine is the treatment of choice for severe disease in the patient unable to tolerate TMP/SMZ. For patients with a Pa_{O2}<70 mmHg or with an a - A gradient >35 mmHg, adjunct glucocorticoid therapy should be used in addition to specific antimicrobials. Overall, treatment should be for 21 days and followed by secondary prophylaxis. Prophylaxis for PCP is indicated for any HIV-infected individual who has experienced a prior bout of PCP, any patient with a CD4+ T cell count of<200/uL or a CD4 percentage<15, any patient with unexplained fever for>2 weeks, and any patient with a history of oropharyngeal candidiasis. The preferred regimen for prophylaxis is TMP/SMZ, one double-strength tablet daily. This regimen also provides protection against toxoplasmosis and some bacterial respiratory pathogens. For patients who cannot tolerate TMP/SMZ, alternatives include dapsone plus pyrimethamine plus leucovorin, aerosolized pentamidine administered by the Respirgard II nebulizer, and atovaquone. Primary prophylaxis for PCP can be discontinued in those patients treated with combination antiretroviral therapy who maintain good suppression of HIV (<500 copies per milliliter) and CD4+ T cell counts >200/uL for at least 3 to 6 months. There is as yet insufficient information to know if the same recommendation will hold for discontinuation of secondary prophylaxis.

M. tuberculosis, once thought to be on its way to extinction in the United States, experienced a resurgence associated with the HIV epidemic (Chap. 169). Worldwide, approximately one-third of all AIDS-related deaths are associated withTB. In the United States approximately 5% of AIDS patients have active TB. HIV infection increases the risk of developing active tuberculosis by a factor of 15 to 30. For the patient with untreated HIV infection and a positivePPD skin test, the rate of reactivation TB is 7 to 10% per year. Untreated TB can accelerate the course of HIV infection. Levels of plasma HIV RNA increase in the setting of active TB and decline in the setting of successful TB treatment. Active TB is most common in patients 25 to 44 years of age, in

African-Americans and Hispanics, in patients in New York City and Miami, and in patients in developing countries. In these demographic groups, 20 to 70% of the new cases of active TB are in patients with HIV infection. The epidemic of TB embedded in the epidemic of HIV infection probably represents the greatest health risk to the general public and the health care profession associated with the HIV epidemic. In contrast to infection with atypical mycobacteria such asMAC, active TB often develops relatively early in the course of HIV infection and may be an early clinical sign of HIV disease. In one study, the median CD4+ T cell count at presentation of TB was 326/uL. The clinical manifestations of TB in HIV-infected patients are quite varied and generally show different patterns as a function of the CD4+ T cell count. In patients with relatively high CD4+ T cell counts, the typical pattern of pulmonary reactivation occurs in which patients present with fever, cough, dyspnea on exertion, weight loss, night sweats, and a chest x-ray revealing cavitary apical disease of the upper lobes. In patients with lower CD4+ T cell counts, disseminated disease is more common. In these patients the chest x-ray may reveal diffuse or lower lobe bilateral reticulonodular infiltrates consistent with miliary spread, pleural effusions, and hilar and/or mediastinal adenopathy. Infection may be present in bone, brain, meninges, gastrointestinal tract, lymph nodes (particularly cervical lymph nodes), and viscera. Approximately 60 to 80% of patients have pulmonary disease, and 30 to 40% have extrapulmonary disease. Respiratory isolation and a negative-pressure room should be used for patients in whom a diagnosis of pulmonary TB is being considered. This approach is critical to limit nosocomial and community spread of infection. Culture of the organism from an involved site provides a definitive diagnosis. Blood cultures are positive in 15% of patients. In the setting of fulminant disease one cannot rely upon the accuracy of a negative PPD skin test to rule out a diagnosis of TB. TB is one of the conditions associated with HIV infection for which cure is possible. Therapy for TB is generally the same in the HIV-infected patient as in the HIV-negative patient (Chap. 169). Due to pharmacokinetic interactions, rifabutin should be substituted for rifampin in patients receiving the HIV protease inhibitors or nonnucleoside reverse transcriptase inhibitors; both drugs should be avoided in patients receiving ritonavir. Treatment is most effective in programs that involve directly observed therapy. Effective prevention of active TB can be a reality if the health care professional is aggressive in looking for evidence of latent TB by making sure that all patients with HIV infection receive a PPD skin test. Anergy testing is not of value in this setting. HIV-infected individuals with a skin test reaction of>5 mm or those who are close household contacts of persons with active TB should receive treatment with 9 months of isoniazid, or 2 months of therapy with rifampin and pyrazinamide, or 2 months of therapy with rifabutin and pyrazinamide.

Atypical mycobacterial infections are also seen with an increased frequency in patients with HIV infection. Infections with at least 12 different mycobacteria have been reported, including *M. bovis* and representatives of all four Runyon groups. The most common atypical mycobacterial infection is with *M. avium* or *M. intracellulare* species, the *M. avium* complex (MAC). Infections with MAC are seen mainly in patients in the United States and are rare in Africa. It has been suggested that prior infection with *M. tuberculosis* decreases the risk of MAC infection. MAC infections probably arise from organisms that are ubiquitous in the environment, including both soil and water. The presumed portals of entry are the respiratory and gastrointestinal tract. MAC infection is a late complication of HIV infection, occurring in patients with CD4+ T cell counts of<50/uL. The average CD4+ T cell count at the time of diagnosis is 10/uL. The most common presentation is disseminated disease with fever, weight loss, and night sweats. At least 85% of patients with MAC infection are mycobacteremic, and large numbers of organisms can often be demonstrated on bone marrow biopsy. The chest x-ray is abnormal in approximately 25% of patients, with the most common pattern being that of a bilateral, lower lobe infiltrate suggestive of miliary spread. Alveolar or nodular infiltrates and hilar and/or mediastinal adenopathy can also occur. Other clinical findings include endobronchial lesions, abdominal pain, diarrhea, and lymphadenopathy. The diagnosis is made by the culture of blood or involved tissue. The finding of two consecutive sputum samples positive for MAC is highly suggestive of pulmonary infection. Cultures may take 2 weeks to turn positive. Therapy consists of a macrolide, usually clarithromycin, with ethambutol. Some physicians elect to add a third drug from among rifabutin, ciprofloxacin, or amikacin in patients with extensive disease. Therapy is generally for life; however, with the advent of highly active antiretroviral therapy (HAART), it may be possible to discontinue therapy in patients with sustained suppression of HIV replication and CD4+ T cell counts >100/uL for>6 months. Primary prophylaxis for MAC is indicated in patients with HIV infection and CD4+ T cell counts<50/uL. This may be discontinued in patients in whom HAART induces a sustained suppression of viral replication and increases in CD4+ T cell counts to >100/uL for 3 to 6 months.

Rhodococcus equi is a gram-positive pleomorphic acid-fast non-spore-forming bacillus that can cause pulmonary and/or disseminated infection in patients with HIV infection. Fever and cough are the most common presenting signs. Radiographically one may see cavitary lesions and consolidation. Blood cultures are often positive. Treatment is based upon antimicrobial sensitivity testing.

Fungal infections of the lung can be seen in patients with AIDS. Patients with pulmonary cryptococcal disease present with fever, cough, dyspnea, and in some cases, hemoptysis. A focal or diffuse interstitial infiltrate is seen on chest x-ray in>90% of patients. In addition, one may see lobar disease, cavitary disease, pleural effusions, and hilar or mediastinal adenopathy. Over half of patients are fungemic, and 90% of patients have concomitantCNSinfection. *Coccidioides immitis* is a mold that is endemic in the southwest United States. It can cause a reactivation pulmonary syndrome in patients with HIV infection. Most patients with this condition will have CD4+ T cell counts <250/uL. Patients present with fever, weight loss, cough, and extensive, diffuse reticulonodular infiltrates on chest x-ray. One may also see nodules, cavities, pleural effusions, and hilar adenopathy. While serologic testing is of value in the immunocompetent host, serologies are negative in 25% of HIV-infected patients with coccidioidal infection. Invasive aspergillosis is not an AIDS-defining illness and is generally not seen in patients with AIDS in the absence of neutropenia or administration of glucocorticoids. *Aspergillus* infection may have an unusual presentation in the respiratory tract of patients with AIDS where it gives the appearance of a pseudomembranous tracheobronchitis. Primary pulmonary infection of the lung may be seen with *histoplasmosis*. The most common pulmonary manifestation of histoplasmosis, however, is in the setting of disseminated disease, presumably due to reactivation. In this setting respiratory symptoms are usually minimal, with cough and dyspnea occurring in 10 to 30% of patients. The chest x-ray is abnormal in about 50% of patients, showing either a diffuse interstitial infiltrate or diffuse small nodules.

Two forms of *idiopathic interstitial pneumonia* have been identified in patients with HIV infection: lymphoid interstitial pneumonitis (LIP) and nonspecific interstitial pneumonitis (NIP). LIP, a common finding in children, is seen in about 1% of adult patients with HIV infection. This disorder is characterized by a benign infiltrate of the lung and is felt to be part of the polyclonal activation of lymphocytes seen in the context of HIV andEBVinfections. Transbronchial biopsy is diagnostic in 50% of the cases, with an open-lung biopsy required for diagnosis in the remainder of cases. This condition is generally self-limited and no specific treatment is necessary. Severe cases have been managed with brief courses of glucocorticoids. Although rarely a clinical problem since the use ofHAART, evidence of NIP may be seen in up to half of all patients with untreated HIV infection. Histologically, interstitial infiltrates of lymphocytes and plasma cells in a perivascular and peribronchial distribution are present. When symptomatic, patients present with fever and nonproductive cough occasionally accompanied by mild chest discomfort. Chest x-ray is usually normal or may reveal a faint interstitial pattern. Similar to LIP, this is a self-limited process for which no therapy is indicated.

Neoplastic diseases of the lung includingKS and lymphoma are discussed below in the section on malignancies.

Diseases of the Cardiovascular System While heart disease is a relatively common postmortem finding in HIV-infected patients (25 to 75% in autopsy series), it is less of a problem clinically. The most common clinically significant finding is a dilated cardiomyopathy associated with congestive heart failure referred to as *HIV-associated cardiomyopathy*. This generally occurs as a late complication of HIV infection and, histologically, displays elements of myocarditis. For this reason some have advocated treatment with intravenous Ig. HIV can be directly demonstrated in cardiac tissue in this setting, and there is debate over whether or not it plays a direct role in this condition. Patients present with typical findings of congestive heart failure, namely edema and shortness of breath. Patients with HIV infection may also develop cardiomyopathy as a side effect ofIFN-aor nucleoside analogue therapy, which is reversible once therapy is stopped.KS, cryptococcosis, Chagas disease, and toxoplasmosis can involve the myocardium, leading to cardiomyopathy. In one series, most patients with HIV infection and a treatable myocarditis were found to have myocarditis associated with toxoplasmosis. Most of these patients also had evidence ofCNStoxoplasmosis. Thus,MRI or double-dose contrastCT scan of the brain should be included in the workup of any patient with advanced HIV infection and cardiomyopathy.

A variety of other cardiovascular problems are found in patients with HIV infection. Pericardial effusions may be seen in the setting of advanced HIV infection. Predisposing factors includeTB, congestive heart failure, mycobacterial infection, cryptococcal infection, pulmonary infection, lymphoma, andKS. While pericarditis is quite rare, in one series 5% of patients with HIV disease had pericardial effusions that were considered to be moderate or severe. Tamponade and death have occurred in association with pericardial KS, presumably owing to acute hemorrhage. A high percentage of patients have hypertriglyceridemia and elevations in serum cholesterol, and coronary artery disease has been a relatively frequent finding at autopsy. This problem appears to becoming even more prevalent as a side effect ofHAART. While clinically significant ischemic heart disease has not been reported to be occurring with an increased frequency in this patient population, many are concerned that it is just a matter of time before this is the case. Nonbacterial thrombotic endocarditis has been reported and should be considered in patients with unexplained embolic phenomena. Intravenous pentamidine, when given rapidly, can result in hypotension as a consequence of cardiovascular collapse.

Diseases of the Oropharynx and Gastrointestinal System Oropharyngeal and gastrointestinal diseases are common features of HIV infection. They are most frequently due to secondary infections. In addition, oral and gastrointestinal lesions may occur withKS and lymphoma.

Oral lesions, including *thrush*, *hairy leukoplakia*, and *aphthous ulcers*, are particularly common in patients with untreated HIV infection. Thrush, due to *Candida* infection, and oral hairy leukoplakia, presumed due toEBV, are usually indicative of fairly advanced immunologic decline; they generally occur in patients with CD4+ T cell counts of <300/uL. In one study, 59% of patients with oral candidiasis went on to develop AIDS in the next year. Thrush appears as a white, cheesy exudate, often on an erythematous mucosa in the posterior oropharynx (seePlate IID-43). While most commonly seen on the soft palate, early lesions are often found along the gingival border. The diagnosis is made by direct examination of a scraping for pseudohyphal elements. Culturing is of no diagnostic value, as most patients with HIV infection will have a positive throat culture for *Candida* even in the absence of thrush. Oral hairy leukoplakia presents as white, frondlike lesions, generally along the lateral borders of the tongue and sometimes on the adjacent buccal mucosa (seePlate IID-42). Despite its name, oral hairy leukoplakia is not considered a premalignant condition. Lesions are associated with florid replication of EBV. While usually more disconcerting as a sign of HIV-associated immunodeficiency than a clinical problem in need of treatment, severe cases have been reported to respond to topical podophyllin or systemic therapy with acyclovir. Aphthous ulcers of the posterior oropharynx are also seen with regularity in patients with HIV infection. These lesions are of unknown etiology and can be quite painful and interfere with swallowing. Topical anesthetics provide immediate symptomatic relief of short duration. The fact that thalidomide is an effective treatment for this condition suggests that the pathogenesis may involve the action of tissue-destructive cytokines. Palatal, glossal, or gingival ulcers may also result from cryptococcal disease or histoplasmosis.

Esophagitis (Fig. 309-29) may present with odynophagia and retrosternal pain. Upper endoscopy is generally required to make an accurate diagnosis. Esophagitis may be due to *Candida*,CMV, orHSV. While CMV tends to be associated with a single large ulcer, HSV infection is more often associated with multiple small ulcers. The esophagus may also be the site ofKS and lymphoma. Like the oral mucosa, the esophageal mucosa may have large, painful ulcers of unclear etiology that may respond to thalidomide. While achlorhydria is a common problem in patients with HIV infection, other gastric problems are generally rare. Among the conditions involving the stomach are KS and lymphoma. Infections of the small and large intestine leading to diarrhea, abdominal pain, and occasionally fever are among the most significant gastrointestinal problems in the HIV-infected patients. They include infections with bacteria, protozoa, and viruses.

Bacteria and fungi may be responsible for secondary infections of the gastrointestinal tract. Infections with enteric pathogens such as *Salmonella*, *Shigella*, and

Campylobacter are more common in homosexual men and are often more severe and more apt to relapse in patients with HIV infection. Patients with untreated HIV have approximately a 20-fold increased risk of infection with *S. typhimurium*. They may present with a variety of nonspecific symptoms including fever, anorexia, fatigue, and malaise of several weeks' duration. Diarrhea is common but may be absent. Diagnosis is made by culture of blood and stool. Long-term therapy with ciprofloxacin is the recommended treatment. HIV-infected patients also have an increased incidence of *S. typhi* infection in areas of the world where typhoid is a problem. *Shigella* spp., particularly *S. flexneri*, can cause severe intestinal disease in HIV-infected individuals. Up to 50% of patients will develop bacteremia. *Campylobacter* infections occur with an increased frequency in patients with HIV infection. While *C. jejuni* is the strain most frequently isolated, infections with many other strains have been reported. Patients usually present with crampy abdominal pain, fever, and bloody diarrhea. Infection may present as proctitis. Stool examination reveals the presence of fecal leukocytes. Systemic infection can occur, with up to 10% of infected patients exhibiting bacteremia. Most strains are sensitive to erythromycin. Abdominal pain and diarrhea may be seen withMACinfection.

Fungal infections may also be a cause of diarrhea in patients with HIV infection. Histoplasmosis, coccidioidomycosis, and penicilliosis have all been identified as a cause of fever and diarrhea in patients with HIV infection. Peritonitis has been seen with *C. immitis*.

Cryptosporidia, microsporidia, and *Isospora belli* (Chap. 218) are the most common opportunistic protozoa that infect the gastrointestinal tract and cause diarrhea in HIV-infected patients. Cryptosporidial infection may present in a variety of ways, ranging from a self-limited or intermittent diarrheal illness in patients in the early stages of HIV infection to a severe, life-threatening diarrhea in severely immunodeficient individuals. In patients with untreated HIV infection and CD4+ T cell counts of <300/uL, the incidence of cryptosporidiosis is approximately 1% per year. In 75% of cases the diarrhea is accompanied by crampy abdominal pain, and 25% of patients have nausea and/or vomiting. Cryptosporidia may also cause biliary tract disease in the HIV-infected patient, leading to cholecystitis with or without accompanying cholangitis. The diagnosis of cryptosporidial diarrhea is made by stool examination. The diarrhea is noninflammatory, and the characteristic finding is the presence of oocysts that stain with acid-fast dyes. Therapy is predominantly supportive, and marked improvements have been reported in the setting of effective antiretroviral therapy. Treatment with up to 2000 mg/d of nitazoxanide (NTZ) is associated with improvement in symptoms or a decrease in shedding of organisms in about half of patients. Its overall role in the management of this condition remains unclear. Patients can minimize their risk of developing cryptosporidiosis by avoiding contact with human and animal feces and by not drinking untreated water from lakes or rivers.

Microsporidia are small, unicellular, obligate intracellular parasites that reside in the cytoplasm of enteric cells (Chap. 218). The main species causing disease in humans is *Enterocytozoon bieneusi*. The clinical manifestations are similar to those described for Cryptosporidia and include abdominal pain and diarrhea. The small size of the organism may make it difficult to detect; however, with the use of chromotrope-based stains, organisms can be identified in stool samples by light microscopy. Definitive diagnosis generally depends on electron microscopic examination of a stool specimen, intestinal aspirate, or intestinal biopsy specimen. In contrast to cryptosporidia, microsporidia have been noted in a variety of extraintestinal locations, including the eye, muscle, and liver, and have been associated with conjunctivitis and hepatitis. Albendazole, 400 mg bid, has been reported to be of benefit in some patients.

I. belli is a coccidian parasite (Chap. 218) most commonly found as a cause of diarrhea in patients from the Caribbean and Africa. Its cysts appear in the stool as large, acid-fast structures that can be differentiated from those of cryptosporidia on the basis of size, shape, and number of sporocysts. The clinical syndromes of *Isospora* infection are identical to those caused by cryptosporidia. The important distinction is that infection with *Isospora* is generally relatively easy to treat withTMP/SMZ. While relapses are common, a thrice-weekly regimen, similar to that used to provide prophylaxis againstPCP, appears adequate to prevent recurrence.

CMVcolitis was once seen in 5 to 10% of patients with AIDS. It is much less common with the advent ofHAART. CMV colitis presents as diarrhea, abdominal pain, weight loss, and anorexia. The diarrhea is usually nonbloody, and the diagnosis is achieved through endoscopy and biopsy. Multiple mucosal ulcerations are seen at endoscopy, and biopsies reveal characteristic intranuclear inclusion bodies. Bacteremia may result as a consequence of thinning of the bowel wall. Treatment is with either ganciclovir or foscarnet for 3 to 6 weeks. Relapses are common, and maintenance therapy is typically necessary in patients whose HIV infection is poorly controlled. Patients with CMV disease of the gastrointestinal tract should be carefully monitored for evidence of retinitis.

In addition to disease caused by specific secondary infections, patients with HIV infection may also experience a chronic diarrheal syndrome for which no etiologic agent other than HIV can be identified. This entity is referred to as *AIDS enteropathy* or *HIV enteropathy*. It is most likely a direct result of HIV infection in the gastrointestinal tract. Histologic examination of the small bowel in these patients reveals low-grade mucosal atrophy with a decrease in mitotic figures, suggesting a hyporegenerative state. Patients often have decreased or absent small-bowel lactase and malabsorption with accompanying weight loss.

The initial evaluation of a patient with HIV infection and diarrhea should include a set of stool examinations, including culture, examination for ova and parasites, and examination for *Clostridium difficile* toxin. Approximately 50% of the time this workup will demonstrate infection with pathogenic bacteria, mycobacteria, or protozoa. If the initial stool examinations are negative, additional evaluation, including upper and/or lower endoscopy with biopsy, will yield a diagnosis of microsporidial or mycobacterial infection of the small intestine ~30% of the time. In patients for whom this diagnostic evaluation is nonrevealing, a presumptive diagnosis of HIV enteropathy can be made if the diarrhea has persisted for >1 month. An algorithm for the evaluation of diarrhea in patients with HIV infection is given inFig. 309-30.

Rectal lesions are common in HIV-infected patients, particularly the perirectal ulcers and erosions due to the reactivation ofHSV(Fig. 309-31). These may appear quite atypical, as denuded skin without vesicles, and they respond well to treatment with acyclovir, famciclovir, or foscarnet. Other rectal lesions encountered in the patients with HIV infection include condylomata acuminata,KS, and intraepithelial neoplasia.

Hepatobiliary Disease Diseases of the hepatobiliary system are a major problem in patients with HIV infection. It has been estimated that approximately one-third of the deaths of patients with HIV infection are in some way related to liver disease. While this is predominantly a reflection of the problems encountered in the setting of co-infection with hepatitis B or C, it is also a reflection of the hepatic injury, predominantly in the form of hepatic steatosis, that can be seen in the context of nucleoside analogue antiretroviral therapy.

Over 95% of HIV-infected individuals have evidence of infection withHBV; 5-40% of patients are co-infected with hepatitis C virus (HCV); and co-infection with hepatitis D, E, and/or G viruses is common. HIV infection has a significant impact on the course of hepatitis virus infection. It is associated with approximately a threefold increase in the development of persistent hepatitis B surface antigenemia. Patients infected with both HBV and HIV have decreased evidence of inflammatory liver disease. The presumption that this is due to the immunosuppressive effects of HIV infection is supported by the observations that this situation can be reversed, and one may see the development of more severe hepatitis following the initiation of effective antiretroviral therapy.IFN-ais less successful as a treatment of HBV in patients with HIV co-infection, and lamivudine is the treatment of choice. It is important to remember that lamivudine is also a potent antiretroviral agent in the setting of combination antiretroviral therapy. It should not be used as a single agent in patients with HIV infection, even if it is only being used to treat HBV, in order to avoid the rapid development of lamivudine-resistant quasispecies of HIV. In contrast to the situation with HBV, HCV infection is more severe in the patient with HIV infection. In the setting of HIV and HCV co-infection, levels of HCV are approximately tenfold higher than in the HIV-negative patient with HCV infection. The incidence of HCV-associated liver failure appears to be higher by a similar factor in patients with HIV infection. Hepatitis A virus infection is not seen with an increased frequency in patients with HIV infection. It is recommended that all patients with HIV infection who have not experienced natural infection be immunized with hepatitis A and/or hepatitis B vaccines.

A variety of other infections may also involve the liver. Granulomatous hepatitis may be seen as a consequence of mycobacterial or fungal infections, particularlyMACinfection. Hepatic masses may be seen in the context ofTB, peliosis hepatis, or fungal infection. Among the fungal opportunistic infection *C. immitis* and *Histoplasma capsulatum* are those most likely to involve the liver. Biliary tract disease in the form of papillary stenosis or sclerosing cholangitis has been reported in the context of cryptosporidiosis,CMVinfection, andKS.

Many of the drugs used to treat HIV infection are metabolized by the liver and can cause liver injury. Nucleoside analogues work by inhibiting DNA synthesis. This can result in toxicity to mitochondria, which can lead to disturbances in oxidative metabolism. This may be manifest as hepatic steatosis and, in severe cases, lactic acidosis and fulminant liver failure. It is important to be aware of this condition and to watch for it in patients with HIV infection receiving nucleoside analogues. It is reversible if diagnosed early and the offending agent(s) discontinued. Indinavir may cause mild to moderate elevations in serum bilirubin in 10 to 15% of patients in a syndrome similar to Gilbert's syndrome.

Pancreatic injury is most commonly a consequence of drug toxicity, notably that secondary to pentamidine or dideoxynucleosides. While up to half of patients in some series have biochemical evidence of pancreatic injury, <5% of patients show any clinical evidence of pancreatitis that is not linked to a drug toxicity.

Diseases of the Kidney and Genitourinary Tract Diseases of the kidney or genitourinary tract may be a direct consequence of HIV infection, due to an opportunistic infection or neoplasm, or related to drug toxicity. *HIV-associated nephropathy* was first described inIDUs and was initially thought to be IDU nephropathy in patients with HIV infection; it is now recognized as a true direct complication of HIV infection. HIV-associated nephropathy can be an early manifestation of HIV infection and is also seen in children. Over 90% of reported cases have been in African-American or Hispanic individuals; the disease is not only more prevalent in these populations but also more severe. Proteinuria is the hallmark of this disorder. Overall, microalbuminuria is seen in ~20% of untreated HIV-infected patients; significant proteinuria is seen in closer to 2%. Edema and hypertension are rare. Ultrasound examination reveals enlarged, hyperechogenic kidneys. A definitive diagnosis is obtained through renal biopsy. Histologically, focal segmental glomerulosclerosis is present in 80%, and mesangial proliferation in 10 to 15% of cases. Prior to effective antiretroviral therapy, this disease was characterized by relatively rapid progression to end-stage renal disease. Treatment with prednisone, 60 mg/d, has been reported to be of benefit in some cases. The incidence of this disease in patients receiving adequate antiretroviral therapy has not been well defined; however, the impression is that it has decreased in frequency. It is the leading cause of end-stage renal disease in patients with HIV infection.

Among the drugs commonly associated with renal damage in patients with HIV disease are pentamidine, amphotericin, adefovir, cidofovir, and foscarnet.TMP/SMZmay compete for tubular secretion with creatinine and cause an increase in the serum creatinine level. Sulfadiazine may crystallize in the kidney and result in an easily reversible form of renal shutdown. One of the most common drug-induced renal complications is indinavir-associated renal calculi. This condition is seen in ~10% of patients receiving this HIV protease inhibitor. It may present with a variety of manifestations, ranging from asymptomatic hematuria to renal colic. Adequate hydration is the mainstay of treatment and prevention for this condition.

Genitourinary tract infections are seen with a high frequency in patients with HIV infection; they present with dysuria, hematuria, and/or pyuria and are managed in the same fashion as in patients with HIV infection. Infections with *T. pallidum*, the etiologic agent of *syphilis*, play an important role in the HIV epidemic (Chap. 172). In HIV-negative individuals, genital syphilitic ulcers as well as the ulcers of chancroid are major predisposing factors for heterosexual transmission of HIV infection. While most HIV-infected individuals with syphilis have a typical presentation, a variety of formerly rare clinical problems may be encountered in the setting of dual infection. Among them are *lues maligna*, an ulcerating lesion of the skin due to a necrotizing vasculitis; unexplained fever; nephrotic syndrome; and neurosyphilis. The most common presentation of syphilis in the HIV-infected patient is that of *condylomata lata*, a form of secondary syphilis. Neurosyphilis may be asymptomatic or may present as acute meningitis, neuroretinitis, deafness, or stroke. The rate of neurosyphilis may be as high as 1% in patients with HIV infection. As a consequence of the immunologic abnormalities seen in the setting of HIV infection, diagnosis of syphilis through standard serologic testing may be challenging. On the one hand, a significant number of patients have false-positive Venereal Disease Research Laboratory (VDRL) tests due to polyclonal B cell activation. On the other hand, the development of a new positive VDRL may be delayed in patients with new infections, and the anti-fluorescent treponema antibody (anti-FTA) test may be negative due to immunodeficiency. Thus, dark-field examination of appropriate specimens should be performed in any patient in whom syphilis is suspected, even if the patient has a negative VDRL. Similarly, any patient with a positive serum VDRL test, neurologic findings, and an abnormal spinal fluid examination should be considered to have neurosyphilis, regardless of theCSF VDRL result. In any setting, patients treated for syphilis need to be carefully monitored to ensure adequate therapy.

Vulvovaginal candidiasis is a common problem in women with HIV infection. Symptoms include pruritus, discomfort, dyspareunia, and dysuria. Vulvar infection may present as a moribilliform rash that may extend to the thighs. Vaginal infection is usually associated with a white discharge, and plaques may be seen along an erythematous vaginal wall. Diagnosis is made by microscopic examination of the discharge for pseudohyphal elements in a 10% potassium hydroxide solution. Mild disease can be treated with topical therapy. More serious disease can be treated with fluconazole. Other causes of vaginitis include *Trichomonas* and mixed bacteria.

Diseases of the Endocrine System and Metabolic Disorders A variety of endocrine and metabolic disorders are seen in the context of HIV infection. Between 33 and 75% of patients with HIV infection receivingHAARTdevelop a syndrome often referred to as *lipodystrophy*, consisting of elevations in plasma triglycerides, total cholesterol, apolipoprotein B, and high-density lipoprotein cholesterol as well as hyperinsulinemia. Many of these patients have been noted to have a characteristic set of body habitus changes associated with fat redistribution, consisting of truncal obesity coupled with peripheral wasting (Fig. 309-32). Truncal obesity is apparent as an increase in abdominal girth related to increases in mesenteric fat, a dorsocervical fat pad ("buffalo hump") reminiscent of patients with Cushing's syndrome, and enlargement of the breasts. The peripheral wasting is particularly noticeable in the face and buttocks and by the prominence of the veins in the legs. Other related problems include insulin-requiring diabetes mellitus and avascular necrosis of the femoral head. These changes may develop at any time ranging from approximately 6 weeks to several years following the initiation of HAART. The syndrome has been reported in association with regimens containing a variety of different drugs, and while initially reported in the setting of protease inhibitor therapy, it appears similar changes can be induced by potent protease-sparing regimens. National Cholesterol Education Program (NCEP) guidelines should be followed in the management of these lipid abnormalities (Chap. 242). Due to concerns regarding drug interactions, the most commonly utilized agents in this setting are gemfibrozil and atorvostatin.

Patients with advanced HIV disease may develop hyponatremia due to the syndrome of inappropriate antidiuretic hormone (vasopressin) secretion (SIADH) as a consequence of increased free water intake and decreased free water excretion. SIADH is usually seen in conjunction with pulmonary orCNSdisease. Low serum sodium may also be due to adrenal insufficiency; concomitant high serum potassium should alert one to this possibility. Adrenal gland disease may be due to mycobacterial infections,CMVdisease, cryptococcal disease, histoplasmosis, or ketoconazole toxicity.

Thyroid function is generally normal in patients with HIV infection although approximately 2 to 3% of patients may have elevations in thyroid stimulating hormone (TSH). In advanced HIV disease, infection of the thyroid gland may occur with opportunistic pathogens, including *P. carinii*,CMV, mycobacteria, *Toxoplasma gondii*, and *Cryptococcus neoformans*. These infections are generally associated with a nontender, diffuse enlargement of the thyroid gland. Thyroid function is usually normal. Diagnosis is made by fine-needle aspirate or open biopsy.

Advanced HIV disease is associated with *hypogonadism* in approximately 50% of men. While this is generally a complication of underlying illness, testicular dysfunction may also be a side effect of ganciclovir therapy. In some surveys, up to two-thirds of patients report decreased libido and one-third complain of impotence. Androgen replacement therapy should be considered in patients with symptomatic hypogonadism. HIV infection does not seem to have a significant effect on the menstrual cycle outside the setting of advanced disease.

Rheumatologic Diseases Immunologic and rheumatologic disorders are common in patients with HIV infection and range from excessive immediate-type hypersensitivity reactions (Chap. 310) to an increase in the incidence of reactive arthritis (Chap. 315) to conditions characterized by a diffuse infiltrative lymphocytosis. These phenomena occur in an apparent paradox to the profound immunodeficiency and immunosuppression that characterizes HIV infection. In addition, following the initiation of antiretroviral therapy, one may see a variety of exaggerated immune responses to existing opportunistic infections referred to as *immune reactivation syndromes*.

Drug allergies are the most significant allergic reactions occurring in HIV-infected patients and appear to become more common as the disease progresses. They occur in 65% of patients who receive therapy withTMP/SMZfor PCP. In general, these drug reactions are characterized by erythematous, morbilliform eruption that are pruritic, tend to coalesce, and are often associated with fever. Nonetheless, ~33% of patients can be maintained on the offending therapy, and thus these reactions are not an immediate indication to stop the drug. Anaphylaxis is extremely rare in patients with HIV infection, and patients who have a cutaneous reaction during a single course of therapy can still be considered candidates for future treatment or prophylaxis with the same agent. The one exception to this is the nucleoside analogue abacavir, where fatal hypersensitivity reactions have been reported with rechallenge. A hypersensitivity reaction to abacavir is an absolute contraindication to future therapy. For other agents, including TMP/SMZ, desensitization regimens are moderatively successful. While the mechanisms underlying these allergic-type reactions remain unknown, patients with HIV infection have been noted to have elevated IgE levels that increase as the CD4+ T cell count declines. The numerous examples of patients with multiple drug reactions suggest that a common pathway is involved.

HIV infection shares many similarities with a variety of autoimmune diseases, including a substantial polyclonal B cell activation that is associated with a high incidence of antiphospholipid antibodies, such as anticardiolipin antibodies,VDRLantibodies, and lupus-like anticoagulants. In addition, HIV-infected individuals have an increased incidence of antinuclear antibodies. Despite these serologic findings, there is no evidence that HIV-infected individuals have an increase in two of the more common autoimmune diseases, i.e., systemic lupus erythematosus and rheumatoid arthritis. In fact, it has been observed that these diseases may be somewhat ameliorated by the concomitant presence of HIV infection, suggesting that an intact CD4+ T cell limb of the immune response plays an integral role in the pathogenesis of these conditions. Similarly, there are anecdotal reports of patients with common variable immunodeficiency (Chap. 308), characterized by hypogammaglobulinemia who have had a normalization of Ig levels following the development of HIV infection, suggesting a possible role for overactive CD4+ T cell immunity in certain forms of that syndrome. The one autoimmune disease that may occur with an increased frequency in patients with HIV infection is a variant of primary Sjogren's syndrome (Chap. 314). Patients with HIV infection may develop a syndrome consisting of parotid gland enlargement, dry eyes, and dry mouth that is associated with lymphocytic infiltrates of the salivary gland and lung. In contrast to Sjogren's syndrome, in which these infiltrates are composed predominantly of CD4+ T cells, in patients with HIV infection the infiltrates are composed predominantly of CD8+ T cells. In addition, while patients with Sjogren's syndrome are mainly women who have autoantibodies to Ro and La and who frequently have HLA-DR3 or -B8,MHChaplotypes, HIV-infected individuals with this syndrome are usually African-American men who do not have anti-Ro or anti-La and who most often are HLA-DR5. This syndrome appears to be less common with the increased use of effective antiretroviral therapy. The term *diffuse infiltrative lymphocytosis syndrome* (DILS) has been proposed to describe this entity and to distinguish it from Sjogren's syndrome.

Approximately one-third of HIV-infected individuals experience arthralgias; furthermore, 5 to 10% are diagnosed as having some form of reactive arthritis, such as Reiter's syndrome or psoriatic arthritis (Chaps. 315 and324). These syndromes occur with increasing frequency as the competency of the immune system declines. This association may be related to an increase in the number of infections with organisms that may trigger a reactive arthritis with progressive immunodeficiency. Reactive arthritides in HIV-infected individuals generally respond well to standard treatment; however, therapy with methotrexate has been associated with an increase in the incidence of opportunistic infections and should be used with caution and only in severe cases.

HIV-infected individuals also experience a variety of joint problems without obvious cause that are referred to generically as *HIV- or AIDS-associated arthropathy*. This syndrome is characterized by subacute oligoarticular arthritis developing over a period of 1 to 6 weeks and lasting 6 weeks to 6 months. It generally involves the large joints, predominantly the knees and ankles, and is nonerosive with only a mild inflammatory response. X-rays of the joint are nonrevealing. Nonsteroidal anti-inflammatory drugs are only marginally helpful; however, relief has been noted with the use of intraarticular glucocorticoids. A second form of arthritis also thought to be secondary to HIV infection is called *painful articular syndrome*. This condition, found in as many as 10% of AIDS patients, presents as an acute, severe, sharp pain in the affected joint. It affects primarily the knees, elbows, and shoulders; lasts 2 to 24 h; and may be severe enough to require narcotic analgesics. The cause of this arthropathy is unclear; however, it is thought to result from a direct effect of HIV on the joint. This condition is reminiscent of the fact that other lentiviruses, in particular the caprine arthritis-encephalitis virus, are capable of directly causing arthritis.

A variety of other immunologic or rheumatologic diseases have been reported in HIV-infected individuals, either de novo or in association with opportunistic infections or drugs. Using the criteria of widespread musculoskeletal pain of at least 3 months' duration and the presence of at least 11 of 18 possible tender points by digital palpation, 11% of an HIV-infected cohort containing 55%IDUswere diagnosed as having *fibromyalgia* (Chap. 325). While the incidence of frank arthritis was less in this population than in other studied populations that consisted predominantly of homosexual men, these data support the concept that there are musculoskeletal problems that occur as a direct result of HIV infection. In addition there have been reports of leukocytoclastic vasculitis in the setting of zidovudine therapy.CNSangiitis and polymyositis have also been reported in HIV-infected individuals. Septic arthritis is surprisingly rare, especially given the increased incidence of staphylococcal bacteremias seen in this population. When septic arthritis has been reported, it has usually been due to systemic fungal infections with *C. neoformans*, *Sporothrix schenckii*, or *H. capsulatum*, or systemic mycobacterial infection with *M. haemophilum*.

Following the initiation of effective antiretroviral therapy, a paradoxical worsening of preexisting, untreated opportunistic infections may be noted. These *immune reactivation syndromes* are particularly common in patients with underlying untreated mycobacterial infections. They appear to be related to a phenomenon similar to type IV hypersensitivity reactions and reflect the immediate improvements in immune function that occur as levels of HIV RNA drop and the immunosuppressive effects of HIV infection are controlled. In severe cases the use of immunosuppressive drugs such as glucocorticoids may be required to blunt the inflammatory component of these reactions while specific antimicrobial therapy takes effect.

Diseases of the Hematopoietic System Disorders of the hematopoietic system including lymphadenopathy, anemia, leukopenia, and/or thrombocytopenia are common throughout the course of HIV infection and may be the direct result of HIV, manifestations of secondary infections and neoplasms, or side effects of therapy (Table 309-12). Direct histologic examination and culture of lymph node or bone marrow tissue are often diagnostic. A significant percentage of bone marrow aspirates from patients with HIV infection have been reported to contain lymphoid aggregates, the precise significance of which is unknown.

Some patients, otherwise asymptomatic, may develop *persistent generalized lymphadenopathy* as an early clinical manifestation of HIV infection. This condition is defined as the presence of enlarged lymph nodes (>1 cm) in two or more extrainguinal sites for >3 months without an obvious cause. The lymphadenopathy is due to marked follicular hyperplasia in the node in response to HIV infection. The nodes are generally discrete and freely movable. This feature of HIV disease may be seen at any point in the spectrum of immune dysfunction and is not associated with an increased likelihood of developing AIDS. Paradoxically, a loss in lymphadenopathy or a decrease in lymph node size outside the setting of antiretroviral therapy may be a prognostic marker of disease progression. In patients with CD4+ T cell counts>200/uL, the differential diagnosis of lymphadenopathy includesKS andTB. In patients with more advanced disease, lymphadenopathy may also be due to lymphoma, atypical mycobacterial infection, toxoplasmosis, systemic fungal infection, or bacillary angiomatosis. While indicated in patients with CD4+ T cell counts <200/uL, lymph node biopsy is not indicated in patients with early-stage disease unless there are signs and symptoms of systemic illness, such as fever and weight loss, or unless the nodes begin to enlarge, become fixed, or coalesce.

Anemia is the most common hematologic abnormality in HIV-infected patients. While generally mild, anemia can be quite severe and require chronic blood transfusions. Among the specific reversible causes of anemia in the setting of HIV infection are drug toxicity, systemic fungal and mycobacterial infections, nutritional deficiencies, and parvovirus B19 infections. Zidovudine has a somewhat selective ability to block erythroid maturation, an effect that precedes effects on other marrow elements. A characteristic feature of zidovudine therapy is an elevated mean corpuscular volume (MCV). Another drug used in patients with HIV infection that has a selective effect on the erythroid series is dapsone. This drug can cause a serious hemolytic anemia in patients who are deficient in glucose-6-phosphate dehydrogenase and can create a functional anemia in others through induction of methemoglobinemia. Folate levels are usually normal in HIV-infected individuals; however, vitamin B_{12}levels may be depressed as a consequence of achlorhydria or malabsorption. True autoimmune hemolytic anemia is rare, although ~20% of patients with HIV infection may have a positive direct antiglobulin test as a consequence of polyclonal B cell activation. Infection with parvovirus B19 may also cause anemia. It is important to recognize this possibility given the fact that it responds well to treatment with intravenous immunoglobulin. Erythropoietin levels in patients with HIV infection and anemia are generally less than expected given the degree of anemia. Treatment with erythropoietin at doses of 100 ug/kg three times a week may result in an increase in hemoglobulin levels. An exception to this is a subset of patients with zidovudine-associated anemia in whom erythropoietin levels may be quite high.

During the course of HIV infection, neutropenia may be seen in approximately half of patients. In most instances it is mild; however, it can be severe and can put patients at risk of spontaneous bacterial infections. This is most frequently seen in patients with severely advanced HIV disease and in patients receiving any of a number of potentially myelosuppressive therapies. In the setting of neutropenia, diseases that are not commonly seen in HIV-infected patients, such as aspergillosis or mucormycosis, may occur. The potential role of colony-stimulating factors in the management of patients with HIV infection has undergone extensive evaluation. Both granulocyte colony stimulating factor (G-CSF) andGM-CSFincrease neutrophil counts in patients with HIV infection regardless of the cause of the neutropenia. Earlier concerns about the potential of these agents to also increase levels of HIV were not confirmed in controlled clinical trials.

Thrombocytopenia may be an early consequence of HIV infection. Approximately 3% of patients with untreated HIV infection and CD4+ T cell counts ³400/uL have platelet counts <150,000/uL. For untreated patients with CD4+ T cell counts <400/uL, this incidence increases to 10%. Thrombocytopenia is rarely a serious clinical problem in patients with HIV infection and generally responds well to antiretroviral therapy. Clinically, it resembles the thrombocytopenia seen in patients with idiopathic thrombocytopenic purpura (Chap. 116). Immune complexes containing anti-gp120 antibodies and anti-anti gp120 antibodies have been noted in the circulation and on the surface of platelets in patients with HIV infection. Patients with HIV infection have also been noted to have a platelet-specific antibody directed towards a 25-kDa component of the surface of the platelet. Other data suggest that the thrombocytopenia in patients with HIV infection may be due to a direct effect of HIV on megakaryocytes. Whatever the cause, it is very clear that the most effective medical approach to this problem has been the use of combination antiretroviral therapy. For patients with platelet counts<20,000/uL a more aggressive approach combining intravenous Ig or anti-Rh Ig for an immediate response with antiretroviral therapy for a more lasting response is appropriate. Splenectomy is a rarely needed option and is reserved for patients refractory to medical management. Because of the risk of serious infection with encapsulated organisms, all patients with HIV infection about to undergo splenectomy should be immunized with pneumococcal polysaccharide. It should be noted that, in addition to causing an increase in the platelet count, removal of the spleen will result in an increase in the peripheral blood lymphocyte count, making CD4+ T cell counts unreliable. In this setting, the clinician should rely on the CD4+ T cell percent for making diagnostic decisions with respect to the likelihood of opportunistic infections. A CD4+ T cell percent of 15 is approximately equivalent to a CD4+ T cell count of 200/uL. In patients with early HIV infection, thrombocytopenia has also been reported as a consequence of classic thrombotic thrombocytopenic purpura (Chap. 116). This clinical syndrome, consisting of fever, thrombocytopenia, hemolytic anemia, and neurologic and renal dysfunction, is a rare complication of early HIV infection. As in other settings, the appropriate management is the use of salicylates and plasma exchange. Other causes of thrombocytopenia include lymphoma, mycobacterial infections, and fungal infections.

Dermatologic Diseases Dermatologic problems occur in >90% of patients with HIV infection. From the macular, roseola-like rash seen with the acute seroconversion syndrome to extensive end-stageKS, cutaneous manifestations of HIV disease can be seen throughout the course of HIV infection. Among the more common nonneoplastic problems are seborrheic dermatitis, eosinophilic pustular folliculitus, and opportunistic infections. Extrapulmonary pneumocystosis may cause a necrotizing vasculitis. Neoplastic conditions are covered below in the section on malignant diseases.

Seborrheic dermatitis occurs in 3% of the general population and in up to 50% of patients with HIV infection. Seborrheic dermatitis increases in prevalence and severity as the CD4+ T cell count declines. In HIV-infected patients, seborrheic dermatitis may be aggravated by concomitant infection with *Pityrosporum*, a yeastlike fungus; use of topical antifungal agents has been recommended in cases refractory to standard topical treatment.

Eosinophilic pustular folliculitis is a rare dermatologic condition that is seen with increased frequency in patients with HIV infection. It presents as multiple, urticarial perifollicular papules that may coalesce into plaquelike lesions. Skin biopsy reveals an eosinophilic infiltrate of the hair follicle, which in certain cases has been associated with the presence of a mite. Patients typically have an elevated serum IgE level and may respond to treatment with topical anthelminthics. Patients with HIV infection have also been reported to develop a severe form of *Norwegian scabies* with hyperkeratotic psoriasiform lesions.

Both *psoriasis* and *ichthyosis*, although they are not reported to be increased in frequency, may be particularly severe when they occur in patients with HIV infection. Preexisting psoriasis may become guttate in appearance and more refractory to treatment in the setting of HIV infection.

Reactivation herpes zoster (*shingles*) is seen in 10 to 20% of patients with HIV infection. This reactivation syndrome of varicella-zoster virus indicates a modest decline in immune function and may be the first indication of clinical immunodeficiency. In one series, patients who developed shingles did so an average of 5 years after HIV infection. In a cohort of patients with HIV infection and localized zoster, the subsequent rate of the development of AIDS was 1% per month. In that study, AIDS was more likely to develop if the outbreak of zoster was associated with severe pain, extensive skin involvement, or involvement of cranial or cervical dermatomes. The clinical manifestations of reactivation zoster in HIV-infected patients, although indicative of immunologic compromise, are not as severe as those seen in other immunodeficient conditions. Thus, while lesions may extend over several dermatomes (see Plate IID-37) and frank cutaneous dissemination may be seen, visceral involvement has not been reported. In contrast to patients without a known underlying immunodeficiency state, patients with HIV infection tend to have recurrences of zoster with a relapse rate of approximately 20%. Acyclovir or famciclovir is the treatment of choice. Foscarnet is of value in patients with acyclovir-resistant virus.

Infection with *herpes simplex virus* in HIV-infected individuals is associated with recurrent orolabial, genital, and perianal lesions as part of recurrent reactivation syndromes (Chap. 182). As HIV disease progresses and the CD4+ T cell count declines, these infections become more frequent and severe. Lesions often appear as beefy red, are exquisitely painful, and have a tendency to occur high in the gluteal cleft (Fig. 309-31). PerirectalHSV may be associated with proctitis and anal fissures. HSV should be high in the differential diagnosis of any HIV-infected patient with a poorly healing, painful perirectal lesion. In addition to recurrent mucosal ulcers, recurrent HSV infection in the form of *herpetic whitlow* can be a problem in patients with HIV infection, presenting with painful vesicles or extensive cutaneous erosion. Acyclovir or famciclovir is the treatment of choice in these settings.

Diffuse skin eruptions due to *Molluscum contagiosum* may be seen in patients with advanced HIV infection. These flesh-colored, umbilicated lesions may be treated with local therapy. They tend to regress with effective antiretroviral therapy. Similarly, *condyloma acuminatum* lesions may be more severe and more widely distributed in patients with low CD4+ T cell counts. Atypical mycobacterial infections may present as erythematous cutaneous nodules as may fungal infections, *Bartonella*, *Acanthamoeba*, andKS.

The skin of patients with HIV infection is often a target organ for drug reactions (Chap.

59). Although most skin reactions are mild and not necessarily an indication to discontinue therapy, patients may have particularly severe cutaneous reactions, including erythroderma and *Stevens-Johnson syndrome*, as a reaction to drugs, particularly sulfa drugs, the nonnucleoside reverse transcriptase inhibitors, abacavir, and amprenavir. Similarly, patients with HIV infection are often quite photosensitive and burn easily following exposure to sunlight or as a side effect of radiation therapy (see Chap. 60).

HIV infection and its treatment may be accompanied by cosmetic changes of the skin that are not of great clinical importance but may be troubling to patients. Yellowing of the nails and straightening of the hair, particularly in African-American patients, have been reported as a consequence of HIV infection. Zidovudine therapy has been associated with elongation of the eyelashes and the development of a bluish discoloration to the nails, again more common in African-American patients. Therapy with clofazimine may cause a yellow-orange discoloration of the skin.

Neurologic Diseases Clinical disease of the nervous system accounts for a significant degree of morbidity in a high percentage of patients with HIV infection (Table 309-13). The neurologic problems that occur in HIV-infected individuals may be either primary to the pathogenic processes of HIV infection or secondary to opportunistic infections or neoplasms (see above). Among the more frequent opportunistic diseases that involve theCNS are toxoplasmosis, cryptococcosis, progressive multifocal leukoencephalopathy, and primary CNS lymphoma. Other less common problems include mycobacterial infections; syphilis; and infection withCMV,HTLV-I, or *Acanthamoeba*. Overall, secondary diseases of the CNS occur in approximately one-third of patients with AIDS. These data antedate the widespread use of combination antiretroviral therapy, and this frequency is considerably less in patients receiving effective antiretroviral drugs. Primary processes related to HIV infection of the nervous system are reminiscent of those seen with other lentiviruses, such as the Visna-Maedi virus of sheep. Neurologic problems occur throughout the course of disease and may be inflammatory, demyelinating, or degenerative in nature. While only one of these, the *AIDS dementia complex*, or *HIV encephalopathy*, is considered an AIDS-defining illness, most HIV-infected patients have some neurologic problem during the course of their disease. As noted in the section on pathogenesis, damage to the CNS may be a direct result of viral infection of the CNS macrophages or glial cells or may be secondary to the release of neurotoxins and potentially toxic cytokines such asIL-1b,TNF-a, IL-6, andTGF-b. Virtually all patients with HIV infection have some degree of nervous system involvement with the virus. This is evidenced by the fact thatCSFfindings are abnormal in approximately 90% of patients, even during the asymptomatic phase of HIV infection. CSF abnormalities include pleocytosis (50 to 65% of patients), detection of viral RNA (~75%), elevated CSF protein (35%), and evidence of intrathecal synthesis of anti-HIV antibodies (90%). It is important to point out that evidence of infection of the CNS with HIV does not imply impairment of cognitive function. The neurologic function of an HIV-infected individual should be considered normal unless clinical signs and symptoms suggest otherwise.

Aseptic meningitis may be seen in any but the very late stages of HIV infection. In the setting of acute primary infection patients may experience a syndrome of headache, photophobia, and meningismus. Rarely, an acute encephalopathy due to encephalitis may occur. Cranial nerve involvement may be seen, predominantly cranial nerve VII but occasionally V and/or VIII.CSFfindings include a lymphocytic pleocytosis, elevated protein level, and normal glucose level. This syndrome, which cannot be clinically differentiated from other viral meningitides (Chap. 373), usually resolves spontaneously within 2 to 4 weeks; however, in some patients, signs and symptoms may become chronic. Aseptic meningitis may occur any time in the course of HIV infection; however, it is rare following the development of AIDS. This fact suggests that clinical aseptic meningitis in the context of HIV infection is an immune-mediated disease.

C. neoformans is the leading infectious cause of meningitis in patients with AIDS (Chap. 204). It is the initial AIDS-defining illness in approximately 2% of patients and generally occurs in patients with CD4+ T cell counts <100/uL. Cryptococcal meningitis is particularly common in patients with AIDS in Africa, occurring in ~20% of patients. Most patients resent with a picture of subacute meningoencephalitis with fever, nausea, vomiting, altered mental status, headache, and meningeal signs. The incidence of seizures and focal neurologic deficits is low. TheCSFprofile may be normal or may show only modest elevations inWBC or protein levels. In addition to meningitis, patients may develop cryptococcomas. Approximately one-third of patients also have pulmonary disease. Uncommon manifestations of cryptococcal infection include skin lesions that resemble *molluscum contagiosum*, lymphadenopathy, palatal and glossal ulcers, arthritis, gastroenteritis, myocarditis, and prostatitis. The prostate gland may serve as a reservoir for smoldering cryptococcal infection. The diagnosis of cryptococcal meningitis is made by identification of organisms in spinal fluid with India ink examination or by the detection of cryptococcal antigen. A biopsy may be needed to make a diagnosis ofCNScryptococcoma. Treatment is with intravenous amphotericin B, at a dose of 0.7 mg/kg daily, with flucytosine, 25 mg/kg qid for 2 weeks, followed by fluconazole, 400 mg/d orally for 8 weeks, and then fluconazole, 200 mg/d for life. Other fungi that may cause meningitis in patients with HIV infection are *C. immitis* and *H. capsulatum*. Meningoencephalitis has also been reported due to *Acanthamoeba* or *Naegleria*.

HIV encephalopathy, also called HIV-associated dementia or AIDS dementia complex, consists of a constellation of signs and symptoms ofCNSdisease that generally occurs late in the course of HIV infection and progresses slowly over months. A major feature of this entity is the development of dementia, defined as a decline in cognitive ability from a previous level. It may present as impaired ability to concentrate, increased forgetfulness, difficulty reading, or increased difficulty performing complex tasks. Initially these symptoms may be indistinguishable from findings of situational depression or fatigue. In contrast to "cortical" dementia (such as Alzheimer's disease), aphasia, apraxia, and agnosia are uncommon, leading some investigators to classify HIV encephalopathy as a "subcortical dementia" (see below). In addition to dementia, patients with HIV encephalopathy may also have motor and behavioral abnormalities. Among the motor problems are unsteady gait, poor balance, tremor, and difficulty with rapid alternating movements. Increased tone and deep tendon reflexes may be found in patients with spinal cord involvement. Late stages may be complicated by bowel and/or bladder incontinence. Behavioral problems include apathy and lack of initiative, with progression to a vegetative state in some instances. Some patients develop a state of agitation or mild mania. These changes usually occur without significant changes in level of alertness. This is in contrast to the finding of somnolence in patients with dementia due to toxic/metabolic encephalopathies.

HIV encephalopathy is the initial AIDS-defining illness in approximately 3% of patients with HIV infection and thus only rarely precedes clinical evidence of immunodeficiency. Clinically significant encephalopathy eventually develops in approximately one-fourth of patients with AIDS. As immunologic function declines, the risk and severity of HIV encephalopathy increases. Autopsy series suggest that 80 to 90% of patients with HIV infection have histologic evidence of CNSinvolvement. Several classification schemes have been developed for grading HIV encephalopathy; a commonly used clinical staging system is outlined inTable 309-14.

The precise cause of HIV encephalopathy remains unclear, although the condition is thought to be a result of direct effects of HIV on theCNS. HIV has been found in the brains of patients with HIV encephalopathy by Southern blot, in situ hybridization,PCR, and electron microscopy. Multinucleated giant cells, macrophages, and microglial cells appear to be the main cell types harboring virus in the CNS. Histologically, the major changes are seen in the subcortical areas of the brain and include pallor and gliosis, multinucleated giant cell encephalitis, and vacuolar myelopathy. Less commonly, diffuse or focal spongiform changes occur in the white matter.

There are no specific criteria for a diagnosis of HIV encephalopathy, and this syndrome must be differentiated from a number of other diseases that affect theCNS of HIV-infected patients (Table 309-13). The diagnosis of dementia depends upon demonstrating a decline in cognitive function. This can be accomplished objectively with the use of a Mini-Mental Status Examination (MMSE) (Table 309-15) in patients for whom prior scores are available. For this reason, it is advisable for all patients with a diagnosis of HIV infection to have a baseline MMSE. However, changes in MMSE scores may be absent in patients with mild HIV encephalopathy. Imaging studies of the CNS, by eitherMRI orCT, often demonstrate evidence of cerebral atrophy (Fig. 309-33). MRI may also reveal small areas of increased density on T2-weighted images. Lumbar puncture is an important element of the evaluation of patients with HIV infection and neurologic abnormalities. It is generally most helpful in ruling out or making a diagnosis of opportunistic infections. In HIV encephalopathy, patients may have the nonspecific findings of an increase inCSFcells and protein level. While HIV RNA can often be detected in the spinal fluid and HIV can be cultured from the CSF, this finding is not specific for HIV encephalopathy. There appears to be no correlation between the presence of HIV in the CSF and the presence of HIV encephalopathy. Elevated levels ofb$_2$-microglobulin, neopterin, and quinolinic acid (a metabolite of tryptophan reported to cause CNS injury) have been noted in the CSF of patients with HIV encephalopathy. These findings suggest that these factors as well as inflammatory cytokines may be involved in the pathogenesis of this syndrome.

Combination antiretroviral therapy is of benefit in patients with HIV encephalopathy. Improvement in neuropsychiatric test scores has been noted for both adult and pediatric patients treated with antiretrovirals. The rapid improvement in cognitive function noted with the initiation of antiretroviral therapy suggests that at least some component of this problem is quickly reversible, again supporting at least a partial role of soluble mediators in the pathogenesis. It should also be noted that these patients have an increased sensitivity to the side effects of neuroleptic drugs. The use of these drugs for symptomatic treatment is associated with an increased risk of extrapyramidal side effects; therefore, patients with HIV encephalopathy who receive these agents must be monitored carefully.

Seizures may be a consequence of opportunistic infections, neoplasms, or HIV encephalopathy (Table 309-16). The seizure threshold is often lower than normal in these patients owing to the frequent presence of electrolyte abnormalities. Seizures are seen in 15 to 40% of patients with cerebral toxoplasmosis, 15 to 35% of patients with primaryCNSlymphoma, 8% of patients with cryptococcal meningitis, and 7 to 50% of patients with HIV encephalopathy. Seizures may also be seen in patients with CNS tuberculosis, aseptic meningitis, and progressive multifocal leukoencephalopathy. Seizures may be the presenting clinical symptom of HIV disease. In one study of 100 patients with HIV infection presenting with a first seizure, cerebral mass lesions were the most common cause, responsible for 32 of the 100 new-onset seizures. Of these 32 cases, 28 were due to toxoplasmosis and 4 to lymphoma. HIV encephalopathy accounted for an additional 24 new-onset seizures. Cryptococcal meningitis was the third most common diagnosis, responsible for 13 of the 100 seizures. In 23 cases, no cause could be found, and it is possible that these cases represent a subcategory of HIV encephalopathy. Of these 23 cases, 16 (70%) had two or more seizures, suggesting that anticonvulsant therapy is indicated in all patients with HIV infection and seizures unless a rapidly correctable cause is found. While phenytoin remains the initial treatment of choice, hypersensitivity reactions to this drug have been reported in >10% of patients with AIDS, and therefore the use of phenobarbital or valproic acid must be considered as alternatives.

Patients with HIV infection may present with *focal neurologic deficits* from a variety of causes. The most common cause are toxoplasmosis, progressive multifocal leukoencephalopathy, andCNSlymphoma. Other causes include cryptococcal infections (discussed above; alsoChap. 204), stroke, and reactivation Chagas' disease.

Toxoplasmosis has been one of the most common causes of secondaryCNSinfections in patients with AIDS, but its incidence is decreasing in the era ofHAART. It is most common in patients from the Caribbean and from France. Toxoplasmosis is generally a late complication of HIV infection and usually occurs in patients with CD4+ T cell counts <200/uL. Cerebral toxoplasmosis is thought to represent a reactivation syndrome. It is 10 times more common in patients with antibodies to the organism than in patients who are seronegative. Patients diagnosed with HIV infection should be screened for IgG antibodies to *T. gondii* during the time of their initial workup. Those who are seronegative should be counseled about ways to minimize the risk of primary infection including avoiding the consumption of undercooked meat and careful hand washing after contact with soil or changing the cat litter box. The most common clinical presentation in patients with HIV infection is fever, headache, and focal neurologic deficits. Patients may present with seizure, hemiparesis, or aphasia as a manifestation of these focal deficits or with a picture more influenced by the accompanying cerebral edema and characterized by confusion, dementia, and lethargy, which can progress to coma. The diagnosis is usually suspected on the basis ofMRIfindings of multiple lesions in multiple locations, although in come cases only a single lesion is seen. Pathologically, these lesions generally exhibit inflammation and central necrosis and, as a result, demonstrate ring enhancement on contrast MRI (Fig. 309-34) or, if MRI is unavailable or contraindicated, on double-dose contrastCT. There is usually evidence of surrounding edema. In addition to toxoplasmosis, the differential diagnosis of single or multiple enhancing mass lesions in the HIV-infected patient includes primary CNS lymphoma (see below) and, less commonly,TB or fungal or bacterial abscesses. The definitive diagnostic procedure is brain biopsy. However, given the morbidity than can accompany this procedure, it is usually reserved for the patient who has failed 2 to 4 weeks of empirical therapy. If the patient is seronegative for *T. gondii*, the likelihood that a mass lesion is due to toxoplasmosis is <10%. In that setting, one may choose to be more aggressive and perform a brain biopsy sooner. Standard treatment is sulfadiazine and pyrimethamine with leucovorin as needed for a minimum of 4 to 6 weeks. Alternative therapeutic regimens include clindamycin in combination with pyrimethamine; atovaquone plus pyrimethamine; and azithromycin plus pyrimethamine plus rifabutin. Relapses are common, and it is recommended that patients with a history of prior toxoplasmic encephalitis receive maintenance therapy with sulfadiazine, pyrimethamine, and leucovorin. Patients with CD4+ T cell counts <100/uL and IgG antibody to *Toxoplasma* should receive primary prophylaxis for toxoplasmosis. Fortunately, the same daily regimen of a single double-strength tablet ofTMP/SMZused for *P. carinii* prophylaxis provides adequate primary protection against toxoplasmosis. It is likely that future recommendations will allow for discontinuation of prophylaxis for toxoplasmosis in the setting of effective antiretroviral therapy and increases in CD4+ T cell counts to>100/uL for 3 to 6 months.

JC virus, a human papilloma virus that is the etiologic agent of *progressive multifocal leukoencephalopathy* (PML), is an important opportunistic pathogen in patients with AIDS (Chap. 373). While approximately 70% of the general adult population have antibodies to JC virus, indicative of prior infection,<10% of healthy adults show any evidence of ongoing viral replication. PML is the only known clinical manifestation of JC virus infection. It is a late manifestation of AIDS and is seen in ~4% of patients with AIDS. The lesions of PML begin as small foci of demyelination in subcortical white matter that eventually coalesce. The cerebral hemispheres, cerebellum, and brainstem may all be involved. Patients typically have a protracted course with multifocal neurologic deficits, with or without changes in mental status. Ataxia, hemiparesis, visual field defects, aphasia, and sensory defects may occur.MRItypically reveals multiple, nonenhancing white matter lesions that may coalesce and have a predeliction for the occipital and parietal lobes. The lesions show signal hyperintensity on T2-weighted images and diminished signal on T1-weighted images. Prior to the availability of potent antiretroviral combination therapy, the majority of patients with PML died within 3 to 6 months of the onset of symptoms. There is no specific treatment for PML; however, regressions of more than 2.5 years in duration have been reported in patients with PML treated withHAARTfor their HIV disease. Factors influencing a favorable prognosis include a CD4+ T cell count >100/uL at baseline and the ability to maintain an HIV viral load of <500 copies per milliliter. Baseline viral load does not have independent predictive value of survival. Of note, PML is one of the few opportunistic infections that continue to occur with some frequency despite the widespread use of HAART.

Reactivation American trypanosomiasis may present as acute meningoencephalitis with focal neurologic signs, fever, headache, vomiting, and seizures. In South America, reactivation of *Chagas' disease* is considered to be an AIDS-defining condition and may be the initial AIDS-defining condition. Lesions appear radiographically as single or multiple hypodense areas, typically with ring enhancement and edema. They are found predominantly in the subcortical areas, a feature that differentiates them from the deeper lesions of toxoplasmosis. *Trypanosoma cruzi* amastigotes, or trypanosomes, can be identified from biopsy specimens orCSF. Other CSF findings include elevated protein and a mild (<100 cells/uL) lymphocytic pleocytosis. Organisms can also be identified by direct examination of the blood. Treatment consists of benzimidazole (2.5 mg/kg bid) or nifurtimox (1 mg/kg tid) for at least 60 days, followed by maintenance therapy for life with either drug at a dose of 5 mg/kg three times a week.

Stroke may occur in patients with HIV infection. In contrast to the other causes of focal neurologic deficits in patients with HIV infection, the symptoms of a stroke are sudden in onset. Among the secondary infectious diseases in patients with HIV infection that may be associated with stroke are vasculitis due to cerebral varicella zoster or neurosyphilis and septic embolism in association with fungal infection. Other elements of the differential diagnosis of stroke in the patient with HIV infection include atherosclerotic cerebral vascular disease, thrombotic thrombocytopenic purpura, and cocaine or amphetamine use.

PrimaryCNSlymphoma is discussed below in the section on neoplastic diseases.

Spinal cord disease, or myelopathy, is present in approximately 20% of patients with AIDS, often as part of HIV encephalopathy. In fact, 90% of the patients with HIV-associated myelopathy have some evidence of dementia, suggesting that similar pathologic processes may be responsible for both conditions. Three main types of spinal cord disease are seen in patients with AIDS. The first of these is a vacuolar myelopathy, as discussed above under HIV encephalopathy. This condition is pathologically similar to subacute combined degeneration of the cord such as occurs with pernicious anemia. Although vitamin B_{12}deficiency can be seen in patients with AIDS, it does not appear to be responsible for the myelopathy seen in the majority of patients. Vacuolar myelopathy is characterized by a subacute onset and often presents with gait disturbances, predominantly ataxia and spasticity; it may progress to include bladder and bowel dysfunction. Physical findings include evidence of increased deep tendon reflexes and extensor plantar responses. The second form of spinal cord disease involves the dorsal columns and presents as a pure sensory ataxia. The third form is also sensory in nature and presents with paresthesias and dysesthesias of the lower extremities. In contrast to the cognitive problems seen in patients with HIV encephalopathy, these spinal cord syndromes do not respond well to antiretroviral drugs, and therapy is mainly supportive.

One important disease of the spinal cord that also involves the peripheral nerves is a *myelopathy* and *polyradiculopathy* seen in association withCMVinfection. This entity is generally seen late in the course of HIV infection and is fulminant in onset, with lower extremity and sacral paresthesias, difficulty in walking, areflexia, ascending sensory loss, and urinary retention. The clinical course is rapidly progressive over a period of weeks.CSFexamination reveals a predominantly neutrophilic pleocytosis, and CMV DNA can be detected by CSFPCR. Therapy with ganciclovir or foscarnet can lead to rapid improvement, and prompt initiation of foscarnet or ganciclovir therapy is important in minimizing the degree of permanent neurologic damage. Combination therapy with both drugs should be considered in patients who have been previously treated for CMV disease.**Other diseases involving the spinal cord in patients with HIV infection include*

HTLV-I-associated myelopathy (HAM) (Chap. 191), neurosyphilis (Chap. 172), infection with herpes simplex (Chap. 182) or varicella-zoster (Chap. 183), TB (Chap. 169), and lymphoma (Chap. 112).

Peripheral neuropathies are common in patients with HIV infection. They occur at all stages of illness and take a variety of forms. Early in the course of HIV infection, an acute inflammatory demyelinating polyneuropathy resembling Guillain-Barre syndrome may occur (Chap. 378). In other patients, a progressive or relapsing-remitting inflammatory neuropathy resembling chronic inflammatory demyelinating polyneuropathy (CIDP) has been noted. Patients commonly present with progressive weakness, areflexia, and minimal sensory changes.CSFexamination often reveals a mononuclear pleocytosis, and peripheral nerve biopsy demonstrates a perivascular infiltrate suggesting an autoimmune etiology. Plasma exchange or intravenous immunoglobulin has been tried with variable success. Because of the immunosuppressive effects of glucocorticoids, they should be reserved for severe cases of CIDP refractory to other measures. Another autoimmune peripheral neuropathy seen in patients with AIDS is mononeuritis multiplex (Chaps. 378 and317) due to a necrotizing arteritis of peripheral nerves. The most common peripheral neuropathy in patients with HIV infection is a *distal sensory polyneuropathy* that may be a direct consequence of HIV infection or a side effect of dideoxynucleoside therapy. Two-thirds of patients with AIDS may be shown by electrophysiologic studies to have some evidence of peripheral nerve disease. Presenting symptoms are usually painful burning sensations in the feet and lower extremities. Findings on examination include a stocking-type sensory loss to pinprick, temperature, and touch sensation and a loss of ankle reflexes. Motor changes are mild and are usually limited to weakness of the intrinsic foot muscles. Response of this condition to antiretrovirals has been variable, perhaps because antiretrovirals are responsible for the problem in some instances. When due to dideoxynucleoside therapy, patients with lower extremity peripheral neuropathy may complain of a sensation that they are walking on ice. Other entities in the differential diagnosis of peripheral neuropathy include diabetes mellitus, vitamin B_{12}deficiency, and side effects from metronidazole or dapsone. For distal symmetric polyneuropathy that fails to resolve following the discontinuation of dideoxynucleosides, therapy is symptomatic; gabapentin, carbamazepine, tricyclics, or analgesics may be effective for dysesthesias. Some patients may respond to combination antiretroviral therapy, and preliminary data suggest that nerve growth factor may benefit some cases.

Myopathy may complicate the course of HIV infection; causes include HIV infection itself, zidovudine, and the generalized wasting syndrome. HIV-associated myopathy may range in severity from an asymptomatic elevation in creatine kinase levels to a subacute syndrome characterized by proximal muscle weakness and myalgias. Quite pronounced elevations in creatine kinase may occur in asymptomatic patients, particularly after exercise. The clinical significance of this as an isolated laboratory finding is unclear. A variety of both inflammatory and noninflammatory pathologic processes have been noted in patients with more severe myopathy, including myofiber necrosis with inflammatory cells, nemaline rod bodies, cytoplasmic bodies, and mitochondrial abnormalities. Profound muscle wasting, often with muscle pain, may be seen after prolonged zidovudine therapy. This toxic side effect of the drug is dose-dependent and is related to its ability to interfere with the function of mitochondrial polymerases. It is reversible following discontinuation of the drug. Red ragged fibers are a histologic hallmark of zidovudine-induced myopathy.

Ophthalmologic Disease Ophthalmologic problems occur in approximately half of patients with advanced HIV infection. The most common abnormal findings on funduscopic examination are cotton-wool spots. These are hard white spots that appear on the surface of the retina and often have an irregular edge. They represent areas of retinal ischemia secondary to microvascular disease. At times they are associated with small areas of hemorrhage and thus can be difficult to distinguish fromCMVretinitis. In contrast to CMV retinitis, however, these lesions are not associated with visual loss and tend to remain stable or improve over time.

One of the most devastating consequences of HIV infection isCMVretinitis. Patients at high risk of CMV retinitis (CD4+ T cell count<100/uL) should undergo an ophthalmologic examination every 3 to 6 months. The majority of cases of CMV retinitis occur in patients with a CD4+ T cell count <50/uL. Prior to the availability ofHAART, this CMV reactivation syndrome was seen in 25 to 30% of patients with AIDS. CMV retinitis usualy presents as a painless, progressive loss of vision. Patients may also complain of blurred vision, "floaters," and scintillations. The disease is usually bilateral, affecting one eye more than the other. The diagnosis is made on clinical grounds by an experienced ophthalmologist. The characteristic retinal appearance is that of perivascular hemorrhage and exudate (seePlate III-1). In situations where the diagnosis is in doubt due to an atypical presentation or an unexpected lack of response to therapy, vitreous or aqueous humor sampling with molecular diagnostic techniques may be of value. CMV infection of the retina results in a necrotic inflammatory process, and the visual loss that develops is irreversible. As a consequence of retinal atrophy in areas or prior inflammation, CMV retinitis may be complicated by rhegmatogenous retinal detachment. Therapy for CMV retinitis consists of intravenous ganciclovir or foscarnet, with cidofovir as an alternative. Combination therapy with ganciclovir and foscarnet has been shown to be slightly more effective than either ganciclovir or foscarnet alone in the patient with relapsed CMV retinitis. A 3-week induction course is followed by maintenance therapy with one of these drugs systemically. While the majority of patients will require intravenous maintenance therapy, a ganciclovir prodrug with better oral bioavailability has shown promise in clinical trials. If CMV disease is limited to the eye, a ganciclovir-releasing intraocular implant, periodic injections of the antisense nucleic acid preparation formivirsen, or intravitreal injections of ganciclovir or foscarnet may be considered; some choose to combine intraocular implants with oral ganciclovir. Intravitreal injections of cidofovir are generally avoided due to the increased risk of uveitis and hypotony. Maintenance therapy is continued until the CD4+ T cell count remains >100 to 150/uL for>6 months. The majority of patients with HIV infection and CMV disease develop some degree of uveitis with the initiation of antiretroviral therapy. The etiology of this is unknown; however, it has been suggested that this may be due to the generation of an enhanced immune response to CMV. In some instances this has required the use of topical glucocorticoids.

BothHSV and varicella zoster virus can cause a rapidly progressing, bilateral necrotizing retinitis referred to as the *acute retinal necrosis syndrome*. This syndrome, in contrast toCMVretinitis, is associated with pain, keratitis, and iritis. It is often associated with orolabial HSV or trigeminal zoster. Ophthalmologic examination reveals widespread pale gray peripheral lesions. This condition is often complicated by retinal detachment. It is important to recognize and treat this condition with intravenous acyclovir as quickly as possible to minimize the loss of vision.

Several other secondary infections may cause ocular problems in HIV-infected patients. *P. carinii* can cause a lesion of the choroid that may be detected as an incidental finding on ophthalmologic examination. These lesions are typically bilateral, are from half to twice the disc diameter in size, and appear as slightly elevated yellow-white plaques. They are usually asymptomatic and may be confused with cotton-wool spots. Chorioretinitis due to toxoplasmosis can be seen alone or, more commonly, in association withCNStoxoplasmosis.

Additional Disseminated Infections and Wasting Syndrome Infections with species of the small, gram-negative rickettsia-like organism *Bartonella* (Chap. 163) are seen with increased frequency in patients with HIV infection. While not considered an AIDS-defining illness by theCDC, many experts view infection with *Bartonella* as indicative of a severe defect in cell-mediated immunity. It is usually seen in patients with CD4+ T cell counts <100/uL. Among the clinical manifestations of *Bartonella* infection are bacillary angiomatosis, cat-scratch disease, and trench fever. *Bacillary angiomatosis* is usually due to infection with *B. henselae*. It is characterized by a vascular proliferation that leads to a variety of skin lesions that have been confused with the skin lesions ofKS. In contrast to the lesions of KS, the lesions of bacillary angiomatosis generally blanch, are painful, and typically occur in the setting of systemic symptoms. Infection can extend to the lymph nodes, liver (peliosis hepatis), spleen, bone, heart,CNS, respiratory tract, and gastrointestinal tract. *Cat-scratch disease* generally begins with a papule at the site of inoculation. This is followed several weeks later by the development of regional adenopathy and malaise. Infection with *B. quintana* is transmitted by lice and has been associated with case reports of trench fever, endocarditis, adenopathy, and bacillary angiomatosis. The organism is quite difficult to culture, and diagnosis often relies upon identifying the organism in biopsy specimens using the Warthin-Starry or similar stains. Treatment is with either erythromycin or doxycyline for at least 3 months.

Histoplasmosis is an opportunistic infection that is seen most frequently in patients in the Mississippi and Ohio River valleys, Puerto Rico, the Dominican Republic, and South America. These are all areas in which infection with *H. capsulatum* is endemic (Chap. 201). Because of this limited geographic distribution, the percentage of AIDS cases in the United States with histoplasmosis is only approximately 0.5. Histoplasmosis is generally a late manifestation of HIV infection; however, it may be the initial AIDS-defining condition. In one study, the median CD4+ T cell count for patients with histoplasmosis and AIDS was 33/uL. While disease due to *H. capsulatum* may present as a primary infection of the lung, disseminated disease, presumably due to reactivation, is the most common presentation in HIV-infected patients. Patients usually present with a 4- to 8-week history of fever and weight loss. Hepatosplenomegaly and lymphadenopathy are each seen in about 25% of patients.CNSdisease, either meningitis or a mass lesion, is seen in 15% of patients. Bone marrow involvement is common, with thrombocytopenia, neutropenia, and anemia occurring in 33% of patients. Approximately 7% of patients have mucocutaneous lesions consisting of a maculopapular rash and skin or oral ulcers. Respiratory symptoms are usually mild, with chest x-ray showing a diffuse infiltrate or diffuse small nodules in approximately half of cases. Diagnosis is made by culturing the organisms from blood, bone marrow, or tissue. Treatment is typically with amphotericin B, 0.7 to 1.0 mg/kg daily to a total dose of 1 g followed by maintenance therapy with itraconazole. In the setting of mild infection, it may be appropriate to treat with itraconazole alone.

Following the spread of HIV infection to southeast Asia, disseminated infection with *Penicillium marneffei* was recognized as a complication of HIV infection and is considered an AIDS-defining condition in those parts of the world where it occurs. *P. marneffei* is the third most common AIDS-defining illness in Thailand, followingTB and cryptococcosis. It is more frequently diagnosed in the rainy than the dry season. Clinical features include fever, generalized lymphadenopathy, hepatosplenomegaly, anemia, thrombocytopenia, and papular skin lesions with central umbilication. Treatment is with amphotericin B followed by itraconazole.

Visceral leishmaniasis (Chap. 215) is recognized with increasing frequency in patients with HIV infection who live in or travel to areas endemic for this protozoal infection transmitted by sandflies. The clinical presentation is one of hepatosplenomegaly, fever, and hematologic abnormalities. Lymphadenopathy and other constitutional symptoms may be present. Organisms can be isolated from cultures of bone marrow aspirates. Histologic stains may be negative, and antibody titers are of little help. Patients with HIV infection usually respond well initially to standard therapy with pentavalent antimony compounds. Eradication of the organism is difficult, however, and relapses are common.

Generalized wasting is an AIDS-defining condition; it is defined as involuntary weight loss of>10% associated with intermittent or constant fever and chronic diarrhea or fatigue lasting >30 days in the absence of a defined cause other than HIV infection. It is the initial AIDS-defining condition in approximately 10% of patients with AIDS in the United States. A constant feature of this syndrome is severe muscle wasting with scattered myofiber degeneration and occasional evidence of myositis. Glucocorticoids may be of some benefit; however, this approach must be carefully weighed against the risk of compounding the immunodeficiency of HIV infection. Androgenic steroids, growth hormone, and total parenteral nutrition have been used as therapeutic interventions with variable success.

Neoplastic Diseases The neoplastic diseases clearly seen with an increased frequency in patients with HIV infection areKS and non-Hodgkin's lymphoma. In addition, there also appears to be an increased incidence of Hodgkin's disease; multiple myeloma; leukemia; melanoma; and cervical, brain, testicular, oral, and anal cancers. Recent years have witnessed a marked reduction in the incidence of KS (Fig. 309-28), felt to be primarily due to the use of potent antiretroviral therapy. Rates of non-Hodgkin's lymphoma have declined as well; however, this decline has not been as dramatic as the decline in rates of KS.

Kaposi's sarcoma is a multicentric neoplasm consisting of multiple vascular nodules appearing in the skin, mucous membranes, and viscera. The course ranges from indolent, with only minor skin or lymph node involvement, to fulminant, with extensive cutaneous and visceral involvement. In the initial period of the AIDS epidemic,KS was a prominent clinical feature of the first cases of AIDS, occurring in 79% of the patients diagnosed in 1981. By 1989 it was seen in only 25% of cases, by 1992 the number had decreased to 9%, and by 1997 the number was<1%.HHV-8 orKSHV has been strongly implicated as a viral cofactor in the pathogenesis of KS (see above).

Clinically,KS has varied presentations and may be seen at any stage of HIV infection, even in the presence of a normal CD4+ T cell count. The initial lesion may be a small, raised reddish-purple nodule on the skin, a discoloration on the oral mucosa, or a swollen lymph node (see Plate IIB-20). Lesions often appear in sun-exposed areas, particularly the tip of the nose, and have a propensity to occur in areas of trauma (Koebner phenomenon). Because of the vascular nature of the tumors and the presence of extravasated red blood cells in the lesions, their color ranges from reddish to purple to brown and often take the appearance of a bruise, with yellowish discoloration and tattooing. Lesions range in size from a few millimeters to several centimeters in diameter and may be either discrete or confluent. KS lesions most commonly appear as raised macules; however, they also can be papular, particularly in patients with higher CD4+ T cell counts. Confluent lesions may give rise to surrounding lymphedema and may be disfiguring when they involve the face and disabling when they involve the lower extremities or the surfaces of joints. Apart from skin, lymph nodes, gastrointestinal tract, and lung are the organ systems most commonly affected by KS. Lesions have been reported in virtually every organ, including the heart and theCNS. In contrast to most malignancies, in which lymph node involvement implies metastatic spread and a poor prognosis, lymph node involvement may be seen very early in Kaposi's sarcoma and is of no special clinical significance. In fact, some patients may present with disease limited to the lymph nodes. These are generally patients with relatively intact immune function and thus the patients with the best prognosis. Pulmonary involvement with KS generally presents with shortness of breath. Some 80% of patients with pulmonary KS also have cutaneous lesions. The chest x-ray characteristically shows bilateral lower lobe infiltrates that obscure the margins of the mediastinum and diaphragm (Fig. 309-35). Pleural effusions are seen in 70% of cases of pulmonary KS, a fact that is often helpful in the differential diagnosis. Gastrointestinal involvement is seen in 50% of patients and usually takes one of two forms. The first is mucosal involvement, which may lead to bleeding that can be severe. These patients sometimes also develop symptoms of gastrointestinal obstruction if lesions become large. The second gastrointestinal manifestation is due to biliary tract involvement. KS lesions may infiltrate the gallbladder and biliary tree, leading to a clinical picture of obstructive jaundice similar to that seen with sclerosing cholangitis. Several staging systems have been proposed for KS. One in common use was developed by the National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group; it distinguishes patients on the basis of tumor extent, immunologic function, and presence or absence of systemic disease (Table 309-17).

A diagnosis ofKS is based upon biopsy of a suspicious lesion. Histologically one sees a proliferation of spindle cells and endothelial cells, extravasation of red blood cells, hemosiderin-laden macrophages, and, in early cases, an inflammatory cell infiltrate. Included in the differential diagnosis are lymphoma (particularly for oral lesions), bacillary angiomatosis, and cutaneous mycobacterial infections.

Management ofKS(Table 309-18) should be carried out in consultation with an expert since definitive treatment guidelines do not exist. In the majority of cases effective antiretroviral therapy will go a long way in achieving control. Indeed, spontaneous regressions have been reported in the setting ofHAART. For patients in whom tumor persists or in whom control of HIV replication is not possible, a variety of options exist. In some cases, lesions remain quite indolent, and many of these patients can be managed with no specific treatment. Fewer than 10% of AIDS patients with KS die as a consequence of their malignancy, and death from secondary infections is considerably more common. Thus, whenever possible one should avoid treatment regimens that may further suppress the immune system and increase susceptibility to opportunistic infections. Treatment is indicated under two main circumstances. The first is when a single lesion or a limited number of lesions are causing significant discomfort or cosmetic problems, such as with prominent facial lesions, lesions overlying a joint, or lesions in the oropharynx that interfere with swallowing or breathing. Under these circumstances, treatment with localized radiation, intralesional vinblastine, or cryotherapy may be indicated. It should be noted that patients with HIV infection are particularly sensitive to the side effects of radiation therapy. This is especially true with respect to the development of radiation-induced mucositis; doses of radiation directed at mucosal surfaces, particularly in the head and neck region, should be adjusted accordingly. The use of systemic therapy, eitherIFN-aor chemotherapy, should be considered in patients with a large number of lesions or in patients with visceral involvement. The single most important determinant of response appears to be the CD4+ T cell count. This relationship between response rate and baseline CD4+ T cell count is particularly true for IFN-a. The response rate for patients with CD4+ T cell counts >600/uL is approximately 80%, while the response rate for patients with counts <150/uL is<10%. In contrast to the other systemic therapies, IFN-a provides an added advantage of having antiretroviral activity; thus, it may be the appropriate first choice for single-agent systemic therapy for early patients with disseminated disease. A variety of chemotherapeutic agents have also been shown to have activity against KS. Three of them, liposomal daunorubicin, liposomal doxorubicin, and paclitaxel have been approved by theFDA for this indication. Liposomal daunorubicin is approved as first-line therapy for patients with advanced KS. It has fewer side effects than conventional chemotherapy. In contrast, liposomal doxorubicin and paclitaxel are only approved for KS patients who have failed standard chemotherapy. Response rates vary from 23 to 88%, appear to be comparable to what had been achieved earlier with combination chemotherapy regimens, and are greatly influenced by CD4+ T cell count.

Lymphomas occur with an increased frequency in patients with congenital or acquired T cell immunodeficiencies (Chap. 308). AIDS is no exception; at least 6% of all patients with AIDS develop lymphoma at some time during the course of their illness. This is a 120-fold increase in incidence compared to the general population. In contrast to the situation withKS and most opportunistic infections, the incidence of AIDS-associated lymphomas has not experienced as dramatic a decrease as a consequence of the widespread use of effective antiretroviral therapy. Lymphoma occurs in all risk groups, with the highest incidence in patients with hemophilia and the lowest incidence in patients from the Caribbean or Africa with heterosexually acquired infection. Lymphoma is a late manifestation of HIV infection, generally occurring in patients with CD4+ T cell counts of <200/uL. As HIV disease progresses, the risk of lymphoma increases. In contrast to KS, which occurs at a relatively constant rate throughout the course of HIV disease, the attack rate for lymphoma increases exponentially with increasing duration of HIV infection and decreasing level of immunologic function. At 3 years following a diagnosis of HIV infection, the risk of lymphoma is 0.8% per year; by 8 years after infection, it is 2.6% per year. As people with HIV infection live longer as a consequence of improved antiretroviral therapy and better treatment and prophylaxis of opportunistic infections, it is anticipated that the incidence of lymphomas will increase.

Three main categories of lymphoma are seen in patients with HIV infection: grade III or IV immunoblastic lymphoma, Burkitt's lymphoma, and primaryCNSlymphoma. Approximately 90% of these lymphomas are B cell in phenotype, and half containEBV DNA. These tumors may be either monoclonal or oligoclonal in nature and are probably in some way related to the pronounced polyclonal B cell activation seen in patients with AIDS.

Immunoblastic lymphomas account for ~60% of the cases of lymphoma in patients with AIDS. These are generally high grade and would have been classified as diffuse histiocytic lymphomas in earlier classification schemes. This tumor is more common in older patients, increasing in incidence from 0% in HIV-infected individuals <1 year old to>3% in those >50. One variant of immunoblastic lymphoma is body cavity lymphoma. This malignancy presents with lymphomatous pleural, pericardial, and/or peritoneal effusions in the absence of discrete nodal or extranodal masses. The tumor cells do not express surface markers for B cells or T cells.HHV-8 DNA sequences have been found in the genome of the malignant cells (see above).

Small non-cleaved cell lymphoma (*Burkitt's lymphoma*) accounts for ~20% of the cases of lymphoma in patients with AIDS. It is most frequent in patients 10 to 19 years old and usually demonstrates characteristic c-*myc* translocations from chromosome 8 to chromosomes 14 or 22. Burkitt's lymphoma is not commonly seen in the setting of immunodeficiency other than HIV-associated immunodeficiency, and the incidence of this particular tumor is over 1000-fold higher in the setting of HIV infection than in the general population. In contrast to African Burkitt's lymphoma, where 97% of the cases contain EBV genome, only 50% of HIV-associated Burkitt's lymphomas are EBV-positive.

Primary CNS lymphoma accounts for approximately 20% of the cases of lymphoma in patients with HIV infection. In contrast to HIV-associated Burkitt's lymphoma, primaryCNSlymphomas are usually positive forEBV. In one study, the incidence of Epstein-Barr positivity was 100%. This malignancy does not have a predilection for any particular age group. The median CD4+ T cell count at the time of diagnosis is approximately 50/uL. Thus, CNS lymphoma generally presents at a later stage of HIV infection than systemic lymphoma. This fact may at least in part explain the poorer prognosis for this subset of patients.

The clinical presentation of lymphoma in patients with HIV infection is quite varied, ranging from focal seizures to rapidly growing mass lesions in the oral mucosa (Fig. 309-36) to persistent unexplained fever. At least 80% of patients present with extranodal disease, and a similar percentage have B-type symptoms of fever, night sweats, or weight loss. Virtually any site in the body may be involved. The most common extranodal site is theCNS, which is involved in approximately one-third of all patients with lymphoma. Approximately 60% of these cases are primary CNS lymphoma. Primary CNS lymphoma generally presents with focal neurologic deficits, including cranial nerve findings, headaches, and/or seizures.MRI orCTgenerally reveals a limited number (one to three) of 3- to 5-cm lesions (Fig. 309-37). The lesions often show ring enhancement on contrast administration and may occur in any location. Locations that are most commonly involved with CNS lymphoma are deep in the white matter. Contrast enhancement is usually less pronounced than that seen with toxoplasmosis. The main diseases in the differential diagnosis are cerebral toxoplasmosis and cerebral Chagas' disease. In addition to the 20% of lymphomas in HIV-infected individuals that are primary CNS lymphomas, CNS disease is also seen in HIV-infected patients with systemic lymphoma. Approximately 20% of patients with systemic lymphoma have CNS disease in the form of leptomeningeal involvement. This fact underscores the importance of lumbar puncture in the staging evaluation of patients with systemic lymphoma.

Systemic lymphoma is seen at earlier stages of HIV infection than primaryCNSlymphoma. In one series the mean CD4+ T cell count was 189/uL. In addition to lymph node involvement, systemic lymphoma may commonly involve the gastrointestinal tract, bone marrow, liver, and lung. Gastrointestinal tract involvement is seen in ~25% of patients. Any site in the gastrointestinal tract may be involved, and patients may complain of difficulty swallowing or abdominal pain. The diagnosis is usually suspected on the basis ofCT orMRI of the abdomen. Bone marrow involvement is seen in ~20% of patients and may lead to pancytopenia. Liver and lung involvement are each seen in ~10% of patients. Pulmonary disease may present as either a mass lesion, multiple nodules, or an interstitial infiltrate.

Both conventional and unconventional approaches have been employed in an attempt to treat HIV-related lymphomas. Systemic lymphoma is generally treated by the oncologist with combination chemotherapy. Earlier disappointing figures are being replaced with more optimistic results for the treatment of systemic lymphoma following the availability of more effective combination antiretroviral therapy. As in most situations in patients with HIV disease, those with the higher CD4+ T cell counts tend to do better. Response rates as high as 72% and disease-free intervals >15 months have been reported. Treatment of primaryCNSlymphoma remains a significant challenge. Treatment is complicated by the fact that this illness usually occurs in patients with advanced HIV disease. Palliative measures such as radiation therapy provide some relief. The prognosis remains poor in this group, with median survival <1 year.

Evidence of infection with *human papilloma virus*, associated with *intraepithelial dysplasia of the cervix or anus*, is approximately twice as common in HIV-infected individuals as in the general population and can lead to intraepithelial neoplasia and eventually invasive cancer. It is anticipated that both anal and cervical carcinomas will be seen with increased frequency in the HIV-infected population as survival is prolonged with combination antiretroviral therapy. In two separate studies, HIV-infected men without anorectal symptoms were studied for evidence of dysplasia, and Papanicolauo (Pap) smears were found to be abnormal in 40%. These changes were persistent at 1 year follow-up, raising the possibility of a subsequent transition to a more malignant condition. While the incidence of an abnormal Pap smear of the cervix is ~5% in otherwise healthy women, the incidence of abnormal cervical smears in women with HIV infection is 60%. Based on this finding, *invasive cervical cancer* was added to the list of AIDS-defining conditions. Thus far, however, only small increases in the incidence of cervical or anal cancer have been seen as a consequence of HIV infection. However, given this high rate of dysplasia, a comprehensive gynecologic and rectal examination, including Pap smear, is indicated at the initial evaluation and 6 months later for all patients with HIV infection. If these examinations are negative at both time points, the patient should be followed with yearly evaluations. If an initial or repeat Pap smear shows evidence of severe inflammation with reactive squamous changes, the next Pap smear should be performed at 3 months. If, at any time, a Pap smear shows evidence of squamous intraepithelial lesions, colposcopic examination with biopsies as indicated should be performed.

IDIOPATHIC CD4+ T LYMPHOCYTOPENIA

A syndrome was recognized in 1992 that was characterized by an absolute CD4+ T cell count of<300/uL or <20% of total T cells on more than one occasion; no evidence of HIV-1, HIV-2,HTLV-I, or HTLV-II on testing; and the absence of any defined immunodeficiency or therapy associated with decreased levels of CD4+ T cells. By mid-1993, approximately 100 patients had been described. After extensive multicenter investigations, a series of reports were published in early 1993, which together allowed a number of conclusions. Idiopathic CD4+ lymphocytopenia (ICL) is a very rare syndrome, as determined by studies of blood donors and cohorts of HIV-seronegative homosexual men. Cases were clearly identified as early as 1983, and cases remarkably similar to ICL had been identified decades ago. The definition of ICL based on CD4+ T cell counts coincided with the ready availability of testing for CD4+ T cells in patients suspected of being immunosuppressed. Although, as a result of immune deficiency, certain patients with ICL develop some of the opportunistic diseases (particularly cryptococcosis) seen in HIV-infected patients, the syndrome is demographically, clinically, and immunologically unlike HIV infection and AIDS. Fewer than half of the reported ICL patients had risk factors for HIV infection, and there were wide geographic and age distributions. The fact that a significant proportion of patients did have risk factors probably reflects a selection bias, in that physicians who take care of HIV-infected patients are more likely to monitor CD4+ T cells. Approximately one-third of the patients are women, compared to 16% of women among HIV-infected individuals in the United States. Many patients with ICL remained clinically stable, and their condition did not deteriorate progressively as is common with seriously immunodeficient HIV-infected patients. Certain patients with ICL even experienced spontaneous reversal of the CD4+ T lymphocytopenia. Immunologic abnormalities in ICL are somewhat different from those of HIV infection. ICL patients often also have decreases in CD8+ T cells and in B cells. Furthermore, immunoglobulin levels were either normal or, more commonly, decreased in patients with ICL, compared to the usual hypergammaglobulinemia of HIV-infected individuals. Finally, virologic studies revealed no evidence of HIV-1, HIV-2, HTLV-I, or HTLV-II or of any other mononuclear cell-tropic virus. Furthermore, there was no epidemiologic evidence to suggest that a transmissible microbe was involved. The cases of ICL were widely dispersed, with no clustering. Close contacts and sexual partners who were studied were clinically well and were serologically, immunologically, and virologically negative for HIV. ICL is a heterogeneous syndrome, and it is highly likely that there is no common cause; however, there may be common causes among subgroups of patients that are currently unrecognized.

Patients who present with laboratory data consistent with ICL should be worked up for underlying diseases that could be responsible for the immune deficiency. If no underlying cause is detected, no specific therapy should be initiated. However, if opportunistic diseases occur, they should be treated appropriately (see above). Depending on the level of the CD4+ T cell count, patients should receive prophylaxis for the commonly encountered opportunistic infections.

TREATMENT

General Principles of Patient Management The treatment of patients with HIV infection requires not only a comprehensive knowledge of the possible disease processes that may occur but also the ability to deal with the problems of a chronic, potentially life-threatening illness. Great advances have been made in the treatment of patients with HIV infection. The appropriate use of potent combination antiretroviral therapy and other treatment and prophylactic interventions is of critical importance in providing each patient with the best opportunity to live a long and healthy life despite the presence of HIV infection. In contrast to the earlier days of this epidemic, a diagnosis of HIV infection need no longer be equated with an inevitably fatal disease. In addition to medical interventions, the health care provider has a responsibility to provide each patient with appropriate counseling and education concerning their disease as part of a comprehensive care plan. Patients must be educated about the potential transmissibility of their infection and about the fact that while health care providers may refer to levels of the virus as "undetectable" this is more a reflection of the sensitivity of the assay being used to measure the virus than a comment on the presence or absence of the virus. It is important for patients to be aware and that the virus is still present and capable of being transmitted at all stages of HIV disease. Thus, there needs to be frank discussions concerning sexual practices and the sharing of needles. The treating physician must not only be aware of the latest medications available for patients with HIV infection but must also educate patients concerning the natural history of their illness and listen and be sensitive to their fears and concerns. As with other diseases, therapeutic decisions should be made in consultation with the patient, when possible, and with the patient's proxy if the patient is incapable of making decisions. In this regard, it is recommended that all patients with HIV infection, and in particular those with CD4+ T cell counts <200/uL, designate a trusted individual with durable power of attorney to make medical decisions on their behalf, if necessary.

No matter how well prepared a patient is for adversity, the discovery of a diagnosis of HIV infection is a devastating event. For this reason, it is recommended that anyone about to undergo testing have "pretest counseling" to prepare him or her at least partially should the results demonstrate the presence of HIV infection. Following a diagnosis of HIV infection, the health care provider should be prepared to activate support systems immediately for the newly diagnosed patient. These should include an experienced social worker or nurse who can spend time talking to the person and ensuring that he or she is emotionally stable. Most communities have HIV crisis centers that can be of great help in these difficult situations.

Following a diagnosis of HIV infection, there are several examinations and laboratory studies that should be performed to help determine the extent of disease and provide baseline standards for future reference (Table 309-19). In addition to routine chemistry and hematology screening panels and chest x-ray, one should also obtain a CD4+ T cell count, two separate plasma HIV RNA levels, aVDRLtest, and an anti-*Toxoplasma* antibody titer. APPD test should be done, and aMMSEperformed and recorded. Patients should be immunized with pneumococcal polysaccharide and, if seronegative for these viruses, with hepatitis A and hepatitis B vaccines. In addition, patients should be counseled with regard to sexual practices and needle sharing, and counseling should be offered to those whom the patient knows or suspects may also be infected. Once these baseline activities are performed, short- and long-term medical management strategies should be developed based uoon the most recent information available and modified as new information becomes available. The field of HIV medicine is changing rapidly, and it is difficult to remain fully up to date. Fortunately there are a series of excellent sites on the World Wide Web that are frequently updated, and they provide the most recent information on a variety of topics, including consensus panel reports on treatment (Table 309-20).

Antiretroviral Therapy Combination antiretroviral therapy, orHAART, is the cornerstone of management of patients with HIV infection. Following the initiation of widespread use of HAART in the United States in 1995 to 1996, marked declines have been noted in the incidence of most AIDS-defining conditions (Fig. 309-28). Suppression of HIV replication is an important component in prolonging life as well as improving the quality of life in patients with HIV infection. Unfortunately, many of the most important questions related to the treatment of HIV disease currently lack definitive answers. Among them are the questions of when should therapy be started, what is the best initial regimen, when should a given regimen be changed, and what should it be changed to when a change is made. Notwithstanding these uncertainties, the physician and patient must come to a mutually agreeable plan based upon the best available data. In an effort to facilitate this process, the United States Department of Health and Human Services has published a series of frequently updated guidelines including the "*Principles of Therapy of HIV Infection*," "*Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents*," and "*Guidelines for the Prevention of Opportunistic Infections in Persons Infected with Human Immunodeficiency Virus*." At present, an extensive clinical trials network, involving both clinical investigators and patient advocates, is in place attempting to develop improved approaches to therapy. Consortia comprising representatives of academia, industry, and the federal government are involved in the process of drug development, including clinical trials. As a result, new therapies and new therapeutic strategies are continually emerging. New drugs are often available through expanded access programs prior to official licensure. Given the complexity of this field, decisions regarding antiretroviral therapy are best made in consultation with experts. Currently licensed drugs for the treatment of HIV infection fall into two main categories: those that inhibit the viral reverse transcriptase enzyme (Table 309-21,Fig. 309-3*B*) and those that inhibit the viral protease enzyme. There are numerous drug-drug interactions that one must take into consideration when using these agents (Table 309-22).

TheFDA-approved reverse transcriptase inhibitors include the *nucleoside analogues* zidovudine, didanosine, zalcitabine, stavudine, lamivudine, and abacavir and the *nonnucleoside reverse transcriptase inhibitors* nevirapine, delavirdine, and efavirenz (Fig. 309-38;Table 309-21). These were the first class of drugs that were licensed for the treatment of HIV infection. They are indicated for this use as part of combination regimens. It should be stressed that none of these drugs should be used as monotherapy for HIV infection. Thus, when lamivudine is used to treat hepatitis B infection in the setting of HIV infection, one should ensure that the patient is also on additional antiretroviral medication. The reverse transcriptase inhibitors block the HIV replication cycle at the point of RNA-dependent DNA synthesis, the reverse transcription step. While the nonnucleoside reverse transcriptase inhibitors are quite selective for the HIV-1 reverse transcriptase, the nucleoside analogues inhibit a variety of DNA polymerization reactions in addition to those of the HIV-1 reverse transcriptase. For this reason, serious side effects are more common with the nucleoside analogues and include mitochondrial damage that can lead to hepatic steatosis and lactic acidosis as well as peripheral neuropathy and pancreatitis.

Zidovudine (AZT; 3¢-azido-2¢,3¢-dideoxythymidine) was the first drug approved for the treatment of HIV infection and is the prototype nucleoside analogue. These compounds, in which the hydroxyl group in the 3¢ position of the ribose moiety is substituted with a hydrogen or other chemical group, act as DNA chain terminators owing to their inability to form a 3¢-5¢phosphodiester linkage with another nucleoside. They bind much more avidly to the active site of the RNA-dependent DNA polymerase of HIV (reverse transcriptase) than to the active site of mammalian cell DNA polymerases; this explains their selective effect on HIV replication. Zidovudine also has a relatively high avidity for the DNA polymerase-g of human mitochondria. This may contribute to the development of the fatty liver and the myopathy sometimes observed in patients taking zidovudine. As with all the nucleoside analogues, the active form of zidovudine is the triphosphate, and the rate of phosphorylation, a thymidine kinase-dependent pathway, may be different in different cells. This may explain why zidovudine is more effective at inhibiting HIV replication in some cells than others. The clinical efficacy of zidovudine was clearly established in 1986 in a phase II, randomized, placebo-controlled trial in patients with advanced HIV disease. However, while treatment of patients with early stages of HIV infection was associated with increases in CD4+ T cell count, it was not associated with a better overall outcome than later intervention. Subsequent trials established the ability of this drug to dramatically decrease the incidence of perinatal transmission of HIV from infected mother to infant. Eventually a series of studies demonstrated the superiority of combination antiretroviral regimens over zidovudine alone, and combination therapy (discussed below) remains the standard of treatment today. Among the side effects of zidovudine at the initiation of therapy are fatigue, malaise, nausea, and headache. These side effects often subside over time. Patients on zidovudine may develop a macrocytic anemia, myopathy, cardiomyopathy, and lactic acidosis associated with fatty infiltration of the liver. As with every antiretroviral drug, HIV has the ability to develop resistance to zidovudine. Zidovudine resistance has been reported to occur ~6 months following the initiation of zidovudine monotherapy. More recently, zidovudine-resistant viruses have been noted in patients with acute infection prior to the initiation of therapy, implying that zidovudine-resistant viruses can be transmitted from person to person. Resistance emerges more rapidly in late-stage patients, presumably as a consequence of a greater degree of viral replication and thus a greater opportunity for mutation. A variety of amino acid changes including substitutions, insertions, and deletions have been reported to confer zidovudine resistance (Fig. 309-39). A combination preparation, Combivir, consists of zidovudine and lamivudine.

Didanosine (ddI; 2¢,3¢-dideoxyinosine) was the second drug licensed for the treatment of HIV infection, followed shortly thereafter by zalcitabine. Didanosine is metabolized to dideoxyadenosine in vivo. It is best absorbed on an empty stomach at a high pH. For this reason, the current formulations of didanosine contain a buffer, and each dose must be administered in no fewer than two tablets to ensure adequate buffering of stomach acid. The toxicity profile of didanosine is quite different from that of zidovudine. The most common toxicity is a painful sensory peripheral neuropathy that occurs in ~30% of patients receiving >400 mg/d. It generally resolves with discontinuation of the drug and may not recur if the drug is resumed at a reduced dose. At higher doses than are currently used one may see pancreatitis in ~10% of patients. Pancreatitis associated with didanosine therapy can be fatal. Didanosine should be discontinued if a patient experiences abdominal pain consistent with pancreatitis or if an elevated serum amylase or lipase is found in association with an edematous pancreas on ultrasound. Didanosine is contraindicated in patients with a prior history of pancreatitis, regardless of etiology.

Zalcitabine (ddC; 2¢,3¢-dideoxycytidine) is rarely used today in the management of patients with HIV infection. Among the nucleoside analogues licensed for the treatment of HIV infection, it is probably the weakest. The main toxicity of ddC is pancreatitis.

Stavudine (d4T; 2¢,3¢-didehydro-3¢-deoxythymidine) was the fourth drug licensed for the treatment of HIV infection. Like zidovudine, stavudine is a thymidine analogue. These two drugs are antagonistic in vitro and in vivo and should not be given together. Peripheral neuropathy and hepatic steatosis are the main toxicities of stavudine. It is commonly used with lamivudine as part of an initial treatment regimen.

Lamivudine (3TC; 2¢,3¢-dideoxy-3¢-thiacytidine) is the fifth of the nucleoside analogues to be licensed in the United States. It is licensed for use in combination with zidovudine in situations where zidovudine is indicated. In actual practice, lamivudine, is a frequent element of many different combination regimens currently in use. It is available either alone or in combination with zidovudine (Combivir). One reason behind the excellent synergy seen between lamivudine and the other nucleoside analogues may be that strains of HIV resistant to lamivudine (M184V substitution) appear to have enhanced sensitivity to other nucleosides, and thus development of dual resistance is quite difficult. In addition, there is a suggestion that 3TC-resistant strains of HIV may be less virulent and are less able to generate new mutants than are strains of HIV that are 3TC-sensitive. Lamivudine is among the best tolerated and least toxic nucleoside analogues.

Abacavir lcub;(1S,cis)-4-[2-amino-6-(cyclopropylamino)-9H-purin-9-yl]-2-cyclopentene-1-methanol sulfate (salt)(2:1)rcub; is a synthetic carbocyclic analogue of the nucleoside guanosine. It is licensed to be used in combination with other antiretroviral agents for the treatment of HIV-1 infection. Hypersensitivity reactions have been reported in ~5% of patients treated with this drug, and patients developing signs or symptoms of hypersensitivity such as fever, skin rash, fatigue, and gastrointestinal symptoms should discontinue the drug and not restart it. Fatal hypersensitivity reactions have been reported with rechallenge. Abacavir-resistant strains of HIV are typically also resistant to lamivudine, didanosine, and zalcitabine.

Nevirapine, *delavirdine*, and *efavirenz* are nonnucleoside inhibitors of the HIV-1 reverse transcriptase. They are licensed for use in combination with nucleoside analogues for the treatment of HIV-infected adults. These agents inhibit reverse transcriptase by binding to regions of the enzyme outside the active site and causing conformational changes in the enzyme that render it inactive. Although these agents are active in the nanomolar range, they are also very selective for the reverse transcriptase of HIV-1, have no activity against HIV-2, and, when used as monotherapy, are associated with the rapid emergence of drug-resistant mutants (Table 309-21;Fig. 309-39). Efavirenz is administered once a day, nevirapine twice a day, and delavirdine three times a day. All three drugs are associated with the development of a maculopapular rash, generally seen within the first few weeks of therapy. While it is possible to treat through this rash, it is important to be sure that one is not dealing with a more severe eruption such as Stevens-Johnson syndrome by looking carefully for signs of mucosal involvement, significant fever, or painful lesions with desquamation. In addition to skin rash, many patients treated with efavirenz note a feeling of light-headedness, dizziness, or out of sorts following the initiation of therapy. Some complain of vivid dreams. These symptoms tend to disappear after several weeks of therapy. Aside from difficulties with dreams, taking efavirenz at bedtime may minimize the side effects. Nevirapine and efavirenz are both commonly used as part of initial treatment regimens in combination with two nucleoside analogues. Another common use of these drugs is as part of salvage regimens in patients whose current regimen is inadequate.

The introduction of the HIV-1 protease inhibitors (saquinavir, indinavir, ritonavir, nelfinavir, and amprenavir) to the therapeutic armamentarium of antiretrovirals has had a profound impact on the efficacy of antiretroviral therapy. When used as part of initial regimens in combination with reverse transcriptase inhibitors, these agents have been shown to be capable of suppressing levels of HIV replication to under 50 copies per milliliter in the majority of patients for a minimum of 3 years. As in the case of reverse transcriptase inhibitors, resistance to protease inhibitors can develop rapidly in the setting of monotherapy, and thus these agents should be used as part of combination therapeutic regimens. A summary of known resistance mutations for reverse transcriptase and protease inhibitors is shown in Fig. 309-39.

Saquinavir was the first of the HIV-1 protease inhibitors to be licensed. Initially provided as a hard gel (Invirase) with poor bioavailability, the current soft-get formulation (Fortavase) provides good plasma levels of drug, particularly when administered in conjunction with ritonavir. Saquinavir is metabolized by the cytochrome P450 system, and ritonavir therapy results in inhibition of cytochrome P450 action. Thus, when both drugs are administered together there is the potential for increases in saquinavir levels. The use of low doses of ritonavir to provide pharmacodynamic boosting of other agents has become a fairly common strategy in HIV therapy. Saquinavir is perhaps the best-tolerated protease inhibitor and the one with the fewest side effects.

Ritonavir was the first protease inhibitor for which clinical efficacy was demonstrated. In a study of 1090 patients with CD4+ T cell counts <100/uL who were randomized to receive either placebo or ritonavir in addition to any other licensed medications, patients receiving ritonavir had a reduction in the cumulative incidence of clinical progression or death from 34% to 17%. Mortality decreased from 10.1% to 5.8%. At full doses, ritonavir is poorly tolerated. Among the main side effects are nausea, diarrhea, abdominal pain, and circumoral paresthesia. Ritonavir has a high affinity for several isoforms of cytochrome P450, and its use can result in large increases in the plasma concentrations of drugs metabolized by this pathway. Among the agents affected in this manner are saquinavir, indinavir, macrolide antibiotics, R-warfarin, ondansetron, rifampin, most calcium channel blockers, glucocorticoids, and some of the chemotherapeutic agents used to treatKS. In addition, ritonavir may increase the activity of glucuronyltransferases, thus decreasing the levels of drugs metabolized by this pathway. Overall, great care must be taken when prescribing additional drugs to patients taking ritonavir. As mentioned above, the pharmacodynamic boosting property of ritonavir, seen with doses as low as 100 to 200 mg twice a day, is often used in the setting of HIV infection to derive more convenient regimens. For example, when given with low-dose ritonavir, saquinavir and indinavir can both be given on twice-a-day schedules and taken with food.

Indinavir is among the best studied of the HIV-1 protease inhibitors. It was the first protease inhibitor used in combination with dual nucleoside therapy. The combination of zidovudine, lamivudine, and indinavir was the first "triple combination" shown to have a profound effect on HIV replication. The main side effects of indinavir are nephrolithiasis (seen in 4% of patients) and asymptomatic indirect hyperbilirubinemia (seen in 10%). Indinavir is predominantly metabolized by the liver. The dose should be lowered in patients with cirrhosis. Indinavir shares metabolic pathways with terfenadine, astemizole, cisapride, triazolam, and midazolam. To avoid the potential for cardiac arrhythmias or prolonged sedation, these drugs should not be administered to patients taking indinavir. Levels of indinavir are decreased during concurrent therapy with rifabutin or nevirapine and increased during concurrent therapy with ketoconazole, delavirdine, efavirenz, or ritonavir. Dosages should be modified appropriately in these circumstances (Table 309-22).

Nelfinavir was approved in 1997 and *amprenavir* was approved in 1999 for the treatment of adult or pediatric HIV infection when antiretroviral therapy is warranted. As with most of the newer antiretroviral agents, these approvals were based on randomized, controlled trials that demonstrated decreases in plasma HIV RNA levels and increases in CD4+ T cell counts. Both agents have unique resistance profiles. Nelfinavir resistance is associated with a D30N substitution in the protease gene. Viruses harboring this single mutation retain sensitivity to other protease inhibitors, and it has been suggested that for this reason nelfinavir is a good initial protease inhibitor. It is not clear, however, whether this theoretical consideration will be borne out in the results of clinical trials. Protease inhibitor resistance typically involves multiple amino acid substitutions and reduced susceptibility across the class. Amprenavir resistance is associated with a unique substitution at amino acid 50 (I50V), and it has been suggested that amprenavir may be of particular value in salvage regimens. This assumption also awaits verification in controlled clinical trials. Nelfinavir and amprenavir are both associated with gastrointestinal side effects. About 1% of patients receiving amprenavir have experienced severe and life-threatening skin reactions. An additional disadvantage of amprenavir is that the current formulation requires the patient to take 8 large capsules twice a day.

One of the main problems that has been encountered with the widespread use ofHAARTtherapy has been a syndrome of hyperlipidemia and fat distribution often referred to as *lipodystrophy syndrome* (discussed above under metabolic abnormalities).

The principles of therapy for HIV infection have been articulated by a panel sponsored by the U.S. Department of Health and Human Services and the Henry J. Kaiser Family Foundation. These principles are summarized in Table 309-23. However, *one element of HIV disease not currently covered by these principles is that eradication of HIV infection has not yet been possible*. Treatment decisions must take into account the fact that one is dealing with a chronic infection. While early therapy is generally the rule in infectious diseases, immediate treatment of every HIV-infected individual upon diagnosis may not be prudent, and therapeutic decisions must take into account the balance between risks and benefits. At present, a reasonable course of action is to initiate antiretroviral therapy in anyone with the acute HIV syndrome; patients with symptomatic disease; patients with asymptomatic disease with CD4+ T cell counts<500/uL or with>20,000 copies of HIV RNA per milliliter (Table 309-24). In addition, one may wish to administer a 6-week course of therapy to uninfected individuals immediately following a high-risk exposure to HIV (see below).

Once the decision has been made to initiate therapy, the health care provider must decide which drugs to use as the first regimen. The decision regarding choice of drugs not only will affect the immediate response to therapy but also will have implications regarding options for future therapeutic regimens. The initial regimen is usually the most effective insofar as the virus has yet to develop significant resistance. The two options for initial therapy most commonly in use today are two different three-drug regimens. The first regimen utilizes two nucleoside analogues (one of which is usually lamivudine) and a protease inhibitor. The second regimen utilizes two nucleoside analogues and a nonnucleoside reverse transcriptase inhibitor. Unfortunately there are no clear data at present on which to base distinctions between these two approaches. Following the initiation of therapy one should expect a 1 log (tenfold) reduction in plasma HIV RNA levels within 1 to 2 months and eventually a decline in plasma HIV RNA levels to <50 copies per milliliter. During this same time there should be a rise in the CD4+ T cell count of 100 to 150/uL that is particularly brisk during the first month of therapy. Many clinicians feel that failure to achieve this endpoint is an indication for a change in therapy. Other reasons for a change in therapy include a persistently declining CD4+ T cell count, clinical deterioration, or drug toxicity (Table 309-25). As in the case of initiating therapy, changing therapy may have a lasting impact on future therapeutic options. When changing therapy because of treatment failure (clinical progression or worsening laboratory parameters), it is important to attempt to provide a regimen with at least two new drugs. In the patient in whom a change is made for reasons of drug toxicity, a simple replacement of one drug is reasonable. It should be stressed that in attempting to sort out a drug toxicity it may be advisable to hold all therapy for a period of time to distinguish between drug toxicity and disease progression. Drug toxicity will usually begin to show signs of reversal within 1 to 2 weeks. Prior to changing a treatment regimen because of drug failure, it is important to ensure that the patient has been adherent to the prescribed regimen. As in the case of initial therapy, the simpler the therapeutic regimen, the easier it is for the patient to be compliant. Plasma HIV RNA levels and CD4+ T lymphocyte counts should be monitored every 3 to 4 months during therapy and more frequently if one is contemplating a change in regimen or immediately following a change in regimen.

In an attempt to determine an optimal therapeutic regimen, one may attempt to measure antiretroviral drug susceptibility through genotyping or phenotyping of HIV quasispecies. Genotyping may be done through dideoxynucleotide sequencing, DNA chip hybridization, or line probe assays. Phenotypic assays measure the performance of reverse transcriptase or protease in the presence or absence of different concentrations of different drugs. These assays will generally detect quasispecies present at a frequency of at least 10%. The precise role of resistance testing in the management of patients with HIV infection is not yet clear. While randomized studies have suggested that information regarding HIV resistance profiles may improve therapeutic outcomes in patients failing their current antiretroviral regimen, the degree of improvement thus far has been small and the duration of the benefit limited. Resistance testing may be of particular value in distinguishing drug-resistant virus from poor patient compliance; it may also be of value to help guide initial therapy in a setting where transmission of a drug-resistant isolate is felt to be likely.

In addition to the licensed medications discussed above, a large number of experimental agents are being evaluated as possible therapies for HIV infection. Therapeutic strategies are being developed that interfere with virtually every step of the replication cycle of the virus (Fig. 309-3). In addition, as more is discovered about the role of the immune system in controlling viral replication, additional strategies, generically referred to as "immune-based therapies," are being developed as a complement to antiviral therapy. Among the antiviral agents in early clinical trials are additional nucleoside analogues, nucleotide analogues, additional protease inhibitors including nonpeptidomimetic compounds, integrase inhibitors, antisense nucleic acids, and fusion inhibitors. Among the immune-based therapies being evaluated areIFN-a, bone marrow transplantation, adoptive transfer of lymphocytes genetically modified to resist infection or enhance HIV-specific immunity, active immunotherapy with inactivated HIV, andIL-2.

HIV AND THE HEALTH CARE WORKER

Health care workers, especially those who deal with large numbers of HIV-infected patients, have a small but definite risk of becoming infected with HIV as a result of professional activities. As of January 1, 2000, 56 health care workers in the United States had been documented as having seroconverted to HIV following occupational exposure; 25 have developed AIDS. The individuals who seroconverted include 19 laboratory workers (16 of whom were clinical laboratory workers), 23 nurses, 6 physicians, 2 surgical technicians, 1 dialysis technician, 1 respiratory therapist, 1 health aide, 1 embalmer/morgue technician, and 2 housekeeper/maintenance workers. The exposures included 48 percutaneous (puncture/cut injury), 5 mucocutaneous (mucous membrane and/or skin), 2 both percutaneous and mucocutaneous, and 1 unknown route of exposure. Fifty exposures were to HIV-infected blood, three to concentrated virus in a laboratory, one to visibly bloody fluid, and one to unspecified fluid. As of January 1, 2000, there had been 136 other cases of HIV infection or AIDS among health care workers who have not reported other risk factors for HIV infection and who report a history of exposure to blood, body fluids, or HIV-infected laboratory material, but for whom seroconversion after exposure was not documented. The number of these workers who actually acquired their infection through occupational exposures is not known. Taken together, the data from several large studies suggest that the risk of HIV infection following a percutaneous injury with an HIV-contaminated hollow-bore needle (in contrast to a solid-bore needle, i.e., a suture needle) is approximately 0.3%. A seroprevalence survey of 3420 orthopedic surgeons, 75% of whom practiced in an area with a relatively high prevalence of HIV infection and 39% of whom reported percutaneous exposure to patient blood, usually through an accident involving a suture needle, failed to reveal any cases of possible occupational infection, suggesting that the risk of infection with a suture needle may be considerably less than that with a blood-drawing needle.

Most cases of health care worker seroconversion occur as a result of needle-stick injuries. When one considers the circumstances that result in needle-stick injuries, it is immediately obvious that adhering to the standard guidelines for dealing with sharp objects would result in a significant decrease in this type of accident. In one study, 27% of needle-stick injuries resulted from improper disposal of the needle (over half of these were due to recapping the needle), 23% occurred during attempts to start an intravenous line, 22% occurred during blood drawing, 16% were associated with an intramuscular or subcutaneous injection, and 12% were associated with giving an intravenous infusion.

Recommendations regarding postexposure prophylaxis must take into account that several circumstances determine the risk of transmission of HIV following occupational exposure. In this regard, five factors have been associated with an increased risk for occupational transmission of HIV infection: deep injury, the presence of visible blood on the instrument causing the exposure, injury with a device that had been placed in the vein or artery of the source patient, terminal illness in the source patient, and lack of postexposure antiretroviral therapy in the exposed health care worker. Other important considerations include pregnancy in the health care worker and the possibility of exposure to drug-resistant virus. Regardless of the decision to use postexposure prophylaxis, the wound should be cleansed immediately and antiseptic applied. If a decision is made to offer postexposure prophylaxis, U.S. Public Health Service guidelines recommend (1) a combination of two nucleoside analogue reverse transcriptase inhibitors given for 4 weeks for routine exposures, or (2) a combination of two nucleoside analogue reverse transcriptase inhibitors plus a protease inhibitor given for 4 weeks for high-risk or otherwise complicated exposures, although most clinicians administer the latter regimen in all cases in which a decision is made to treat. Further details are available from the U.S. Public Health Service *Guidelines for the Management of Health-Care Worker Exposures to HIV and Recommendations for Postexposure Prophylaxis* (CDC, 1998).

Health care workers can minimize their risk of occupational HIV infection by following theCDCguidelines of July 1991, which include adherence to universal precautions, refraining from direct patient care if one has exudative lesions or weeping dermatitis, and disinfecting and sterilizing reusable devices employed in invasive procedures. The premise of universal precautions is that every specimen should be handled as if it came from someone infected with a bloodborne pathogen. All samples should be double-bagged, gloves should be worn when drawing blood, and spills should be immediately disinfected with bleach.

In attempting to put this small but definite risk to the health care worker in perspective, it is important to point out that approximately 200 health care workers die each year as a result of occupationally acquired hepatitis B infection. The tragedy in this instance is that these infections and deaths due toHBVcould be greatly decreased by more extended use of the HBV vaccine. The risk of HBV infection following a needle-stick injury from a hepatitis antigen-positive patient is much higher than the risk of HIV infection (see "Transmission," above). There are multiple examples of needle-stick injuries where the patient was positive for both HBV and HIV and the health care worker became infected only with HBV. For these reasons, it is advisable, given the high prevalence of HBV infection in HIV-infected individuals, that all health care workers dealing with HIV-infected patients be immunized with the HBV vaccine.

TBis another infection common to HIV-infected patients that can be transmitted to the health care worker. For this reason, all health care workers should know theirPPDstatus, have it checked yearly, and receive one year of isoniazid treatment if their skin test converts to positive. In addition, all patients in whom a diagnosis of TB is being entertained should be placed immediately in respiratory isolation, pending results of the diagnostic evaluation. The emergence of drug-resistant organisms has made TB an increasing problem for health care workers. This is particularly true for the health care worker with preexisting HIV infection.

One of the most charged issues ever to come between health care workers and patients is that of transmission of infection from HIV-infected health care workers to their patients. This is discussed under "Occupational Transmission of HIV: Health Care Workers and Laboratory Workers," p. 1857. Theoretically, the same universal precautions that are used to protect the health care worker from the HIV-infected patient will also protect the patient from the HIV-infected health care worker.

VACCINES

Historically, vaccines have provided a safe, cost-effective, and efficient means of preventing illness, disability, and death from infectious diseases. Given the fact that human behavior, especially human sexual behavior, is extremely difficult to change, the best hope for preventing the spread of HIV infection rests with the development of a safe and effective vaccine. This task is problematic for a number of reasons, including the high mutability of the virus, the fact that the infection can be transmitted by cell-free or cell-associated virus, the likely need for the development of effective mucosal immunity, and the fact that it has been difficult to establish the precise correlates of protective immunity to HIV infection. Some HIV-infected individuals are long-term nonprogressors (see above), and a number of individuals have been exposed to HIV multiple times but remain uninfected; these facts suggest that there are protective elements of an HIV-specific immune response. In addition, studies using animal models, specificallySIV in the monkey and HIV-1 in the chimpanzee, have been encouraging and suggest that an HIV vaccine is possible. It should be pointed out that while the ideal goal of an HIV vaccine is to prevent infection, a vaccine given to an uninfected individual that significantly alters the course of disease or the infectivity of the individual, should that person become infected, could have an impact not only on the individual in question but also on the spread of infection in the community.

A number of clinical trials ranging from several small phase I trials to determine safety, to fewer intermediate-sized phase II trials to determine safety and immunogenicity, to a single phase III trial to determine efficacy have been or are currently being conducted in humans. The single phase III trial is testing a bivalent gp 120 protein; this product has been shown to induce antibodies but not cytolytic T cells responses in phase I and II trials. The furthest advanced among phase II trials involves a combination approach using a live canarypox vector expressing one or multiple HIV epitopes given together with gp120 or using the gp120 as a boost. This approach has resulted in neutralizing antibodies in virtually all recipients and HIV-specific cytolytic T cells in approximately 30% of individuals at any given time during the course of the trial.

Other approaches currently being tested in phase I and/or phase II trials in humans include naked DNA; vaccines employing vectors such as modified vaccinia Ankara (MVA), salmonella, Venezuela equine encephalitis (VEE) virus, among others; peptide and subunit vaccines; and pseudovirions (Fig. 309-40). Live attenuated HIV vaccines have not proceeded into human trials at this time because of safety concerns. It is clear that it will take several years of clinical trials to establish the efficacy or lack thereof of a candidate vaccine for HIV.

PREVENTION

Education, counseling, and behavior modification are the cornerstones of an HIV prevention strategy. Widespread voluntary testing of individuals who have practiced or are practicing high-risk behavior, together with counseling of infected individuals, is recommended. Information gathered from such an approach should serve as the basis for behavior-modification programs, both for infected individuals who may be unaware of their HIV status and who could infect others and for uninfected individuals practicing high-risk behavior. The practice of "safer sex" is the most effective way for sexually active uninfected individuals to avoid contracting HIV infection and for infected individuals to avoid spreading infection. Abstinence from sexual relations is the only absolute way to prevent sexual transmission of HIV infection. However, this may not be feasible, and there are a number of relatively safe practices that can markedly decrease the chances of transmission of HIV infection. Partners engaged in monogamous sexual relationships who wish to be assured of safety should both be tested for HIV antibody. If both are negative, it must be understood that any divergence from monogamy puts both partners at risk; open discussion of the importance of honesty in such relationships should be encouraged. When the HIV status of either partner is not known, or when one partner is positive, there are a number of options. Use of condoms can markedly decrease the chance of HIV transmission. It should be remembered that condoms are not 100% effective in preventing transmission of HIV infection, and there is an ~10% failure rate of condoms used for contraceptive purposes. Most condom failures result from breakage or improper usage, such as not wearing the condom for the entire period of intercourse. Latex condoms are preferable, since virus has been shown to leak through natural skin condoms. Petroleum-based gels should never be used for lubrication of the condom, since they increase the likelihood of condom rupture. There has been a tendency among homosexual men to practice fellatio as a "minimal risk" activity compared to receptive anal intercourse. It should be emphasized that receptive oral fellatio is definitely not safe sex, and there has been clear-cut documentation of transmission of HIV where receptive fellatio was the only sexual act performed (see "Transmission," above). Topical microbicides for vaginal and anal use are being pursued actively as a means by which individuals could avoid infection when the insertive partner cannot be relied on to use a condom. Kissing is considered safe, although there is a theoretical possibility of transmission via virus in saliva. The low concentration of virus in saliva of infected individuals, as well as the presence in saliva of HIV-inhibitory proteins (see above), lessens any risk of transmission by kissing.

The most effective way to prevent transmission of HIV infection amongIDUs is to stop the use of injectable drugs. Unfortunately, that is extremely difficult to accomplish unless the individual enters a treatment program. For those who will not or cannot participate in a drug treatment program and who will continue to inject drugs, the avoidance of sharing of needles and other paraphernalia ("works") is the next best way to avoid transmission of infection. The cultural and social factors that contribute to the sharing of paraphernalia are complex and difficult to overcome. In addition, needles and syringes may be in short supply. Under these circumstances, paraphernalia should be cleaned after each usage with a virucidal solution, such as undiluted sodium hypochlorite (household bleach). Data from a number of studies have indicated that programs that provide sterile needles to addicts in exchange for used needles have resulted in a decrease in HIV transmission without increasing the use of injection drugs. It is important for IDUs to be tested for HIV infection and counseled, to avoid transmission to their sexual partners. Secondary and tertiary spread of HIV infection by the heterosexual route within settings of a high level of injection drug use has increased greatly in the United States (see above).

Transmission of HIV via transfused blood or blood products has been decreased dramatically by a combination of screening of all blood donors for HIV infection by assays for both HIV antibody and p24 antigen and self-deferral of individuals at risk for HIV infection. In addition, clotting factor concentrates are heat-treated, essentially eliminating the risk to hemophiliacs who require these products. Autologous transfusions are preferable to transfusions from another individual. However, logistic constraints as well as the unpredictability of the need for most transfusions limit the feasibility of this approach. At present the risk of becoming HIV-infected from a contaminated blood transfusion is approximately 1 in 676,000 donations.

HIV can be transmitted via breast milk and colostrum. The avoidance of breast feeding may not be practical in developing countries, where nutritional concerns override the risk of HIV transmission. However, it is becoming appreciated that from 5 to 15% of infants who were born of HIV-infected mothers and who were fortunate enough not to have been infected intrapartum or peripartum become infected via breast feeding. Therefore, even in developing countries, breast feeding from an infected mother should be avoided if at all possible. Unfortunately, this is rarely the case, and given the disadvantages of withholding breast feeding in developing countries (see above), health authorities in most developing countries continue to recommend breast feeding despite the potential for HIV transmission. In developed countries such as the United States, where bottled formula and milk are readily accessible, breast feeding is absolutely contraindicated when a mother is HIV positive.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 2 -DISORDERS OF IMMUNE-MEDIATED INJURY

310. ALLERGIES, ANAPHYLAXIS, AND SYSTEMIC MASTOCYTOSIS - *K. Frank Austen*

The term *atopic allergy* implies a familial tendency to manifest such conditions as asthma, rhinitis, urticaria, and eczematous dermatitis (atopic dermatitis) alone or in combination. However, individuals without an atopic background may also develop hypersensitivity reactions, particularly urticaria and anaphylaxis, associated with the same class of antibody, IgE, found in atopic individuals. Inasmuch as the mast cell is the key effector cell of the biologic response in allergic rhinitis, urticaria, anaphylaxis, and systemic mastocytosis, the introduction to these clinical problems will consider the developmental biology, activation pathway, product profile, and target tissues for this cell type.

The fixation of IgE to human mast cells and basophils, a process termed *sensitization*, prepares these cells for subsequent antigen-specific activation. The interaction of the high-affinity Fc receptor for IgE, designated FceRI, upregulates the cellular expression of the receptor, possibly by ligand-mediated stabilization. FceRI is composed of onea, one b, and two disulfide-linked g chains, which together cross the plasma membrane seven times. The a chain is solely responsible for IgE binding, and the b andg chains are responsible for signal transduction that results from the aggregation of the tetrameric receptors by polymeric antigen.

The interaction of specific multivalent antigen with receptor-bound IgE results in clustering of the receptors to initiate signal transduction through the action of a *src* family-related tyrosine kinase, termed *Lyn*, that is constitutively associated with the b chain. Lyn transphosphorylates the canonical immunoreceptor tyrosine-based activation motifs (ITAMs) of theb and g chains of the receptor, resulting in recruitment of more active Lyn to theb chain and of the Syk/zap-70 family tyrosine kinases. The two phosphorylated tyrosines in the ITAMs function as binding sites for the tandem *src* homology two (SH2) domains within these kinases. It appears that Syk activates not only phospholipase Cg but also phosphatidylinositol-3-kinase to provide phosphatidyl-3,4,5-triphosphate, which allows membrane targeting of the Tec family kinases (Btk and Itk) and their activation by Lyn. The resulting Tec kinase-dependent phosphorylation of phospholipase Cg with cleavage of its phospholipid membrane substrate provides inositol-1,4,5-triphosphate (IP$_3$) and 1,2-diacylglycerols (1,2-DAGs) so as to mobilize intracellular calcium and activate protein kinase C. The subsequent opening of calcium-regulated activated channels provides the sustained elevations of intracellular calcium required to recruit the mitogen-activated protein kinases, JNK and p38 (serine/threonine kinases), which provide cascades to augment arachidonic acid release and to mediate nuclear translocation of transcription factors for various cytokines. The calcium ion-dependent activation of phospholipases cleaves membrane phospholipids to generate lysophospholipids, which, like 1,2-DAG, are fusogenic and may facilitate the fusion of the secretory granule perigranular membrane with the cell membrane, a step that releases the membrane-free granule containing the preformed or primary mediators of mast cell effects.

The secretory granule of the human mast cell has a crystalline structure, unlike mast cells of lower species, and IgE-dependent cell activation can be characterized morphologically by solubilization and swelling of the granule contents within the first minute of receptor perturbation; this reaction is followed by the ordering of intermediate filaments about the swollen granule, movement toward the cell surface, and fusion of the perigranular membrane with that of other granules and with the plasmalemma to form extracellular channels for mediator release while maintaining cell viability.

In addition to exocytosis, aggregation of FceRI initiates two other pathways for generation of bioactive products, namely, lipid mediators and cytokines. The biochemical steps involved in expression of such cytokines as tumor necrosis factor a(TNF-a), interleukin (IL) 6, IL-4, IL-5, granulocyte-macrophage colony-stimulating factor (GM-CSF), and others have not been specifically defined for mast cells. Nonetheless, inhibition studies of cytokine production (IL-1b, TNF-a, and IL-6) in mouse mast cells with cyclosporine or FK506 reveal binding to the ligand-specific immunophilin and attenuation of the calcium ion- and calmodulin-dependent serine/threonine phosphatase, calcineurin.

Lipid mediator generation (Fig. 310-1) involves translocation of calcium ion-dependent cytosolic phospholipase A_2 to the perinuclear membrane, with subsequent release of arachidonic acid for metabolic processing by the distinct prostanoid and leukotriene pathways. The constitutive prostaglandin endoperoxide synthase (PGHS-1/cyclooxygenase-1) and the de novo inducible PGHS-2 (cyclooxygenase-2) convert released arachidonic acid to the sequential intermediates prostaglandin (PG) G_2 and PGH_2. The glutathione-dependent hematopoietic PGD_2synthase then converts PGH_2 to PGD_2, the predominant mast cell prostanoid.

For processing by the leukotriene pathway, the released arachidonic acid is translocated to an integral perinuclear membrane protein, the 5-lipoxygenase activating protein (FLAP). The calcium ion-dependent activation of 5-lipoxygenase involves translocation to the perinuclear membrane, which allows conversion of the arachidonic acid to the sequential intermediates, 5-hydroperoxyeicosatetraenoic acid and leukotriene (LT) A_4. LTA_4 is conjugated with reduced glutathione by LTC_4synthase, an integral membrane protein with significant homology to FLAP. Intracellular LTC_4 is released by a carrier-specific export step for extracellular conversion to the receptor-active cysteinyl leukotrienes LTD_4 and LTE_4 by sequential removal of glumatic acid and glycine. A cytosolic LTA_4hydrolase converts some LTA_4 to the dihydroxy leukotriene LTB_4, which then undergoes specific export for extracellular receptor-mediated actions. The lysophospholipid formed during release of arachidonic acid from 1-*O*-alkyl-2-acyl-*sn*-glyceryl-3-phosphorylcholine can be acetylated in the second position to form platelet-activating factor (PAF).

Unlike other cells of bone marrow origin, mast cells leave the marrow and circulate as committed progenitors lacking their definitive secretory granules. These committed progenitors express the receptor, c-*kit*, for stem cell factor (SCF) before the expression of FceRI. Whereas c-*kit* is lost or markedly diminished in expression by other cell types, it is retained by mature, differentiated mast cells and is an absolute requirement for the development of constitutive tissue mast cells residing in skin and connective tissue sites and for the T cell-dependent mast cells residing in mucosal surfaces or undergoing reactive hyperplasia. Indeed, in clinical T cell deficiencies, mast cells are absent from the intestinal mucosa but are present in the submucosa. It is thus assumed that unrecognized mast cell progenitors enter the tissue and undergo regulated proliferation, differentiation, and maturation. Based on the immunodetection of secretory granule neutral proteases, mast cells in the lung parenchyma and intestinal mucosa selectively express tryptase; those in the intestinal and airway submucosa, skin, lymph nodes, and breast parenchyma express tryptase, chymase, and carboxypeptidase A (CPA); and occasional mast cells in intestinal submucosa express chymase and CPA but not tryptase. The secretory granules of mast cells selectively positive for tryptase in lung and intestinal mucosa exhibit closed scrolls with a periodicity suggestive of a crystalline structure by electron microscopy; whereas the secretory granules of mast cells with mutiple proteases residing in skin, lymph nodes, breast parenchyma, and submucosa of airways and intestine are scroll-poor, with an amorphous or latticelike appearance.

Mast cells are distributed at cutaneous and mucosal surfaces and in deeper tissues about venules and could regulate the entry of foreign substances by their rapid response capability (Fig. 310-2). Upon stimulus-specific activation in vitro, histamine and secretory granule-associated acid hydrolases are solubilized, whereas the neutral proteases, which are cationic, remain largely complexed to the anionic proteoglycans, heparin and chrondroitin sulfate E. The macromolecular complex serves to deliver the neutral proteases so that the endo- and exoproteases can function in concert at the substrate site to clear damaged tissue and facilitate repair. Histamine and the various lipid mediators (PGD_2,LTD_4/E_4,PAF) alter venular permeability, thereby allowing influx of plasma proteins such as complement and immunoglobulins, whereas LTB_4mediates leukocyte-endothelial cell adhesion with subsequent directed migration (chemotaxis). The accumulation of leukocytes and opsonins would facilitate defense of the microenvironment. The cysteinyl leukotrienes constrict both vascular and nonvascular smooth muscle and are much more potent than histamine in constricting human airway smooth muscle when administered by aerosol.

The cellular component of the inflammatory response elicited by preformed secretory granule-associated and membrane-derived lipid mediators would be augmented and sustained by the addition of cytokines of mast cell or T cell origin to the microenvironment. Activation of human skin mast cells in situ elicitsTNF-aproduction and release, which in turn induces endothelial cell responses favoring leukocyte adhesion. Activation of purified human lung mast cells in vitro results in substantial production ofIL-5 and lesser quantities of IL-4. Bronchial biopsies of patients with bronchial asthma reveal that mast cells are immunohistochemically positive for IL-4 and IL-5, but that the predominant localization of IL-4, IL-5, andGM-CSFis to T cells, defined as T_H2 by this profile. It is speculated that IL-4 modulates the T cell phenotype to the T_H2 subtype, and that IL-5 or GM-CSF converts infiltrating eosinophils to an activated, autoaggressive phenotype with augmented capacity for cytotoxicity and generation of $O_{2-}$and the cysteinyl leukotrienes.

The view of immediate and late cellular phase of allergic inflammation is supported by the response of the skin, nose, or lung of allergic humans to local allergen challenge; greater quantities of allergen are needed to elicit the cellular phase. In the immediate phase of a local challenge, there is pruritus and watery discharge from the nose, bronchospasm and mucous secretion in the lungs, and a wheal-and-flare response with pruritus in the skin. The reduced nasal patency, reduced pulmonary function, or evident erythema with swelling at the skin site in a late-phase response at 6 to 8 h are associated with biopsy findings of infiltrating and activated TH2 type T cells, eosinophils, basophils, and even some neutrophils. This allergic inflammation proceeding from early mast cell activation to late cellular infiltration is believed to promote end-organ hyperresponsivity, as would be characteristic of perennial rhinitis or bronchial asthma; for attenuation, it requires introduction of an anti-inflammatory agent such as a glucocorticoid. The particular chemokines responsible for directed migration of eosinophils and T cells after their integrin-dependent endothelial cell adhesion are not yet defined, although eotaxin is a likely contributor since both cell types, as well as basophils, express the selective receptor CCR-3.

Consideration of the mechanism of immediate type hypersensitivity diseases in the human has focused largely on the IgE-dependent recognition of otherwise nontoxic substances. A region of chromosome 5 (5q23-31) contains genes implicated in the control of IgE levels including IL-4 and IL-13, as well as IL-3 and IL-9 involved in reactive mast cell hyperplasia and IL-5 and GM-CSF central to eosinophil development and their enhanced tissue viability. Genes with linkage to the specific IgE response to particular allergens include those encoding the major histocompatibility complex (MHC) and certain chains of the T cell receptor (TCR-ad). The complexity of atopy and the associated diseases is such that susceptibility, severity, and therapeutic responses most likely relate not only to specific IgE but also to constitutive target tissue reactivity and the superimposed effects of the local inflammatory response mediated by TH2 cells, mast cells, basophils, and eosinophils.

The induction of allergic disease requires sensitization of a predisposed individual to specific allergen. This sensitization can occur anytime in life, although the greatest propensity for the development of allergic disease appears to occur in childhood and early adolescence. Exposure of a susceptible individual to an allergen results in processing of the allergen by antigen-presenting cells, including macrophage-like cells located throughout the body at surfaces that contact the outside environment, such as the nose, lungs, eyes, skin, and intestine. These antigen-presenting cells process the allergen protein and present the epitope-bearing peptides via theirMHC to particular T cell subsets. The T cell response depends both on cognate recognition through various ligand/receptor interactions and on the cytokine microenvironment, withIL-4 directing a TH2 response and interferon (IFN)g a TH1 profile. T cells can potentially induce several responses to an allergen, including those typical of contact dermatitis, known as the TH1 type response, and those mediated by IgE, known as the TH2 allergic response. The TH2 response is associated with activation of specific B cells that transform into plasma cells. Synthesis and release into the serum of allergen-specific IgE by plasma cells result in sensitization of IgE Fc receptor-bearing cells including mast cells and basophils, which subsequently are capable of becoming activated upon exposure to the specific allergen. In certain diseases, including those associated with atopy, the monocyte and eosinophil populations can express a trimeric high-affinity receptor, FceRI, which lacks the b chain, and yet respond to its aggregation.

ANAPHYLAXIS

DEFINITION

The life-threatening anaphylactic response of a sensitized human appears within minutes after administration of specific antigen and is manifested by respiratory distress often followed by vascular collapse or by shock without antecedent respiratory difficulty. Cutaneous manifestations exemplified by pruritus and urticaria with or without angioedema are characteristic of such systemic anaphylactic reactions. Gastrointestinal manifestations include nausea, vomiting, crampy abdominal pain, and diarrhea.

PREDISPOSING FACTORS AND ETIOLOGY

There is no convincing evidence that age, sex, race, occupation, or geographic location predisposes a human to anaphylaxis except through exposure to some immunogen. According to most studies, atopy does not predispose individuals to anaphylaxis from penicillin therapy or venom of a stinging insect but is a risk factor for allergens in food or latex.

The materials capable of eliciting the systemic anaphylactic reaction in humans include the following: heterologous proteins in the form of hormones (insulin, vasopressin, parathormone), enzymes (trypsin, chymotrypsin, penicillinase, streptokinase), pollen extracts (ragweed, grass, trees), nonpollen extracts (dust mites, dander of cats, dogs, horses, and laboratory animals), food (milk, eggs, seafood, nuts, grains, beans, gelatin in capsules), antiserum (antilymphocyte gamma globulin), occupation-related proteins (latex rubber products), and Hymenoptera venom (yellow jacket, yellow and baldfaced hornets, paper wasp, honey bee, imported fire ants); polysaccharides such as dextran and thiomerosal as a vaccine preservative; and most commonly drugs such as protamine and antibiotics (penicillins, cephalosporins, amphotericin B, nitrofurantoin, quinolones), local anesthetics (procaine, lidocaine), muscle relaxants (suxamethonium, gallamine, pancuronium), vitamins (thiamine, folic acid), diagnostic agents (sodium dehydrocholate, sulfobromophthalein), and occupation-related chemicals (ethylene oxide), which are considered to function as haptens that form immunogenic conjugates with host proteins. The conjugating hapten may be the parent compound, a nonenzymatically derived storage product, or a metabolite formed in the host.

PATHOPHYSIOLOGY AND MANIFESTATIONS

Individuals differ in the time of appearance of symptoms and signs, but the hallmark of the anaphylactic reaction is the onset of some manifestation within seconds to minutes after introduction of the antigen, generally by injection or less commonly by ingestion. There may be upper or lower airway obstruction or both. Laryngeal edema may be experienced as a "lump" in the throat, hoarseness, or stridor, while bronchial obstruction is associated with a feeling of tightness in the chest and/or audible wheezing. Patients with bronchial asthma are predisposed to severe involvement of the lower airways. A characteristic feature is the eruption of well-circumscribed, discrete cutaneous wheals with erythematous, raised, serpiginous borders and blanched centers. These urticarial eruptions are intensely pruritic and may be localized or disseminated. They may coalesce to form giant hives, and they seldom persist beyond 48 h. A localized, nonpitting, deeper edematous cutaneous process, angioedema, may also be present. It may be asymptomatic or cause a burning or stinging sensation.

In fatal cases with clinical bronchial obstruction, the lungs show marked hyperinflation on gross and microscopic examination. The microscopic findings in the bronchi, however, are limited to luminal secretions, peribronchial congestion, submucosal edema, and eosinophilic infiltration, and the acute emphysema is attributed to intractable bronchospasm that subsides with death. The angioedema resulting in death by mechanical obstruction occurs in the epiglottis and larynx, but the process is also evident in the hypopharynx and to some extent in the trachea; on microscopic examination there is wide separation of the collagen fibers and the glandular elements; vascular congestion and eosinophilic infiltration are also present. Patients dying of vascular collapse without antecedent hypoxia from respiratory insufficiency have visceral congestion with a presumptive loss of intravascular blood volume. The associated electrocardiographic abnormalities, with or without infarction, noted in some patients may reflect a primary cardiac event or be secondary to a critical reduction in blood volume.

The angioedematous and urticarial manifestations of the anaphylactic syndrome have been attributed to release of endogenous histamine. A role for the cysteinyl leukotrienes in altering pulmonary mechanics by causing marked bronchiolar constriction seems likely. Vascular collapse without respiratory distress in response to experimental challenge with the sting of a hymenopteran was associated not only with marked and prolonged elevations in blood histamine but also with evidence of intravascular coagulation and kinin generation. The findings that patients with systemic mastocytosis and episodic hypotension proceeding to vascular collapse excrete large amounts ofPGD$_2$metabolites in addition to histamine and that these events are controlled by administration of a nonsteroidal agent but not by antihistamines alone suggest that PGD$_2$ is also of importance in the hypotensive anaphylactic reactions. The cysteinyl leukotrienes may be involved in the pathobiologic process in patients with myocardial ischemia without or with infarction.

DIAGNOSIS

The diagnosis of an anaphylactic reaction depends largely on an accurate history revealing the onset of the appropriate symptoms and signs within minutes after the responsible material is encountered. When only a portion of the full syndrome is present, such as isolated urticaria, sudden bronchospasm in a patient with asthma, or vascular collapse after intravenous administration of an agent, it may be appropriate to consider a complement-mediated immune complex reaction, an idiosyncratic response to any of the nonsteroidal anti-inflammatory agents, or the direct effect of certain drugs or diagnostic agents on mast cells. Intravenous administration of a chemical mast cell-degranulating agent, including opiate derivatives and radiographic contrast media, may elicit generalized urticaria, angioedema, and a sensation of retrosternal oppression with or without clinically detectable bronchoconstriction or hypotension. Aspirin and other nonsteroidal anti-inflammatory agents such as indomethacin, aminopyrine, and mefenamic acid may precipitate a life-threatening episode of obstruction of upper or lower airways, especially in patients with asthma, that is clinically reminiscent of anaphylaxis but is not associated with a detectable IgE response. This syndrome, which is commonly associated with nasal polyposis, is due to inhibition of PGHS-1 with corresponding unregulated, amplified generation of the cysteinyl leukotrienes via the 5-lipoxygenase/LTC$_4$synthase pathway. In the transfusion anaphylactic reaction that occurs in patients with IgA deficiency, the responsible specificity resides in IgG or IgE anti-IgA; the mechanism of the reaction mediated by IgG anti-IgA is presumed to be complement activation with secondary mast cell participation.

The presence of specific IgE in the heart blood of patients dying of systemic anaphylaxis has been demonstrated at postmortem by passive transfer of the serum intradermally into a normal recipient, followed in 24 h by antigen challenge into the same site, with subsequent development of a wheal and flare, the Prausnitz-Kustner reaction. To avoid the hazards of transferring hepatitis or other infections to a recipient, it is preferable to use the serum to seek passive sensitization of a human leukocyte suspension enriched with basophils for subsequent antigen-induced histamine release. Furthermore, radioimmunoassays have demonstrated specific IgE antibodies in patients with anaphylactic reactions, but such approaches require purified antigens. Elevations of b-tryptase levels in serum implicate mast cell activation in an adverse systemic reaction and are particularly informative with episodes of hypotension during general anesthesia or when there has been a fatal outcome.

TREATMENT

Early recognition of an anaphylactic reaction is mandatory, since death occurs within minutes to hours after the first symptoms. Mild symptoms such as pruritus and urticaria can be controlled by administration of 0.2 to 0.5 mL of 1:1000 epinephrine subcutaneously, with repeated doses as required at 20-min intervals for a severe reaction. If the antigenic material was injected into an extremity, the rate of absorption may be reduced by prompt application of a tourniquet proximal to the reaction site, administration of 0.2 mL of 1:1000 epinephrine into the site, and removal without compression of an insect stinger, if present. An intravenous infusion should be initiated to provide a route for administration of 2.5 mL epinephrine, diluted 1:10,000, at 5- to 10-min intervals, volume expanders such as normal saline, and vasopressor agents such as dopamine if intractable hypotension occurs. Replacement of intravascular volume due to postcapillary venular leakage may require several liters of saline. Epinephrine provides both a- andb-adrenergic effects, resulting in vasoconstriction, bronchial smooth-muscle relaxation, and attenuation of enhanced venular permeability. Beta blockers are relatively contraindicated in persons at risk for anaphylactic reactions, especially those sensitive to Hymenoptera venom or those undergoing immunotherapy for respiratory system allergy. When epinephrine fails to control the anaphylactic reaction, hypoxia due to airway obstruction or related to a cardiac arrhythmia, or both, must be considered. Oxygen via a nasal catheter or intermittent positive-pressure breathing of oxygen with 0.5 mL isoproterenol diluted 1:200 in saline may be helpful, but either endotracheal intubation or a tracheostomy is mandatory for oxygen delivery if progressive hypoxia develops. Ancillary agents such as the antihistamine diphenhydramine, 50 to 100 mg intramuscularly or intravenously, and aminophylline, 0.25 to 0.5 g intravenously, are appropriate for urticaria-angioedema and bronchospasm, respectively. Intravenous glucocorticoids are not effective for the acute event but may alleviate later recurrence of bronchospasm, hypotension, or urticaria. Furthermore, in a syndrome termed *idiopathic anaphylaxis* with recurrent angioedema of the upper airways, glucocorticoid administration may be beneficial by reducing the frequency of attacks and/or the severity of episodes.

PREVENTION

Prevention of anaphylaxis must take into account the sensitivity of the recipient, the dose and character of the diagnostic or therapeutic agent, and the effect of the route of administration on the rate of absorption. If there is a definite history of a past anaphylactic reaction, even though mild, it is advisable to select another agent or procedure. A knowledge of cross-reactivity among agents is critical since, for example, cephalosporins share a commonb-lactam ring with the penicillins. A skin test should be performed before the administration of certain materials that are likely to elicit anaphylactic reactions, such as allergenic extracts, or when the nature of the past adverse reaction is unknown. A scratch test should precede an intradermal test in very sensitive patients. With regard to penicillin, two-thirds of patients with a positive reaction history and positive skin tests to benzylpenicilloyl-polylysine (BPL) and/or the minor determinant mixture (MDM) of benzylpenicillin products experience allergic reactions with treatment, and these are almost uniformly of the anaphylactic type in those patients with minor determinant reactivity. Even patients without a history of previous clinical reactions have a 2 to 6 percent incidence of positive skin tests to the two test materials, and about 3 per 1000 with a negative history experience anaphylaxis with therapy, with a mortality of about 1 per 100,000. Skin testing for antibiotics should be performed only on patients with a positive clinical history consistent with an IgE-mediated reaction and in imminent need of the antibiotic in question; skin testing is of no value for non-IgE-mediated eruptions. Desensitization with most antibiotics can proceed by the intravenous, subcutaneous, or oral route. Typically, graded quantities of the antibiotic are given by the selected route using double doses until a therapeutic dosage is achieved. Due to the risk of systemic anaphylaxis during the course of desensitization, such a procedure should be performed only in a setting in which resuscitation equipment is at hand and an intravenous line is in place. It is critical to give the therapeutic agent at regular intervals to prevent the reestablishment of a sensitized cell pool of large size.

A different form of protection involves the development of blocking antibody of the IgG class, which is protective against Hymenoptera venom-induced anaphylaxis by interacting with antigen so that less reaches the sensitized tissue mast cells; to be effective, this immunotherapy requires the use of specific or cross-reacting Hymenoptera venom. Because sensitization can be transient, the maximal risk for systemic anaphylactic reactions in persons with Hymenoptera sensitivity occurs in association with a currently positive skin test. Although there is only low-grade cross-reactivity between honey bee and yellow jacket venoms, there is a high degree of cross-reactivity between yellow jacket venom and the rest of the vespid venoms (yellow or baldfaced hornets and wasps). Prevention involves modification of outdoor activities to exclude bare feet, wearing perfumed toiletries, eating in areas attractive to insects, clipping hedges or grass, and hauling away trash or fallen fruit. As with each anaphylactic sensitivity, the individual should wear an informational bracelet and have immediate access to an unexpired epinephrine kit. The limitations of lifestyle and the psychological duress can be addressed by venom immunotherapy to achieve a venom-specific IgG titer. Although it has been recommended that venom therapy be continued indefinitely or until the skin and specific serum IgE tests are unremarkable, there is evidence that 5 years of treatment induces a state of resistance to sting reactions that is independent of serum levels of specific IgG or IgE. This contrasts with the definite relation of sting immunity to specific IgG earlier in the treatment regime. For children with a systemic reaction limited to skin, the likelihood of progression to more serious respiratory or vascular manifestations is low, and thus immunotherapy is not recommended.

URTICARIA AND ANGIOEDEMA

DEFINITION

Urticaria and angioedema may appear separately or together as cutaneous manifestations of localized nonpitting edema; a similar process may occur at mucosal surfaces of the upper respiratory or gastrointestinal tract. *Urticaria* involves only the superficial portion of the dermis, presenting as well-circumscribed wheals with erythematous raised serpiginous borders with blanched centers that may coalesce to become giant wheals. *Angioedema* is a well-demarcated localized edema involving the deeper layers of the skin, including the subcutaneous tissue. Recurrent episodes of urticaria and/or angioedema of less than 6 weeks' duration are considered acute, whereas attacks persisting beyond this period are designated chronic.

PREDISPOSING FACTORS AND ETIOLOGY

The occurrence of urticaria and angioedema is probably more frequent than usually described because of the evanescent, self-limited nature of such eruptions, which seldom require medical attention when limited to the skin. Although persons in any age group may experience acute or chronic urticaria and/or angioedema, these lesions increase in frequency after adolescence, with the highest incidence occurring in persons in the third decade of life; indeed, one survey of college students indicated that 15 to 20% had experienced a pruritic wheal reaction.

The classification of urticaria-angioedema presented inTable 310-1 focuses on the different mechanisms for eliciting clinical disease and can be useful for differential diagnosis; nonetheless, most cases of chronic urticaria are idiopathic. Urticaria and/or angioedema occurring during the appropriate season in patients with seasonal respiratory allergy or as a result of exposure to animals or molds is attributed to inhalation or physical contact with pollens, animal dander, and mold spores, respectively. However, urticaria and angioedema secondary to inhalation are relatively uncommon compared to urticaria and angioedema elicited by ingestion of fresh fruits, shellfish, fish, milk products, chocolate, legumes including peanuts, and various drugs that may elicit not only the anaphylactic syndrome with prominent gastrointestinal complaints but also chronic urticaria.

Additional etiologies include physical stimuli such as cold, heat, solar rays, exercise, and mechanical irritation. The physical urticarias can be distinguished by the precipitating event and other aspects of the clinical presentation. *Dermographism*, which occurs in 1 to 4% of the population, is defined by the appearance of a linear wheal at the site of a brisk stroke with a firm object or by any configuration appropriate to the eliciting event. Dermographism has a prevalence that peaks in the second to third decades. It is not influenced by an atopic diathesis and has a duration generally of less than 5 years. *Pressure urticaria*, which often accompanies dermographism or chronic idiopathic urticaria, presents in response to a sustained stimulus such as a shoulder strap or belt, running (feet), or manual labor (hands). *Cholinergic urticaria* is distinctive in that the pruritic wheals are of small size (1 to 2 mm) and are surrounded by a large area of erythema; attacks are precipitated by fever, a hot bath or shower, or exercise and are presumptively attributed to a rise in core body temperature. *Exercise-related anaphylaxis* can be limited to erythema and pruritic urticaria but may progress to angioedema of the face, oropharynx, larynx, or intestine or to vascular collapse; it is distinguished from cholinergic urticaria by presenting with wheals of conventional size and by not occurring with fever or a hot bath. *Cold urticaria*, either acquired or hereditary, is local at body areas exposed to low ambient temperature or cold objects (ice cube) but can progress to vascular collapse with immersion in cold water (swimming). *Solar urticaria* is subdivided into three groups by the response to specific portions of the light spectrum. *Vibratory angioedema* may occur after years of occupational exposure or can be idiopathic; it may be accompanied by cholinergic urticaria. Other rare forms of physical allergy, always defined by stimulus-specific elicitation, include *local heat urticaria*, *aquagenic urticaria* from contact with water of any temperature (sometimes associated with polycythemia vera), and *contact urticaria* from direct interaction with some chemical substance.

Angioedema without urticaria occurs with C1 inhibitor (C1INH) deficiency that may be inborn as an autosomal dominant characteristic or may be acquired. The urticaria and angioedema associated with classic serum sickness or with hypocomplementemic cutaneous necrotizing angiitis are believed to be immune-complex diseases. The drug reactions to mast cell granule-releasing agents and to nonsteroidal anti-inflammatory drugs may be systemic, resembling anaphylaxis, or limited to cutaneous sites.

PATHOPHYSIOLOGY AND MANIFESTATIONS

Urticarial eruptions are distinctly pruritic, involve any area of the body from the scalp to the soles of the feet, and appear in crops of 24- to 72-h duration, with old lesions fading as new ones appear. The most common sites for urticaria are the extremities and face, with angioedema often being periorbital and in the lips. Although self-limited in duration, angioedema of the upper respiratory tract may be life-threatening due to laryngeal obstruction, while gastrointestinal involvement may present with abdominal colic, with or without nausea and vomiting, and may precipitate unnecessary surgical intervention. No residual discoloration occurs with either urticaria or angioedema unless there is an underlying process leading to superimposed extravasation of erythrocytes.

The pathology of urticaria and angioedema is usually characterized by edema of the dermis in urticaria and of the subcutaneous tissue as well as the dermis in angioedema. Collagen bundles in affected areas are widely separated, and the venules are sometimes dilated. The perivenular infiltrate may consist of lymphocytes, eosinophils, and neutrophils that are present in varying combination and number throughout the dermis.

Perhaps the best-studied example of IgE- and mast cell-mediated urticaria and angioedema is *cold urticaria*. Cryoglobulins may be recognized, but not in the majority of patients. Immersion of an extremity in an ice bath precipitates angioedema of the distal portion with urticaria at the air interface within minutes of the challenge. Histologic studies reveal marked mast cell degranulation with associated edema of the dermis and subcutaneous tissues. The venous effluent of the cold-challenged and angioedematous extremity reveals a marked rise in plasma content of histamine, whereas the venous effluent of the contralateral normal extremity contains none of this mediator. Elevated levels of histamine have been found in the plasma of venous effluent and in the fluid of suction blisters at experimentally induced lesional sites in patients with dermographism, pressure urticaria, vibratory angioedema, light urticaria, and heat urticaria. By ultrastructural analysis, the pattern of mast cell degranulation in cold urticaria resembles an IgE-mediated response with solubilization of granule contents, fusion of the perigranular and cell membranes, and discharge of granule contents, whereas in a dermographic lesion there is an additional superimposed zonal (piecemeal) degranulation. Elevations of plasma histamine levels with biopsy-proven mast cell degranulation have also been demonstrated with systemic attacks of *cholinergic urticaria* and *exercise-related anaphylaxis* precipitated experimentally in subjects exercising on a treadmill while wearing a wet suit; however, only in cholinergic urticaria is there a concomitant decrease in pulmonary function.

DIAGNOSIS

The rapid onset and self-limited nature of urticarial and angioedematous eruptions are distinguishing features. Additional characteristics are the occurrence of the urticarial crops in various stages of evolution and the asymmetric distribution of the angioedema. Urticaria and/or angioedema involving IgE-dependent mechanisms are often appreciated by historic considerations implicating specific allergens or physical stimuli, by seasonal incidence, and by exposure to certain environments. Direct reproduction of the lesion with physical stimuli is particularly valuable because it so often establishes the cause of the lesion. The diagnosis of an environmental allergen based on the clinical history can be confirmed by skin testing or assay for allergen-specific IgE in serum. IgE-mediated urticaria and/or angioedema may or may not be associated with an elevation of total IgE or with peripheral eosinophilia. Fever, leukocytosis, and an elevated sedimentation rate are absent.

The classification of urticarial and angioedematous states noted in Table 310-1 in terms of possible mechanisms necessarily includes some differential diagnostic points. Hypocomplementemia is not observed in IgE-mediated mast cell disease and may reflect either an acquired abnormality generally attributed to the formation of immune complexes or a genetic deficiency ofC1INH. Chronic recurrent urticaria, generally in females, associated with arthralgias, an elevated sedimentation rate, and normo- or hypocomplementemia suggests an underlying cutaneous necrotizing angiitis. Vasculitic urticaria typically persists longer than 72 h, whereas conventional urticaria often has a duration of less than 24 to 48 h. Confirmation depends on a biopsy that reveals cellular infiltration, nuclear debris, and fibrinoid necrosis of the venules. The same pathobiologic process accounts for the urticaria in association with such diseases as systemic lupus erythematosus or viral hepatitis with or without an associated arteritis. Serum sickness per se or a similar clinical entity due to drugs includes not only urticaria but also pyrexia, lymphadenopathy, myalgia, and arthralgia or arthritis. Urticarial reactions to blood products or intravenous administration of immunoglobulin are defined by the event and generally are not progressive unless the recipient is IgA-deficient in the former case or the reagent is aggregated in the latter.

Hereditary angioedema is an autosomal dominant disease due to a deficiency of antigenic and/or functionalC1INH. The diagnosis is suggested not only by family history but also by the lack of pruritus and of urticarial lesions, the prominence of recurrent gastrointestinal attacks of colic, and episodes of laryngeal edema. Laboratory diagnosis depends on demonstrating a deficiency of C1INH antigen (type 1) in most kindreds, but some kindreds have an antigenically intact nonfunctional protein (type 2) and require a functional assay to establish the diagnosis. The natural substrates of uninhibited C1 protease, C4 and C2, are chronically depleted but fall further during attacks due to the activation of additional C1. Because the C1INH protein also regulates the Hageman factor-initiated activation of kallikrein and of plasmin, the vasoactive peptides responsible for the angioedema are likely some combination of bradykinin and a plasmin-derived fragment of C1-cleaved C2. An acquired form of C1INH deficiency, associated with lymphoproliferative disorders, has the same clinical manifestations but differs in the lack of a familial element; in the reduction of C1 function and C1q protein as well as C1INH, C4, and C2; and in the presence of an anti-idiotypic antibody to the monoclonal immunoglobulin expressed on the B cells. A second acquired form of C1INH deficiency with angioedema due to the appearance of IgG anti-C1INH may be associated with systemic lupus erythematosus.

Urticaria and angioedema must be differentiated from contact sensitivity, a vesicular eruption that progresses to chronic thickening of the skin with continued allergenic exposure. They must also be differentiated from atopic dermatitis, a condition that may present as erythema, edema, papules, vesiculation, and oozing proceeding to a subacute and chronic stage in which vesiculation is less marked or absent and scaling, fissuring, and lichenification predominate in a distribution that characteristically involves the flexor surfaces. In cutaneous mastocytosis, the reddish brown macules and papules, characteristic of urticaria pigmentosa, urticate with pruritus upon trauma; and in systemic mastocytosis, without or with urticaria pigmentosa, there is an episodic systemic flushing with or without urticaria but no angioedema.

TREATMENT

Identification of the etiologic factor(s) and their elimination provide the most satisfactory therapeutic program; this approach is feasible to varying degrees with IgE-mediated reactions to allergens or physical stimuli. For most forms of urticaria, H_1antihistamines such as chlorpheniramine or diphenhydramine, and including the nonsedating class such as loratadine or cetirizine, are effective in attenuating both urtication and pruritus. Cyproheptadine and especially hydroxyzine have proven effective when H_1antihistamines have been inadequate. Doxepin, a dibenzoxepin tricyclic compound with both H_1 and H_2receptor antagonist activity, is yet another alternative. Topical glucocorticoids are of no value in the management of urticaria and/or angioedema. Systemic glucocorticoids are generally avoided in idiopathic, allergen-induced, or physical urticarias due to their long-term toxicity. However, systemic glucocorticoids are useful in the management of patients with pressure urticaria, with vasculitic urticaria (especially with eosinophil prominence), with idiopathic angioedema with or without urticaria, or with chronic urticaria that responds poorly to conventional treatment. With persistent vasculitic urticaria, hydroxychloroquine or colchicine may be added to the regimen after hydroxyzine and before or along with systemic glucocorticoids.

The therapy of inbornC1INHdeficiency has been simplified by the finding that attenuated androgens correct the biochemical defect and afford prophylactic protection. Since the affected individuals are heterozygous, with the depletion of C1INH being due to deficient synthesis and consequent excessive utilization of the limited amount available, the efficacy of the attenuated androgens is attributed to production by the normal gene of an amount of functional C1INH sufficient to control the spontaneous activation of C1 to C1 protease. Since the use of such agents for children and pregnant women is not yet accepted, the antifibrinolytic agente-aminocaproic acid may be used occasionally to control spontaneous attacks or for preoperative prophylaxis in some patients. This agent should not be used in patients with thrombotic tendencies or ischemia due to arterial athrosclerosis. Infusion of isolated C1INH protein appears useful in prophylaxis and to ameliorate an attack but is not yet widely available.

SYSTEMIC MASTOCYTOSIS

DEFINITION

Systemic mastocytosis is defined by mast cell hyperplasia that in most instances is indolent and nonneoplastic. Since human mast cells originate from pluripotent bone marrow cells (CD34+), circulate as nonmetachromatically staining, c-*kit*-positive mononuclear cells, and undergo tissue-specific proliferation and maturation, the hyperplasia is generally recognized only in bone marrow and in the normal peripheral distribution sites of the cells, such as skin (Fig. 310-CD1), gastrointestinal mucosa, liver, and spleen. Mastocytosis occurs at any age and has a slight preponderance in males. The prevalence of systemic mastocytosis is not known, a familial occurrence has not been established, and atopy is not increased.

CLASSIFICATION, PATHOPHYSIOLOGY, AND CLINICAL MANIFESTATIONS

A recent consensus classification for systemic mastocytosis recognizes four forms (Table 310-2). The form designated as *indolent* accounts for the majority of patients and is not known to alter life expectancy. When a patient is classified as having indolent systemic mastocytosis, the concomitant clinical findings must be carefully noted, since they define the complications and directions for management. In systemic mastocytosis *associated with hematologic disorders*, the prognosis is determined by the nature of that disorder, which can range from dysmyelopoiesis to leukemia. In *aggressive* systemic mastocytosis, mast cell proliferation in parenchymal organs such as liver, spleen, and lymph nodes is marked and in a subset of patients is associated with prominent eosinophilia in affected organs or peripheral blood; the prognosis is poor due to widespread tissue infiltration. *Mast cell leukemia* is the rarest form of the disease and is invariably fatal at present; in contrast to the other forms, the peripheral blood contains circulating, metachromatically staining, atypical mast cells. In types II and IV systemic mastocytosis there is a point mutation of the c-*kit* tyrosine kinase in the leukocytes and mast cells, respectively; this mutation can also be detected in lesional tissue such as the small, reddish brown macules or papules, termed *urticaria pigmentosa*, at skin sites of patients with type I. The most common mutation, a substitution of valine for aspartate in codon 816 (V816D), leads to constitutively activated c-*kit*, which then drives proliferation independently of SCF. More than half of the cases of type II systemic mastocytosis with a V816D mutation exhibit cytogenetic abnormalities on routine karyotyping. In types I and III there is excessive production of the c-*kit* ligand (SCF) in the microenvironment of the mast cells, and this may be autocrine in type III. In infants and children (type I) with cutaneous manifestations, namely, urticaria pigmentosa or bullous lesions, visceral involvement is usually lacking, and resolution is common because there are no mutations to activate the tyrosine kinase. The clinical manifestations of systemic mastocytosis, particularly types I and II, distinct from a leukemic complication, are due to tissue occupancy by the mast cell mass, the tissue response to that mass, and the release of bioactive substances acting at both local and distal sites. The pharmacologically induced manifestations are pruritus, flushing, palpitations and vascular collapse, gastric distress, lower abdominal crampy pain, and recurrent headache. The increase in cell burden is evidenced by the lesions of urticaria pigmentosa at skin sites, but it also contributes to bone pain and malabsorption. The mast cell-mediated fibrotic changes are limited to liver, spleen, and bone marrow and presumably relate to the functional characteristics of mast cells developing at those sites, as opposed to those at sites without fibrosis, such as the gastrointestinal tissue or skin. Immunofluorescent analysis of bone marrow and skin lesions in indolent mastocytosis and of spleen, lymph node, and skin in aggressive systemic mastocytosis has revealed only one mast cell phenotype, namely, scroll-poor cells expressing tryptase, chymase, and CPA.

The cutaneous lesions of urticaria pigmentosa respond to trauma with urtication and erythema (Darier's sign). The apparent incidence of these lesions is 90 percent or greater in patients with indolent systemic mastocytosis. Approximately 1 percent of patients with indolent mastocytosis have skin lesions that appear as tan-brown macules with striking patchy erythema and associated telangiectasia (telangiectasia macularis eruptiva perstans). In the upper gastrointestinal tract, histamine-mediated hypersecretion is the most common problem, with resultant gastritis and peptic ulcer. In the lower intestinal tract, the occurrence of diarrhea and abdominal pain is attributed to increased motility due to mast cell mediators, and this can be aggravated by malabsorption with secondary nutritional insufficiency and osteomalacia. The periportal fibrosis associated with mast cell infiltration and a prominence of eosinophils may lead to portal hypertension and ascites. In some patients, flushing and recurrent vascular collapse are markedly aggravated by an idiosyncratic response to a minimal dosage of nonsteroidal anti-inflammatory agents. The neuropsychiatric disturbances are clinically most evident as impaired recent memory, decreased attention span, and "migraine-like" headaches. Patients in every category of systemic mastocytosis may experience exacerbation of a specific clinical sign or symptom with alcohol ingestion, use of mast cell-interactive narcotics, or ingestion of nonsteroidal anti-inflammatory agents.

DIAGNOSIS

Although the diagnosis is generally suspected on the basis of the clinical history and physical findings, the contention can be strengthened by certain laboratory procedures and established only by a tissue diagnosis. A 24-h urine collection for measurement of histamine, histamine metabolites, or metabolites of$PGD_2$is currently the most common noninvasive approach. A convenient alternative is to measure blood levels of the mast cell-derived neutral protease tryptase. Thea form of tryptase is elevated in more than one-half of patients with type I systemic mastocytosis and in virtually all those with types II and III, whereas the b form is increased in patients undergoing an anaphylactic reaction. Additional studies directed by the presentation include a bone scan or skeletal survey; contrast studies of the upper gastrointestinal tract with small-bowel follow-through, computed tomography scan, or endoscopy; and a neuropsychiatric evaluation, including an electroencephalogram. The tissue diagnosis is straightforward if there are lesions of urticaria pigmentosa, but the diagnosis of systemic mastocytosis requires involvement of other organs and is most frequently established by bone marrow biopsy and aspiration. The bone marrow lesions consist of focal and paratrabecular aggregates of spindle-shaped mast cells, often mixed with eosinophils, lymphocytes, and, on occasion, plasma cells, histiocytes, and fibroblasts.

The differential diagnosis requires the exclusion of other flushing disorders. The 24-h urine assessment of 5-hydroxy-indoleacetic acid and metanephrines should exclude a carcinoid tumor or a pheochromocytoma. Most patients with recurrent anaphylaxis, including the idiopathic group, present with angioedema and/or wheezing, which are not manifestations of systemic mastocytosis.

TREATMENT

The management of systemic mastocytosis uses a stepwise and symptom/sign-directed approach that includes an H_1antihistamine for flushing and pruritus, an H_2antihistamine or proton pump inhibitor for gastric acid hypersecretion, oral cromolyn sodium for diarrhea and abdominal pain, and a nonsteroidal anti-inflammatory agent for severe flushing associated with vascular collapse despite use of H_1 and H_2antihistamines to block biosynthesis ofPGD$_2$. Systemic glucocorticoids appear to alleviate the malabsorption. Headaches are generally managed with tricyclic antidepressants and other neurotransmitter-modifying agents. Ketotifen has been used to alleviate flushing in patients with gastric intolerance to nonsteroidal anti-inflammatory agents and in patients with bone pain or intractable headaches. The efficacy ofIFN-ain aggressive systemic mastocytosis is controversial, and this may relate to the difficulty in achieving the necessary dosage in some patients due to the attendant side effects. Treatment with hydroxyurea to reduce the mast cell lineage progenitors may have merit in type III systemic mastocytosis. Chemotherapy is appropriate for the frank leukemias in types II and IV.

ALLERGIC RHINITIS

DEFINITION

Allergic rhinitis is characterized by sneezing; rhinorrhea; obstruction of the nasal passages; conjunctival, nasal, and pharyngeal itching; and lacrimation, all occurring in a temporal relationship to allergen exposure. Although commonly seasonal due to elicitation by airborne pollens, it can be perennial in an environment of chronic exposure. The incidence of allergic rhinitis in North America is about 7%, with the peak occurring in childhood and adolescence.

PREDISPOSING FACTORS AND ETIOLOGY

Allergic rhinitis generally presents in atopic individuals, i.e., in persons with a family history of a similar or related symptom complex and a personal history of collateral allergy expressed as eczematous dermatitis, urticaria, and/or asthma (Chap. 252). Symptoms generally appear before the fourth decade of life and tend to diminish gradually with aging, although complete spontaneous remissions are uncommon. A relatively small number of weeds that depend on wind rather than insects for cross-pollination, as well as certain grasses and trees, produce sufficient quantities of pollen suitable for wide distribution by air currents to elicit seasonal allergic rhinitis. The dates of pollination of these species generally vary little from year to year in a particular locale but may be quite different in another climate. In the temperate areas of North America, trees typically pollinate from March through May, grasses in June and early July, and ragweed from mid-August to early October. Molds, which are widespread in nature because they occur in soil or decaying organic matter, may propagate spores in a pattern dependent on climatic conditions. Perennial allergic rhinitis occurs in response to allergens that are present throughout the year such as in desquamating epithelium in animal dander or cockroach-derived proteins, the processed materials or chemicals utilized in an industrial setting, or the dust accumulating at work or at home. Dust has a diverse allergen content including *Dermatophagoides farinae* and *D. pteronyssinus*, which may be present alone or together in house dust. Dust mites are scavengers of flecks of human skin and coat the digestate with mite-specific protein for subsequent excretion as part of a fecal ball. In up to two-thirds of patients with perennial rhinitis, no clear-cut allergen can be demonstrated. The ability of allergens to cause rhinitis rather than lower respiratory symptoms may be attributed to their size, 10 to 100 um, and retention within the nose.

PATHOPHYSIOLOGY AND MANIFESTATIONS

Episodic rhinorrhea, sneezing, obstruction of the nasal passages with lacrimation, and pruritus of the conjunctiva, nasal mucosa, and oropharynx are the hallmarks of allergic rhinitis. The nasal mucosa is pale and boggy, the conjunctiva congested and edematous, and the pharynx is generally unremarkable. Swelling of the turbinates and mucous membranes with obstruction of the sinus ostia and eustachian tubes precipitates secondary infections of the sinuses and middle ear, respectively, commonly in perennial but rarely in seasonal disease. Nasal polyps, representing mucosal protrusions containing edema fluid with variable numbers of eosinophils, arise concurrently with edema and/or infection within the sinuses and increase obstructive symptoms.

The nose presents a large mucosal surface area through the folds of the turbinates and serves to adjust the temperature and moisture content of inhaled air and to filter out particulate materials above 10 um in size by impingement in a mucous blanket; ciliary action moves the entrapped particles toward the pharynx. Entrapment of pollen and digestion of the outer coat by mucosal enzymes such as lysozymes release protein allergens generally of 10,000 to 40,000 molecular weight. The initial interaction occurs between the allergen and intraepithelial mast cells and then proceeds to involve deeper perivenular mast cells, both of which are sensitized with specific IgE. During the symptomatic season when the mucosae are already swollen and hyperemic, there is enhanced adverse reactivity to the seasonal pollen as well as to antigenically unrelated pollens for which there is underlying hypersensitivity due to improved penetration of the allergens. Biopsy specimens of nasal mucosa during seasonal rhinitis show submucosal edema with infiltration by eosinophils, along with some basophils and neutrophils.

The mucosal surface fluid contains IgA that is present because of its secretory piece and also IgE, which apparently arrives by diffusion from plasma cells in proximity to mucosal surfaces. IgE fixes to mucosal and submucosal mast cells, and the intensity of the clinical response to inhaled allergens is quantitatively related to the naturally occurring pollen dose. Specific IgE is distributed also to circulating basophilic leukocytes; patients with more severe clinical disease have basophils that release histamine in response to lesser concentrations of allergen in vitro than do cells from patients with milder disease. In sensitive individuals, the introduction of allergen into the nose is associated with sneezing, "stuffiness," and discharge, and the fluid contains histamine,PGD_2, and leukotrienes. Thus the mast cells of the nasal mucosa and submucosa generate and release mediators through IgE-dependent reactions that are capable of producing tissue edema and eosinophilic infiltration.

DIAGNOSIS

The diagnosis of seasonal allergic rhinitis depends largely on an accurate history of occurrence coincident with the pollination of the offending weeds, grasses, or trees. The continuous character of perennial allergic rhinitis due to contamination of the home or place of work makes historic analysis difficult, but there may be a variability in symptoms that can be related to exposure to animal dander, dust mite and/or cockroach allergens, or work-related allergens such as latex. Patients with perennial rhinitis commonly develop the problem in adult life, are more often women than men, and manifest nasal polyps and thickening of the sinus membranes demonstrated by radiography. The term *vasomotor rhinitis* designates a condition of enhanced reactivity of the nasopharynx in which a symptom complex resembling perennial allergic rhinitis occurs with nonspecific stimuli. Other entities to be excluded are structural abnormalities of the nasopharynx; exposure to irritants; upper respiratory infection; pregnancy with prominent nasal mucosal edema; prolonged topical use of a-adrenergic agents in the form of nose drops (rhinitis medicamentosa); and the use of certain therapeutic agents such as rauwolfia, b-adrenergic antagonists, or estrogens.

The nasal secretions of allergic patients are rich in eosinophils, and peripheral eosinophilia is a common feature. Local or systemic neutrophilia implies infection. Total serum IgE is frequently elevated, but the demonstration of immunologic specificity for IgE is critical to an etiologic diagnosis. A skin test by the epicutaneous route (scratch or prick) with the allergens of interest provides a rapid and reliable approach to identifying allergen-specific IgE that has sensitized cutaneous mast cells. An intradermal test may follow if indicated by history when the epicutaneous test is negative, but it is less reliable due to the reactivity of some asymptomatic individuals at the test dose. Skin testing by scratch or prick for food allergens is controversial but does seem to have predictive value for the absence of specific IgE sensitivity. A double-blind, placebo-controlled challenge may document a food allergy, but such a procedure does bear the risk of an anaphylactic reaction. An elimination diet is safer but is tedious and less definitive. Food allergy is uncommon as a cause of allergic rhinitis.

Newer methodology for detecting total IgE, including the development of enzyme-linked immunosorbent assays (ELISA) employing anti-IgE bound to either a solid-phase or a liquid-phase paramagnetic particle, provides rapid and cost effective determinations.

Measurements of specific anti-IgE in serum are obtained by its binding to a solid-phase allergen and quantitation by subsequent uptake of radiolabeled anti-IgE. This radioallergosorbent technique correlates with the bioassay of specific IgE by skin test, which is mast cell-dependent, and by histamine release from peripheral blood leukocytes, which is basophil-dependent. As compared to the skin test, the assay of specific IgE in serum is less sensitive but has high specificity. Furthermore, ELISA utilizing reactions that generate visible light or fluorescence have replaced the radioimmunoassays, and newer chemiluminescent tracers provide additional sensitivity for detection of minute quantities of allergen-specific IgE.

PREVENTION

Avoidance of exposure to the offending allergen is the most effective means of controlling allergic diseases; removal of pets from the home to avoid animal danders, utilization of air filtration devices to minimize the concentrations of airborne pollens, elimination of cockroach-derived proteins by chemical destruction of the pest and careful food storage, travel to nonpollinating areas during the critical periods, and even a change of domicile to eliminate a mold spore problem may be necessary. Control of dust mites by allergen avoidance includes use of plastic-lined covers for mattresses, pillows, and comforters, and elimination of carpets and drapes.

TREATMENT

Management with pharmacologic agents represents the standard approach to seasonal or perennial allergic rhinitis. Antihistamines of the H_1 class are effective for nasopharyngeal itching, sneezing, and watery rhinorrhea and for such ocular manifestations as itching, tearing, and erythema, but they are not efficacious for the nasal congestion. The older antihistamines are sedating, and their anticholinergic (muscarinic) effects include visual disturbance, urinary retention, and even arrhythmias. Because the newer H_1antihistaminics such as loratadine and cetirizine are less lipophilic, their ability to cross the blood-brain barrier is reduced, and thus their sedating and anticholinergic side effects are minimized. Because life-threatening ventricular arrhythmias with some fatalities have been caused by prolongation of the QT interval resulting from inhibition of the metabolism of terfenadine and astemizole by their interactions with macrolide antibiotics, these agents have been subtantially replaced by loratadine, fexofenadine, and cetirizine.a-Adrenergic agents such as phenylephrine or oximetazoline are generally used topically to alleviate nasal congestion and obstruction, but the duration of efficacy is limited because of rebound rhinitis and such systemic responses as insomnia, irritability, and hypertension. The latter are more frequent with use of oral a-adrenergic agonists, which nonetheless are useful in relieving nasal congestion and diminishing the sedating effects of conventional antihistamines. Cromolyn sodium, a nasal spray, is essentially without side effects and is used prophylactically on a continuous basis during the season to attenuate allergen activation of nasal mast cells. The clinical efficacy of cromolyn sodium and that of nonsedating antihistamines are roughly equivalent. Intranasal high-potency glucocorticoids are the most potent drugs available for the relief of established rhinitis, seasonal or perennial, and even vasomotor rhinitis; they provide efficacy with substantially reduced side effects as compared with this same class of agent administered orally. Their most frequent side effect is local irritation, with *Candida* overgrowth being a rare occurrence. The topical-to-systemic activity of flunisolide or budesonide is significantly greater than for beclomethasone or triamcinolone with much less systemic absorption. For patients who do not benefit adequately from a full dosage of a nonsedating H_1antihistamine and a maintenance dosage of cromolyn sodium, ana-adrenergic agent for short-term relief should be replaced by high-potency topical glucocorticoids. For systemic symptoms not related to the nasopharynx, such as allergic conjunctivitis, treatment may be local.

Immunotherapy, often termed *hyposensitization*, consists of repeated subcutaneous injections of gradually increasing concentrations of the allergen(s) considered to be specifically responsible for the symptom complex. Controlled studies of ragweed, grass, dust mite, and cat dander allergens administered for treatment of allergic rhinitis have demonstrated at least partial relief of symptoms and signs. The duration of such immunotherapy is 3 to 5 years, with discontinuation being based on minimal symptoms over two consecutive seasons of exposure. Clinical benefit appears related to the administration of a high dose of allergen at weekly or biweekly intervals. Patients should remain at the treatment site for at least 20 min after allergen administration so that any anaphylactic consequence can be managed. Local reactions with erythema and induration are not uncommon and may persist for 1 to 3 days. Immunotherapy is contraindicated in patients with significant cardiovascular disease or unstable asthma and should be conducted with particular caution in any patient requiringb-adrenergic blocking therapy because of the difficulty in managing an anaphylactic complication. The immunologic characteristics of a response include a rise in antibodies of the IgG class, a small increase in specific IgE early in the treatment course followed by a plateau or decline, and a decline in the percentage of histamine released from peripheral blood basophilic leukocytes challenged with a fixed concentration of the allergen. The antibodies of the IgG class might well reduce or neutralize the quantity of allergen available for interaction with the tissue mast cells but, more important, could modify the seasonal booster response in specific IgE synthesis. None of the individual parameters of the response to immunotherapy correlates well with the assessments of clinical efficacy, suggesting that benefit is derived from a complex of effects that likely includes a reduction in T cell cytokine production. Immunotherapy should be reserved for clearly documented seasonal or perennial rhinitis, clinically related to defined allergen exposure with confirmation by the presence of allergen-specific IgE, which has failed management by allergen avoidance and pharmacotherapy due to lack of efficacy or side effects. A sequence for the management of allergic or perennial rhinitis based on an allergen-specific diagnosis and stepwise management as required for symptom control would include the following: (1) identification of the offending allergen(s) by history with confirmation of the presence of allergen-specific IgE by skin test (epicutaneous) and/or serum assay; (2) avoidance of the offending allergen; (3) for mild symptoms, prophylactic management with topical cromolyn sodium or treatment with a single bedtime dose of chlorpheniramine (if the latter is associated with undue side effects, substitute a second-generation nonsedating antihistimine); combination with an oral decongestant such as pseudoephedrine can be beneficial; (4) for prominent symptoms, utilization of topical beclomethasone, budesonide, fluticasone, momestasone, or triamcinalone may be needed for a satisfactory clinical outcome; and (5) for management failures despite avoidance and pharmacotherapy, progression to immunotherapy.

(Bibliography omitted in Palm version)

Back to Table of Contents

311. SYSTEMIC LUPUS ERYTHEMATOSUS - *Bevra Hannahs Hahn*

DEFINITION AND PREVALENCE

Systemic lupus erythematosus (SLE) is a disease of unknown etiology in which tissues and cells are damaged by pathogenic autoantibodies and immune complexes. Ninety percent of cases are in women, usually of child-bearing age, but children, men, and the elderly can be affected. In the United States, the prevalence of SLE in urban areas varies from 15 to 50 per 100,000 population; it is more common in blacks than in whites. Hispanic and Asian populations are also susceptible.

PATHOGENESIS AND ETIOLOGY

SLEresults from tissue damage caused by pathogenic subsets of autoantibodies and immune complexes. The abnormal immune responses include (1) polyclonal and antigen-specific T and B lymphocyte hyperactivity, and (2) inadequate regulation of that hyperactivity. These abnormal immune responses probably depend upon interactions between susceptibility genes and environment. Evidence for genetic predisposition includes increased concordance for disease in monozygotic (24 to 58%) compared with dizygotic (0 to 6%) twins, al$_s$> 10, and a 10 to 15% frequency of patients with more than one affected family member. Studies of association, linkage, and genome scanning show complex genetic susceptibility. Most people with homozygous deficiencies of early components of complement (C1q, C2, C4) have SLE or similar disease (accounting for <5% of SLE patients), suggesting that these genes are major predisposing factors. Most patients must inherit multiple susceptibility genes, and probably experience environmental stimuli as well, to develop clinical disease. A defective or deleted class III allele, C4AQO, is the most common genetic marker associated with SLE in many ethnic groups (40 to 50% of patients compared with 15% of healthy controls). One extended haplotype, B8.DR3.DQw2.C4AQO, predisposes to SLE in populations with Northern European heritage. SLE is associated with HLA-DR2 or -DR3 in many groups, and single-gene associations occur between HLA class II (especially DQ$_b$) and autoantibodies that associate with clinical subsets of lupus. For example, antibodies to Ro/La (SS-A/SS-B) are associated with subacute cutaneous lupus and certain DQA and DQB genes that are usually inherited with DR3. Normal alleles of FcgRIIA or of FcgRIIIA that bind IgG2 or IgG1 and IgG3, respectively, less efficiently than other alleles are associated with SLE, particularly with nephritis. FcgRIIA predisposes to SLE in African Americans and South Koreans; FcgRIIIA predisposes across different ethnic groups. Such alleles might account for impaired clearing of autoantibodies and immune complexes, thus predisposing to their deposition in tissues. Genome scanning from several laboratories has shown two regions of chromosome 1 that link to disease in sibpairs or multiplex families. One region, 1q23, contains the FcgRIIA gene; the other, 1q41-42, contains poly (ADP-ribosyl) polymerase (PARP), which may be another predisposing gene that plays a role in DNA repair and apoptosis. Other results of genome scanning suggest that at least 10 other regions on various chromosomes, in addition to HLA and the two regions on chromosome 1 discussed above, participate in susceptibility. Some genes may be "autoimmunity" genes common to different autoimmune diseases across different ethnic groups; others are likely to be restricted to a single disease and/or a single ethnic group. Family studies suggest that females are more likely than males to express the autoimmune manifestations of their genotypes.

Environmental factors that cause flares ofSLE are largely unknown, with the exception of ultraviolet (UV)-B (and sometimes UV-A) light. As many as 70% of patients are photosensitive. Other factors, such as ingested alfalfa sprouts, and chemicals, such as hydrazines, have been implicated. Searches for viral/retroviral disease inducers have been inconclusive. Although some drugs can induce lupus-like disease, there are notable clinical and autoantibody differences between drug-induced and spontaneous lupus. Femaleness is clearly a susceptibility factor, since the prevalence in women of child-bearing years is seven to nine times higher than in men, whereas the female:male ratio is 3:1 in pre- and postmenopausal years. Metabolism of estrogenic and androgenic hormones may be abnormal in lupus patients. Sex hormones also influence immune tolerance.

Abnormal immune responses permit sustained production of pathogenic subsets of autoantibodies and immune complexes. Some autoantibodies, such as anti-DNA, can bind to tissue via charge or cross-reactivity, or in immune complexes, and cause complement-mediated damage. Some subsets of anti-DNA and anti-RNP can bind and enter living cells, altering their function. Other autoantibodies cause damage by direct binding to cell membranes (erythrocytes, platelets) that cause those cells to be phagocytized and destroyed. T cell help is critical to development of full-blown disease; cells of CD4+CD8-, CD4-CD8+, and CD4-CD8- phenotypes all help autoantibody production inSLE. The abnormalities that permit hyperactivated self-reactive B and T cells to dominate immune repertoires in murine and human SLE are multiple and include defects in cell activation, tolerance, apoptosis, idiotypic networks, immune complex clearance, and generation of regulatory cells. The structure of antigens that stimulate autoantibodies is under investigation. Some are clearly derived from self (nucleosomes, ribonucleoprotein, erythrocyte and lymphocyte surface antigens); others may be from the external environment and mimic self (e.g., components of vesicular stomatitis virus mimic peptides in Sm antigen). Many DNA/protein and RNA/protein antigens may be presented to the immune system in surface blebs of apoptotic cells. SinceUV light induces apoptosis in skin cells, this might be a mechanism for flaring disease. Autoantibodies characteristic of SLE are listed inTable 311-1.

In summary, some individuals are genetically predisposed toSLE. Under the influence of multiple genes, possibly triggered by environmental challenges and highly influenced by sex, they may develop a number of different clinical syndromes that fulfill diagnostic criteria for SLE. The etiology of these syndromes is complex and probably differs between patients.

CLINICAL MANIFESTATIONS

At onset,SLE may involve only one organ system (additional manifestations occur later) or may be multisystemic. Clinical manifestations are listed inTable 311-2; those that fulfill American Rheumatism Association (currently the American College of Rheumatology) updated criteria for a diagnosis of SLE are listed inTable 311-3. Autoantibodies are detectable at disease onset. Severity varies from mild and intermittent to persistent and fulminant. Most patients experience exacerbations interspersed with periods of relative quiescence. True remissions with no symptoms and requiring no therapy occur in up to 20% but are usually not permanent. Systemic symptoms are usually prominent and include fatigue, malaise, fever, anorexia, and weight loss.

Musculoskeletal Manifestations Almost all patients experience arthralgias and myalgias; most develop intermittent arthritis. Pain is often out of proportion to physical findings. Symmetric fusiform swelling in joints [most frequently proximal interphalangeal (PIP) and metacarpophalangeal (MCP) joints of the hands, wrists, and knees], diffuse puffiness of hands and feet, and tenosynovitis can be seen. Joint deformities are unusual, with 10% of patients developing swan-neck deformities of fingers and ulnar drift at MCP joints. Erosions are rare; subcutaneous nodules occur. Myopathy can be inflammatory (during periods of active disease), or secondary to treatment (hypokalemia, glucocorticoid myopathy, hydroxychloroquine myopathy). Ischemic necrosis of bone is a common cause of hip, knee, or shoulder pain in patients receiving glucocorticoids.

Cutaneous Manifestations The malar ("butterfly") rash is a photosensitive, fixed erythematous rash, flat or raised, over the cheeks and bridge of the nose, often involving the chin and ears. Scarring is absent; telangiectases may develop. A more diffuse maculopapular rash, predominant in sun-exposed areas, is also common and usually indicates disease flare. Loss of scalp hair is usually patchy but can be extensive; hair often regrows inSLElesions but not in lesions of discoid lupus erythematosus (DLE). DLE occurs in about 20% of patients with SLE and can be disfiguring, since the lesions have central atrophy and scarring, with permanent loss of appendages. DLE lesions are circular with an erythematous raised rim, scaliness, follicular plugging, and telangiectasia. They occur over the scalp, ears, face, and sun-exposed areas of the arms, back, and chest. Only 5% of patients with DLE subsequently develop SLE. Less frequent SLE skin lesions include urticaria, bullae, erythema multiforme, lichen planus-like lesions, and panniculitis ("lupus profundus").

Patients with subacute cutaneous lupus erythematosus (SCLE) are a distinct subset with recurring extensive dermatitis. Arthritis and fatigue are frequent; central nervous system and renal involvement are not. Some patients are antinuclear antibody (ANA)-negative. Most have antibodies to Ro (SS-A) or to single-stranded (ss) DNA. Skin lesions are photosensitive and either annular or papulosquamous psoriasiform; they occur over the arms, trunk, and face but do not scar.

Patients withSLE,DLE, orSCLE can develop vasculitic skin lesions. These include purpura, subcutaneous nodules, nail fold infarcts, ulcers, vasculitic urticaria, panniculitis, and gangrene of digits. Shallow, slightly painful ulcers in the mouth and nose are frequent in patients with SLE.

Renal Manifestations Most patients withSLE have immunoglobulins deposited in glomeruli, but only one-half have clinical nephritis, defined by proteinuria. Early in the disease most are asymptomatic, although some develop the edema of nephrotic syndrome. Urinalysis shows hematuria, cylindruria, and proteinuria. Most patients with mesangial or mild focal proliferative nephritis (see discussion under "Pathology," below) maintain good renal function. Patients with diffuse proliferative nephritis develop renal failure if untreated. Because severe nephritis requires aggressive immunosuppression with high-dose glucocorticoids and cytotoxic drugs and mild lesions do not, renal biopsy may provide information that affects therapy. Patients with rapidly deteriorating renal function and active urine sediment require prompt, aggressive therapy; biopsy is not necessary unless they fail to respond. However, patients with a slow rise in serum creatinine to levels>265 umol/L (>3 mg/dL) should be biopsied; a high proportion of sclerotic glomeruli on biopsy suggests that these patients are unlikely to respond to immunosuppressive therapies and are candidates for dialysis or transplantation. Patients with persistently abnormal urinalyses, high titers of anti-dsDNA, and/or hypocomplementemia are at risk for severe nephritis; kidney biopsy may guide therapy.

Nervous System Any region of the brain can be involved inSLE, as can the meninges, spinal cord, and cranial and peripheral nerves. Central nervous system (CNS) events may be single or multiple and often occur when SLE is active in other organ systems. Mild cognitive dysfunction is the most frequent manifestation. Headaches are common and may be migraine-like or nonspecific. Seizures of any type may occur. Less frequent manifestations include psychosis, acute confusional states, demyelinating disorders, cerebrovascular disease, movement disorders, aseptic meningitis, myelopathy, mononeuropathy or polyneuropathy of cranial or peripheral nerves, autonomic dysfunction, acute demyelinating polyneuropathy (Guillain-Barre), mood disorders, optic neuritis, subarachnoid hemorrhage, pseudotumor cerebri, and hypothalamic dysfunction with inappropriate secretion of vasopressin. Depression and anxiety are frequent.

Laboratory diagnosis of CNSlupus can be difficult. Abnormal electroencephalograms occur in about 70% of patients with neurologic complaints and usually show diffuse slowing or focal abnormalities. Cerebrospinal fluid (CSF) shows elevated protein levels in 50% and increased mononuclear cells in 30% of patients; oligoclonal bands and increased Ig synthesis may be found. Lumbar puncture is recommended when the diagnosis of CNS lupus is in doubt or when infection is a possible cause of symptoms. Magnetic resonance imaging (MRI) with contrast is the most sensitive radiographic technique to detect acute and chronic lesions ofSLE; changes are often nonspecific. Patients with focal neurologic lesions are more likely to have positive MRI scans than those with diffuse manifestations. Computed tomography (CT) scans are useful to rule out bleeding or mass lesions, if indicated. Angiograms can detect vasculitis and vascular occlusions or emboli; they cannot visualize vessels smaller than 50 um; lupus vasculitis usually involves smaller vessels. Laboratory measures of disease activity often do not correlate with neurologic manifestations. Neurologic problems (with the exception of deficits resulting from large infarcts) usually improve with immunosuppressive therapy and/or time; recurrences are seen in approximately one-third of patients.

Vascular System Thrombosis in vessels of any size can be a major problem. Although vasculitis may underly thrombosis, there is increasing evidence that antibodies against phospholipids [lupus anticoagulant (LA), anticardiolipin (aCL)] are associated with clotting without inflammation. The source of cerebral emboli may be the lesions of Libman-Sacks endocarditis. In addition, degenerative vascular changes after years of exposure of blood vessels to circulating immune complexes and hyperlipidemia from glucocorticoid therapy predispose to degenerative cerebral and coronary artery disease in lupus patients. Therefore, anticoagulation is more appropriate than immunosuppression in some patients.

Hematologic Manifestation Anemia of chronic disease occurs in most patients when lupus is active. Hemolysis occurs in a small proportion of those with positive Coombs' tests; it is usually responsive to high-dose glucocorticoids; resistant cases may respond to splenectomy. Leukopenia (usually lymphopenia) is common but is rarely associated with recurrent infections and does not require treatment. Mild thrombocytopenia is common; severe thrombocytopenia with bleeding and purpura occurs in 5% of patients and should be treated with high-dose glucocorticoids. Short-term improvement can be achieved by administration of intravenous gamma globulin. If the platelet count fails to reach acceptable levels in 2 weeks, addition of cytotoxic drugs, cyclosporine, danazole, and/or splenectomy should be considered.

TheLAbelongs to a family of antiphospholipid antibodies. It is recognized by prolongation of the partial thromboplastin time and failure of added normal plasma to correct the prolongation. More sensitive tests include the Russell viper venom time.aCL are detected in enzyme-linked immunosorbent assays. Clinical manifestations of LA and aCL include thrombocytopenia, recurrent venous or arterial clotting, recurrent fetal loss, and valvular heart disease. If the LA is associated with hypoprothrombinemia or thrombocytopenia, bleeding may occur. Less commonly, antibodies to clotting factors (VIII, IX) arise; they cause bleeding. Bleeding syndromes usually respond to glucocorticoids; clotting syndromes do not.

Cardiopulmonary System Pericarditis is the most frequent manifestation of cardiac lupus; effusions can occur and occasionally lead to tamponade; constrictive pericarditis is rare. Myocarditis can cause arrhythmias, sudden death, and/or heart failure. Valvular insufficiency (usually aortic or mitral) can occur, with or without Libman-Sacks endocarditis. Lesions on valves are best detected by transesophageal echocardiography. Myocardial infarcts usually result from degenerative disease, although they can result from vasculitis.

Pleurisy and pleural effusions are common manifestations ofSLE. Lupus pneumonitis causes fever, dyspnea, and cough; x-rays show fleeting infiltrates and/or areas of platelike atelectasis; this syndrome responds to glucocorticoids. However, *the most common cause of pulmonary infiltrates in patients with SLE is infection*. Interstitial pneumonitis leading to fibrosis occurs occasionally; the inflammatory phase may respond to treatment; the fibrosis does not. Pulmonary hypertension is an uncommon, grave manifestation of SLE. Infrequent pulmonary manifestations with high mortality rates include adult respiratory distress syndrome and massive intraalveolar hemorrhage.

Gastrointestinal System Common gastrointestinal (GI) symptoms include nausea, diarrhea, and vague discomfort. Symptoms may result from lupus peritonitis and may herald a flare ofSLE. Vasculitis of the intestine is the most dangerous manifestation, presenting with acute crampy abdominal pain, vomiting, and diarrhea. Intestinal perforation can occur and usually requires immediate surgery. Patients with pseudoobstruction have abdominal pain; x-rays show dilated loops of small bowel which may be edematous; surgery should be avoided unless frank obstruction is present. Glucocorticoid therapy is useful for all these GI syndromes. Some patients have GI motility disorders similar to those in scleroderma; they are not benefited by steroids. Acute pancreatitis occurs and can be severe, resulting from active SLE or from therapy with glucocorticoids or azathioprine. Elevated amylase levels may reflect pancreatitis, salivary gland inflammation, or macroamylasemia. Elevated serum transaminase levels are common in patients with active SLE but are not associated with significant hepatic damage; they return to normal as the disease is treated.

Ocular Manifestation Retinal vasculitis is a serious manifestation; blindness can develop over a few days, and aggressive immunosuppression should be instituted. Examination shows areas of sheathed, narrow retinal arterioles and cytoid bodies (white exudates) adjacent to vessels. Other ocular abnormalities include conjunctivitis, episcleritis, optic neuritis, and the sicca syndrome.

PATHOLOGY

Cutaneous Lesions Lesions of acuteSLE,DLE, andSCLEshow similar histopathology, with degeneration of the basal layer of the epidermis, disruption of the dermal-epidermal junction (DEJ), and mononuclear infiltrates around vessels and appendages in the upper dermis. In DLE, follicular plugging and hyperkeratosis are prominent. Deposits of Ig and C¢ are seen in the DEJ in 80 to 100% of lesional and 50% of nonlesional skin in active SLE; the proportions are lower during remissions. Only 50% of SCLE lesions are positive for Ig and C¢ deposits. Ig deposition in the DEJ is not specific for SLE. Vasculitic skin lesions usually show leukocytoclastic angiitis.

Renal Lesions Glomerulonephritis (GN) is caused by deposition of circulating immune complexes or in situ complex formation in mesangium and glomerular basement membrane. Renal biopsy should be considered when results would affect therapy. Information regarding location of immune deposits, histologic pattern of renal damage, and activity and chronicity of lesions are all useful in predicting prognosis and selecting appropriate treatment. In mild GN unlikely to lead to renal failure, Ig deposits are confined to the mesangium, and histology shows no changes or mesangial proliferation. If Ig and C¢ are deposited outside the mesangium in capillary glomerular basement membrane, prognosis worsens. Histologic changes that should be treated with aggressive immunosuppression include focal proliferative, membranoproliferative, and diffuse proliferative GN (Chap. 275). Progression from focal to diffuse lesions can occur. Membranous changes without proliferation are uncommon but have a better prognosis than proliferative GN. Activity and chronicity scores indicate severity and reversibility of lesions. *Reversible "active" lesions* associated with high risk of progression to renal failure are glomerular necrosis, cellular epithelial crescents, hyaline thrombi, interstitial inflammatory infiltrates, and necrotizing vasculitis. *Irreversible changes unlikely to respond to immunosuppression* and highly associated with renal failure include glomerular sclerosis, fibrous crescents, interstitial fibrosis, and tubular atrophy. In patients with high chronicity scores, treatment of lupus should be determined by extrarenal disease.

LABORATORY MANIFESTATIONS

The presence of characteristic antibodies (Table 311-1) confirms the diagnosis ofSLE.ANAs are the best screening test. If the test substrate contains human nuclei (WIL-2 or HEP-2 cells), more than 95% of lupus patients will be positive. A positive ANA test is not specific for SLE; ANAs occur in some normal individuals (usually in low titer);

the frequency increases with aging. Other autoimmune diseases, viral infections, chronic inflammatory processes, and several drugs induce ANAs. Therefore, a positive ANA test supports the diagnosis of SLE but is not specific; a negative ANA test makes the diagnosis unlikely but not impossible. Antibodies to double-stranded DNA (dsDNA) and to Sm are relatively specific for SLE; other autoantibodies listed inTable 311-1 are not. However, determining the complete autoantibody profile of each patient helps predict clinical subsets. High serum levels of ANAs and anti-dsDNA and low levels of complement usually reflect disease activity, especially in patients with nephritis. Total functional hemolytic complement (CH_{50}) levels are the most sensitive measure of complement activation but are also most subject to laboratory error. Quantitative levels of C3 and C4 are widely available. Very low levels of $CH_{50}$with normal levels of C3 suggest inherited deficiency of a complement component, which is highly associated with SLE and with ANA negativity.

Hematologic abnormalities include anemia (usually normochromic normocytic but occasionally hemolytic), leukopenia, lymphopenia, and thrombocytopenia. The Westergren erythrocyte sedimentation rate correlates with disease activity in some patients.

Urinalysis should be performed and serum creatinine levels should be measured periodically in patients withSLE. With active nephritis, the urinalysis usually shows proteinuria, hematuria, and cellular or granular casts. Urinary protein excretion measured over 24 h increases during periods of activity. (See the discussion under "Pathology" for a description of renal biopsy.)

PREGNANCY

Fertility rates are normal in patients withSLE, but spontaneous abortion and stillbirths are frequent (10 to 30%), especially in women withLAand/oraCL. The treatment of choice for pregnant women with prior fetal loss and antiphospholipid antibodies is low-dose heparin, e.g., 5000 units subcutaneously twice a day. This may be associated with maternal bone loss. If there are contraindications to heparin therapy, low-dose aspirin or low- to moderate-dose glucocorticoids may be used.

Pregnancy has varied effects onSLEactivity. Disease flares in a small proportion, especially during the 6 weeks postpartum. If severe renal or cardiac disease is absent and SLE activity is controlled, most patients complete pregnancy safely and deliver normal infants. Glucocorticoids (except dexamethasone and betamethasone) are inactivated by placental enzymes and do not cause fetal abnormalities in humans; they should be used to suppress disease activity. Neonatal lupus, caused by transmission of maternal anti-Ro across the placenta, consists of transient skin rash and (rarely) permanent heart block. Transient thrombocytopenia from maternal antiplatelet antibodies also occurs.

DIFFERENTIAL DIAGNOSIS

The American Rheumatism Association published diagnostic criteria forSLE(Table 311-3) which were updated in 1997. Any four of the manifestations listed establish a diagnosis of SLE. Early disease confined to a few systems is more difficult to classify; it may take several years for a patient to fulfill criteria. Disorders with which SLE can be confused include rheumatoid arthritis; various forms of dermatitis; neurologic disorders such as epilepsy, multiple sclerosis, and psychiatric disorders; and hematologic diseases such as idiopathic thrombocytopenic purpura. Many autoimmune disorders have overlapping features so that exact classification may be difficult. Mixed connective tissue disease has features of SLE, rheumatoid arthritis, polymyositis, and scleroderma, accompanied by high titers of anti-ribonucleoprotein antibodies (Chap. 313); patients have a low incidence of nephritis and CNSdisease and a high incidence of pulmonary manifestations and evolution into scleroderma. The possibility of drug-induced lupus should always be considered.Figure 311-1 presents an algorithm for diagnosis of SLE.

DRUG-INDUCED LUPUS

Several drugs can cause a syndrome resemblingSLE, including procainamide, hydralazine, isoniazid, chlorpromazine, D-penicillamine, practolol, methyldopa, quinidine, interferon a, and possibly hydantoin, ethosuximide, and oral contraceptives. The syndrome is rare with all but procainamide, the most frequent offender, and hydralazine. There is genetic predisposition to drug-induced lupus, partly determined by drug acetylation rates. Procainamide inducesANA in 50 to 75% of individuals within a few months; hydralazine induces ANA in 25 to 30%. Between 10 and 20% of ANA-positive individuals develop lupus-like symptoms. Most common are systemic complaints and arthralgias; polyarthritis and pleuropericarditis occur in 25 to 50%. Renal andCNSinvolvement are rare. All patients have ANA and most have antibodies to histones. Antibodies to dsDNA and hypocomplementemia are rare -- a helpful point in distinguishing drug-induced from idiopathic lupus. Anemia, leukopenia,LA,aCL, thrombocytopenia, cryoglobulins, rheumatoid factors, false-positive VDRL, and positive direct Coombs' tests can occur. The initial therapeutic approach is withdrawal of the offending drug; most patients improve in a few weeks. If symptoms are severe, a short course (2 to 10 weeks) of glucocorticoids is indicated. Symptoms rarely persist more than 6 months; ANA may persist for years. Most lupus-inducing drugs can be used safely in patients with idiopathic SLE.

PROGNOSIS

Survival in patients withSLE is 90 to 95% at 2 years, 82 to 90% at 5 years, 71 to 80% at 10 years, and 63 to 75% at 20 years. The following factors have been associated with poor prognosis (approximately 50% mortality in 10 years): high serum creatinine levels [>124 umol/L(>1.4 mg/dL)], hypertension, nephrotic syndrome (24-h urine protein excretion>2.6 g), anemia [hemoglobin< 124 g/L(<12.4g/dL)], hypoalbuminemia and hypocomplementemia at the time of diagnosis, and low socioeconomic status. Other factors associated with a poor prognosis in most studies include thrombocytopenia, seriousCNSinvolvement, antibodies to phospholipids, and African American race. Disability in SLE patients is common. However, approximately 20% of patients experience disease remissions (usually transient), and the likelihood of remission increases with each decade after diagnosis. Infections and active SLE, especially renal failure, are the leading causes of death in the first decade of disease. Thromboembolic events are frequent causes of death in the second decade.

TREATMENT

There is no cure forSLE. Complete sustained remissions are rare. Therefore, patient and physician should plan to control acute, severe flares and to develop maintenance strategies in which symptoms are suppressed to an acceptable level, usually at the cost of some drug side effects. Approximately 25% of SLE patients have mild disease with no life-threatening manifestations, although pain and fatigue may be disabling. These patients should be managed without glucocorticoids. Arthralgias, arthritis, myalgias, fever, and mild serositis may improve on nonsteroidal anti-inflammatory drugs (NSAIDs) including salicylates. However, NSAID toxicities such as elevated serum transaminases, aseptic meningitis, and renal impairment are especially frequent in SLE. The role of NSAIDs, which are primarily COX-2 inhibitors, in treatment of SLE has not been studied; they are likely to be useful. The dermatitides of SLE, fatigue, and lupus arthritis may respond to antimalarials. Doses of 400 mg hydroxychloroquine daily may improve skin lesions in a few weeks. Side effects are uncommon and include retinal toxicity, rash, myopathy, and neuropathy. Regular ophthalmologic examinations should be performed at least annually, since retinal toxicity is related to cumulative dose. Other therapies include sunscreens (an SPF rating³ 15 is recommended), topical or intralesional glucocorticoids, quinacrine, retinoids, and dapsone. Recent studies suggest that daily oral doses of dihydroepiandrosterone may lower disease activity in patients with mild SLE. Systemic glucocorticoids should be reserved for patients with disabling disease unresponsive to these conservative measures.

Life-threatening, severely disabling manifestations ofSLE that are responsive to immunosuppression should be treated with high doses of *glucocorticoids* (1 to 2 mg/kg per day). When disease is active, glucocorticoids should be given in divided doses every 8 to 12 h. After the disease is controlled, therapy should be consolidated to one morning dose; thereafter the daily dose should be tapered as rapidly as clinical disease permits. Ideally, patients should be slowly converted to alternate-day therapy with a single morning dose of short-acting glucocorticoid (prednisone, prednisolone, methylprednisolone) to minimize side effects. However, the disease may flare on the day off steroids, in which case the lowest single daily dose that suppresses disease should be used. Acutely ill lupus patients, including those with proliferativeGN, can be treated with 3 to 5 days of 1000 mg intravenous "pulses" of methylprednisolone, followed by maintenance daily or alternate-day glucocorticoids. Disease flares are probably controlled more rapidly by this approach, but it is unclear whether long-term outcome is changed.

Undesirable effects of chronic glucocorticoid therapy include cushingoid habitus, weight gain, hypertension, infection, capillary fragility, acne, hirsutism, accelerated osteoporosis, ischemic necrosis of bone, cataracts, glaucoma, diabetes mellitus, myopathy, hypokalemia, irregular menses, irritability, insomnia, and psychosis. Prednisone doses of 15 mg daily (or less) given before noon usually do not suppress the hypothalamic-pituitary axis. Some side effects can be minimized; sustained hyperglycemia, hypertension, edema, and hypokalemia should be treated; infections should be identified and treated early; immunizations with influenza and pneumococcal vaccines are safe if disease is stable. To minimize osteoporosis, supplemental calcium (1000 mg daily) should be added in most patients; in those with 24-h urinary calcium excretion<120 mg, vitamin D, 50,000 units one to three times weekly, can be added (monitor for hypercalcemia). Estrogen replacement therapy (ERT) should be considered at menopause. There is debate regarding the ability of oral contraceptives or ERT to cause flares ofSLE in some patients; these therapies should be withheld from patients with a history of thrombosis. Calcitonin and bisphosphonates (alendronate, didronel, or acetonel) are also useful in preventing and treating osteoporosis.

The use of *cytotoxic agents* (azathioprine, chlorambucil, cyclophosphamide, methotrexate, mycophenolate mofetil) inSLE is probably beneficial in controlling active disease, reducing the rate of disease flares, and reducing steroid requirements. Patients with lupus nephritis have significantly less renal failure and better survival if treated with combinations of glucocorticoids plus intravenous cyclophosphamide; azathioprine as the second drug is less beneficial but is also effective in preventing renal failure. Open trials suggest that mycophenolate, and possibly methotrexate, are also effective second drugs and sometimes benefit patients who fail to respond to cyclophosphamide plus glucocorticoids. Undesirable side effects of cytotoxic drugs include bone marrow suppression, increased infection with opportunistic organisms such as herpes zoster, irreversible ovarian failure, hepatotoxicity (azathioprine, methotrexate, and mycophenolate), bladder toxicity (cyclophosphamide), alopecia, and increased risk for malignancy. Azathioprine is the least toxic; recommended doses are 2 to 3 mg/kg per day orally. Cyclophosphamide is the most effective and the most toxic. Intravenous pulse doses (10 to 15 mg/kg) once every 4 weeks have less urinary bladder toxicity than daily oral doses. Cyclophosphamide can also be used in daily oral doses (1.5 to 2.5 mg/kg per day of each). Mycophenolate can be given orally (1 to 2.5 g a day in divided doses) or methotrexate (5 to 20 mg once a week, orally or subcutaneously). After disease activity has been controlled for a few months, tapering of cytotoxic agents and attempts to discontinue them are appropriate.Figure 311-2 presents an algorithm for treatment of SLE.

Some manifestations ofSLE do not respond to immunosuppression, including clotting disorders, some behavioral abnormalities, and end-stageGN. Anticoagulation is the therapy of choice for prevention of clotting; chronic warfarin therapy in relatively high doses (maintaining INR at 2.5 to 3.0) is effective in preventing venous and arterial clotting in patients with antiphospholipid syndromes; the effects of aspirin, ticlopidine, and heparin on arterial thrombosis are unclear. Psychoactive drugs should be used when appropriate. "Pure" membranous GN may not respond to immunosuppression; several weeks of therapy can be tried but should be abandoned if improvement is not obvious. With regard to renal transplantation, patients with SLE have about twice the rate of allograft failure as do patients with renal failure due to other diseases; the 5-year rate of allograft loss is about 50%. However, overall patient survival is good, >90% at 5 years.

Alternatives to therapy with glucocorticoids plus cytotoxic agents for patients who do not respond to or cannot tolerate these regimens include addition of high-dose intravenous pulse therapy with methylprednisolone, which is ultimately converted to daily prednisone, plus cyclophosphamide, and combinations of cytotoxic drugs; there is some evidence for efficacy of high-dose intravenous glucocorticoids plus intravenous cyclophosphamide and of azathioprine plus cyclophosphamide. All of these regimens increase infection rates. Less well studied, but effective in some patients, are plasmapheresis (which must be accompanied by cytotoxic treatment to prevent rebound of undesirable autoantibodies), cyclosporine, and intravenous immunoglobulin.

Experimental therapies in progress include studies of efficacy of inducing tolerance to DNA, interruption of T and B cell second signals with antibodies to CD40L, and immunoablation with high-dose cyclophosphamide with or without autologous stem cell transplantation.

(Bibliography omitted in Palm version)

Back to Table of Contents

312. RHEUMATOID ARTHRITIS - *Peter E. Lipsky*

Rheumatoid arthritis (RA) is a chronic multisystem disease of unknown cause. Although there are a variety of systemic manifestations, the characteristic feature of RA is persistent inflammatory synovitis, usually involving peripheral joints in a symmetric distribution. The potential of the synovial inflammation to cause cartilage destruction and bone erosions and subsequent changes in joint integrity is the hallmark of the disease. Despite its destructive potential, the course of RA can be quite variable. Some patients may experience only a mild oligoarticular illness of brief duration with minimal joint damage, whereas others will have a relentless progressive polyarthritis with marked functional impairment.

EPIDEMIOLOGY AND GENETICS

The prevalence ofRA is approximately 0.8% of the population (range 0.3 to 2.1%); women are affected approximately three times more often than men. The prevalence increases with age, and sex differences diminish in the older age group. RA is seen throughout the world and affects all races. However, the incidence and severity seem to be less in rural sub-Saharan Africa and in Caribbean blacks. The onset is most frequent during the fourth and fifth decades of life, with 80% of all patients developing the disease between the ages of 35 and 50. The incidence of RA is more than six times as great in 60- to 64-year-old women compared to 18- to 29-year-old women.

Family studies indicate a genetic predisposition. For example, severeRA is found at approximately four times the expected rate in first-degree relatives of individuals with disease associated with the presence of the autoantibody, rheumatoid factor; approximately 10% of patients with RA will have an affected first-degree relative. Moreover, monozygotic twins are at least four times more likely to be concordant for RA than dizygotic twins, who have a similar risk of developing RA as nontwin siblings. Only 15 to 20% of monozygotic twins are concordant for RA, however, implying that factors other than genetics play an important etiopathogenic role. Of note, the highest risk for concordance of RA is noted in twins who have two HLA-DRB1 alleles known to be associated with RA. The class II major histocompatibility complex allele HLA-DR4. (DRB1*0401) and related alleles are known to be major genetic risk factors for RA. Early studies showed that as many as 70% of patients with classic or definite RA express HLA-DR4 compared with 28% of control individuals. An association with HLA-DR4 has been noted in many populations, including North American and European whites, Chippewa Indians, Japanese, and native populations in India, Mexico, South America, and southern China. In a number of groups, including Israeli Jews, Asian Indians, and Yakima Indians of North America, however, there is no association between the development of RA and HLA-DR4. In these individuals, there is an association between RA and HLA-DR1 in the former two groups and HLA-Dw16 in the latter. Molecular analysis of HLA-DR antigens has provided insight into these apparently disparate findings. The HLA-DR molecule is composed of two chains, a nonpolymorphic a chain and a highly polymorphic b chain. Allelic variations in the HLA-DR molecule reflect differences in the amino acids of theb chain, with the major amino acid changes occurring in the three hypervariable regions of the molecule. Each of the HLA-DR molecules that is associated with RA has the same or a very similar sequence of amino acids in the third hypervariable region of the b chain of the molecule. Thus the b chains of the HLA-DR molecules associated with RA, including HLA-Dw4 (DRb1*0401), HLA-Dw14 (DRb1*0404), HLA-Dw15 (DRb1*0405), HLA-DR1 (DRb1*0101), and HLA-Dw16 (DRb1*1402), contain the same amino acids at positions 67 through 74, with the exception of a single change of one basic amino acid for another (arginine ® lysine) in position 71 of HLA-Dw4. All other HLA-DR b chains have amino acid changes in this region that alter either their charge or hydrophobicity. These results indicate that a particular amino acid sequence in the third hypervariable region of the HLA-DR molecule is a major genetic element conveying susceptibility to RA, regardless of whether it occurs in HLA-DR4, HLA-Dw16, or HLA-DR1. It has been estimated that the risk of developing RA in a person with HLA-Dw4 (DRb1*0401) or HLA-Dw14 (DRb1*0404) is 1 in 35 and 1 in 20, respectively, whereas the presence of both alleles puts persons at an even greater risk. The lack of association of HLA-DR4 and RA in certain populations is explained by the major member of the DR4 family found in that population. HLA-DR4 is a family of closely related, serologically defined molecules, including HLA-Dw4, -Dw10, -Dw13, and -Dw15. Different members of the HLA-DR family of molecules are found to predominate in different ethnic groups. Thus, in HLA-DR4-positive North American whites, HLA-Dw4 and -Dw14 are the most frequent, whereas HLA-Dw15 is most frequent in Japanese and southern Chinese. Each of these is associated with RA. By contrast, HLA-Dw10, which is not associated with RA and contains nonconservative amino acid changes in positions 70 and 71 of theb chain, is most common in Israeli Jews. Therefore, HLA-DR4 is not associated with RA in this population. In certain groups of patients, there does not appear to be a clear association between HLA-DR4-related epitopes and RA. Thus, nearly 75% of African American RA patients do not have this genetic element. Moreover, there is an association with HLA-DR10 (DRB1*1001) in Spanish and Italian patients, with HLA-DR9 (DRB1*0901) in Chileans, and with HLA-DR3 (DRB1*0301) in Arab populations.

Additional genes in the HLA-D complex may also convey altered susceptibility toRA. Certain HLA-DR alleles, including HLA-DR5 (DRB1*1101), HLA-DR2 (DRB1*1501), HLA-DR3 (DRB1*0301), and HLA-DR7 (DRB1*0701), may protect against the development of RA in that they tend to be found at lower frequency in RA patients than in controls. Moreover, the HLA-DQ alleles, DQB1*0301 and DQB1*0302, that are in linkage disequilibrium with HLA-DR4 and DQB1*0501, have also been associated with RA. This has raised the possibility that HLA-DQ alleles may represent the actual RA susceptibility genes, whereas specific HLA-DR alleles may convey protection. In this model, the complement of HLA-DR and DQ alleles determines RA susceptibility. Disease manifestations have also been associated with HLA phenotype. Thus, early aggressive disease and extraarticular manifestations are more frequent in patients with DRB1*0401 or DRB1*0404, and more slowly progressive disease in those with DRB1*0101. The presence of both DRB1*0401 and DRB1*0404 appears to increase the risk for both aggressive articular and extraarticular disease. It has been estimated that HLA genes contribute only a portion of the genetic susceptibility to RA. Thus genes outside the HLA complex also contribute. These include genes controlling the expression of the antigen receptor on T cells and both immunoglobulin heavy and light chains. Moreover, polymorphisms in the tumor necrosis factor (TNF)a and the interleukin (IL) 10 genes are also associated with RA, as is a region on chromosome 3 (3q13).

Genetic risk factors do not fully account for the incidence ofRA, suggesting that environmental factors also play a role in the etiology of the disease. This is emphasized by epidemiologic studies in Africa that have indicated that climate and urbanization have a major impact on the incidence and severity of RA in groups of similar genetic background.

ETIOLOGY

The cause ofRAremains unknown. It has been suggested that RA might be a manifestation of the response to an infectious agent in a genetically susceptible host. Because of the worldwide distribution of RA, it has been hypothesized that if an infectious agent is involved, the organism must be ubiquitous. A number of possible causative agents have been suggested, including *Mycoplasma*, Epstein-Barr virus (EBV), cytomegalovirus, parvovirus, and rubella virus, but convincing evidence that these or other infectious agents cause RA has not emerged. The process by which an infectious agent might cause chronic inflammatory arthritis with a characteristic distribution also remains a matter of controversy. One possibility is that there is persistent infection of articular structures or retention of microbial products in the synovial tissues which generates a chronic inflammatory response. Alternatively, the microorganism or response to the microorganism might induce an immune response to components of the joint by altering its integrity and revealing antigenic peptides. In this regard, reactivity to type II collagen and heat shock proteins has been demonstrated. Another possibility is that the infecting microorganism might prime the host to cross-reactive determinants expressed within the joint as a result of "molecular mimicry." Recent evidence of similarity between products of certain gram-negative bacteria and EBV and the HLA-DR4 molecule itself has supported this possibility. Finally, products of infecting microorganisms might induce the disease. Recent work has focused on the possible role of "superantigens" produced by a number of microorganisms, including staphylococci, streptococci and *M. arthritidis*. Superantigens are proteins with the capacity to bind to HLA-DR molecules and particular V_b segments of the heterodimeric T cell receptor and stimulate specific T cells expressing the V_b gene products (Chap. 305). The role of superantigens in the etiology of RA remains speculative. Of all the potential environmental triggers, the only one clearly associated with the development of RA is cigarette smoking.

PATHOLOGY AND PATHOGENESIS

Microvascular injury and an increase in the number of synovial lining cells appear to be the earliest lesions in rheumatoid synovitis. The nature of the insult causing this response is not known. Subsequently, an increased number of synovial lining cells is seen along with perivascular infiltration with mononuclear cells. Before the onset of clinical symptoms, the perivascular infiltrate is predominantly composed of myeloid cells, whereas in symptomatic arthritis, T cells can also be found, although their number does not appear to correlate with symptoms. As the process continues, the synovium becomes edematous and protrudes into the joint cavity as villous projections.

Light-microscopic examination discloses a characteristic constellation of features, which include hyperplasia and hypertrophy of the synovial lining cells; focal or segmental vascular changes, including microvascular injury, thrombosis, and neovascularization; edema; and infiltration with mononuclear cells, often collected into aggregates around small blood vessels (Fig. 312-1). The endothelial cells of the rheumatoid synovium have the appearance of high endothelial venules of lymphoid organs and have been altered by cytokine exposure to facilitate entry of cells into tissue. Rheumatoid synovial endothelial cells express increased amounts of various adhesion molecules involved in this process. Although this pathologic picture is typical ofRA, it can also be seen in a variety of other chronic inflammatory arthritides. The mononuclear cell collections are variable in composition and size. The predominant infiltrating cell is the T lymphocyte. CD4+ T cells predominate over CD8+ T cells and are frequently found in close proximity to HLA-DR+ macrophages and dendritic cells. Increased numbers of a separate population of T cells expressing the gd form of the T cell receptor have also been found in the synovium, although they remain a minor population there and their role in RA has not been delineated. The major population of T cells in the rheumatoid synovium is composed of CD4+ memory T cells that form the majority of cells aggregated around postcapillary venules. Scattered throughout the tissue are CD8+ T cells. Both populations express the early activation antigen, CD69. Besides the accumulation of T cells, rheumatoid synovitis is also characterized by the infiltration of variable numbers of B cells and antibody-producing plasma cells. In advanced disease, structures similar to germinal centers of secondary lymphoid organs may be observed in the synovium. Both polyclonal immunoglobulin and the autoantibody rheumatoid factor are produced within the synovial tissue, which leads to the local formation of immune complexes. Finally, the synovial fibroblasts in RA manifest evidence of activation in that they produce a number of enzymes such as collagenase and cathepsins that can degrade components of the articular matrix. These activated fibroblasts are particularly prominent in the lining layer and at the interface with bone and cartilage. Osteoclasts are also prominent at sites of bone erosion.

The rheumatoid synovium is characterized by the presence of a number of secreted products of activated lymphocytes, macrophages, and fibroblasts. The local production of these cytokines and chemokines appears to account for many of the pathologic and clinical manifestations ofRA. These effector molecules include those that are derived from T lymphocytes such as interleukinIL-2, interferon (IFN) g, IL-6, IL-10, granulocyte-macrophage colony stimulating factor (GM-CSF),TNF-a, transforming growth factor b(TGF-b); IL-13, IL-16, and IL-17; those originating from activated myeloid cells, including IL-1, TNF-a, IL-6, IL-8, IL-10, IL-12, GM-CSF, macrophage CSF, platelet-derived growth factor, insulin-like growth factor, and TGF-b; as well as those secreted by other cell types in the synovium, such as fibroblasts and endothelial cells, including IL-1, IL-6, IL-8, GM-CSF, IL-15, IL-16, IL-18, and macrophage CSF. The activity of these chemokines and cytokines appears to account for many of the features of rheumatoid synovitis, including the synovial tissue inflammation, synovial fluid inflammation, synovial proliferation, and cartilage and bone damage, as well as the systemic manifestations of RA. In addition to the production of effector molecules that propagate the inflammatory process, local factors are produced that tend to slow the inflammation, including specific inhibitors of cytokine action and additional cytokines, such as TGF-b, which inhibits many of the features of rheumatoid synovitis including T cell activation and proliferation, B cell differentiation, and migration of cells into the inflammatory site.

These findings have suggested that the propagation ofRA is an immunologically mediated event, although the original initiating stimulus has not been characterized.

One view is that the inflammatory process in the tissue is driven by the CD4+ T cells infiltrating the synovium. Evidence for this includes (1) the predominance of CD4+ T cells in the synovium; (2) the increase in solubleIL-2 receptors, a product of activated T cells, in blood and synovial fluid of patients with active RA; and (3) amelioration of the disease by removal of T cells by thoracic duct drainage or peripheral lymphapheresis or suppression of their function by drugs, such as cyclosporine or nondepleting monoclonal antibodies to CD4. In addition, the association of RA with certain HLA-DR or -DQ alleles, whose only known functions are to shape the repertoire of CD4+ T cells during ontogeny in the thymus and bind and present antigenic peptides to CD4+ T cells in the periphery, strongly implies a role for CD4+ T cells in the pathogenesis of the disease. Finally, patients with established RA who become infected with HIV also have been noted to improve, although this has not been a uniform finding. Within the rheumatoid synovium, the CD4+ T cells differentiate predominantly into Th1-like effector cells producing the proinflammatory cytokineIFN-gand appear to be deficient in differentiation into Th2-like effector cells capable of producing the anti-inflammatory cytokine IL-4. As a result of the ongoing secretion of IFN-g without the regulatory influences of IL-4, macrophages are activated to produce the proinflammatory cytokines IL-1 andTNF-aand also increase expression of HLA molecules. Moreover, T lymphocytes express surface molecules such as CD154 (CD40 ligand) and also produce a variety of cytokines that promote B cell proliferation and differentiation into antibody-forming cells and therefore also may promote local B cell stimulation. The resultant production of immunoglobulin and rheumatoid factor can lead to immune-complex formation with consequent complement activation and exacerbation of the inflammatory process by the production of the anaphylatoxins, C3a and C5a, and the chemotactic factor C5a. The tissue inflammation is reminiscent of delayed type hypersensitivity reactions occurring in response to soluble antigens or microorganisms, although it has become clear that the number of T cells producing cytokines such as IFN-g is less than is found in typical delayed type hypersensitivity reactions, perhaps owing to the large amount of reactive oxygen species produced locally in the synovium that can dampen T cell function. It remains unclear whether the persistent T cell activity represents a response to a persistent exogenous antigen or to altered autoantigens such as collagen, immunoglobulin, or one of the heat shock proteins, or perhaps both. Alternatively, it could represent persistent responsiveness to activated autologous cells such as might occur as a result ofEBVinfection or persistent response to a foreign antigen or superantigen in the synovial tissue. Finally, rheumatoid inflammation could reflect persistent stimulation of T cells by synovial-derived antigens that cross-react with determinants introduced during antecedent exposure to foreign antigens or infectious microorganisms.

Overriding the chronic inflammation in the synovial tissue is an acute inflammatory process in the synovial fluid. The exudative synovial fluid contains more polymorphonuclear leukocytes than mononuclear cells. A number of mechanisms play a role in stimulating the exudation of synovial fluid. Locally produced immune complexes can activate complement and generate anaphylatoxins and chemotactic factors. Local production, by a variety of cells, of chemokines and cytokines with chemotactic activity as well as inflammatory mediators such as leukotriene B_4 and products of complement activation can attract neutrophils. Moreover, many of these same agents can also stimulate the endothelial cells of postcapillary venules to become more efficient at binding circulating cells. The net result is the enhanced migration of polymorphonuclear leukocytes into the synovial site. In addition, vasoactive mediators such as histamine produced by the mast cells that infiltrate the rheumatoid synovium may also facilitate the exudation of inflammatory cells into the synovial fluid. Finally, the vasodilatory effects of locally produced prostaglandin E_2 may also facilitate entry of inflammatory cells into the inflammatory site. Once in the synovial fluid, the polymorphonuclear leukocytes can ingest immune complexes, with the resultant production of reactive oxygen metabolites and other inflammatory mediators, further adding to the inflammatory milieu. Locally produced cytokines and chemokines can additionally stimulate polymorphonuclear leukocytes. The production of large amounts of cyclooxygenase and lipoxygenase pathway products of arachidonic acid metabolism by cells in the synovial fluid and tissue further accentuates the signs and symptoms of inflammation.

The precise mechanism by which bone and cartilage destruction occurs has not been completely resolved. Although the synovial fluid contains a number of enzymes potentially able to degrade cartilage, the majority of destruction occurs in juxtaposition to the inflamed synovium, or pannus, that spreads to cover the articular cartilage. This vascular granulation tissue is composed of proliferating fibroblasts, small blood vessels, and a variable number of mononuclear cells and produces a large amount of degradative enzymes, including collagenase and stromelysin, that may facilitate tissue damage. The cytokinesIL-1 andTNF-aplay an important role by stimulating the cells of the pannus to produce collagenase and other neutral proteases. These same two cytokines also activate chondrocytes in situ, stimulating them to produce proteolytic enzymes that can degrade cartilage locally and also inhibiting synthesis of new matrix molecules. Finally, these two cytokines may contribute to the local demineralization of bone by activating osteoclasts that accumulate at the site of local bone resorption. Prostaglandin E_2produced by fibroblasts and macrophages may also contribute to bone demineralization. The common final pathway of bone erosion is likely to involve the activation of osteoclasts that are present in large numbers at these sites. Systemic manifestations ofRA can be accounted for by release of inflammatory effector molecules from the synovium. These include IL-1, TNF-a, and IL-6, which account for many of the manifestations of active RA, including malaise, fatigue, and elevated levels of serum acute-phase reactants. The importance of TNF-a in producing these manifestations is emphasized by the prompt amelioration of symptoms following administration of a monoclonal antibody to TNF-a or a soluble TNF-a receptor Ig construct to patients with RA. In addition, immune complexes produced within the synovium and entering the circulation may account for other features of the disease, such as systemic vasculitis.

As shown inFig. 312-2, the pathology ofRAevolves over the duration of this chronic disease. The earliest event appears to be a nonspecific inflammatory response initiated by an unknown stimulus and characterized by accumulation of macrophages and other mononuclear cells within the sublining layer of the synovium. The activity of these cells is demonstrated by the increased appearance of macrophage-derived cytokines, includingTNF-a,IL-1b, and IL-6. Subsequently, activation of CD4+ T cells is induced, presumably in response to antigenic peptides displayed by a variety of antigen-presenting cells in the synovial tissue. The activated T cells are capable of producing cytokines, especiallyIFN-g, which amplify and perpetuate the inflammation. The presence of activated T cells expressing CD154 (CD40 ligand) can induce polyclonal B cell stimulation and the local production of rheumatoid factor. The cascade of cytokines produced in the synovium activates a variety of cells in the synovium, bone, and cartilage to produce effector molecules that can cause tissue damage characteristic of chronic inflammation. It is important to emphasize that there is no current way to predict the progress from one stage of inflammation to the next, and once established, each can influence the other. Important features of this model include the following: (1) the major pathologic events vary with time in this chronic disease; (2) the time required to progress from one step to the next may vary in different patients and the events, once established, may persist simultaneously; (3) once established, the major pathogenic events operative in an individual patient may vary at different times; and (4) the process is chronic and reiterative, with successive events stimulating progressive amplification of inflammation. These considerations have important implications with regard to appropriate treatment.

CLINICAL MANIFESTATIONS

Onset Characteristically,RA is a chronic polyarthritis. In approximately two-thirds of patients, it begins insidiously with fatigue, anorexia, generalized weakness, and vague musculoskeletal symptoms until the appearance of synovitis becomes apparent. This prodrome may persist for weeks or months and defy diagnosis. Specific symptoms usually appear gradually as several joints, especially those of the hands, wrists, knees, and feet, become affected in a symmetric fashion. In approximately 10% of individuals, the onset is more acute, with a rapid development of polyarthritis, often accompanied by constitutional symptoms, including fever, lymphadenopathy, and splenomegaly. In approximately one-third of patients, symptoms may initially be confined to one or a few joints. Although the pattern of joint involvement may remain asymmetric in a few patients, a symmetric pattern is more typical.

Signs and Symptoms of Articular Disease Pain, swelling, and tenderness may initially be poorly localized to the joints. Pain in affected joints, aggravated by movement, is the most common manifestation of establishedRA. It corresponds in pattern to the joint involvement but does not always correlate with the degree of apparent inflammation. Generalized stiffness is frequent and is usually greatest after periods of inactivity. Morning stiffness of greater than 1-h duration is an almost invariable feature of inflammatory arthritis and may serve to distinguish it from various noninflammatory joint disorders. Notably, however, the presence of morning stiffness may not reliably distinguish between chronic inflammatory and noninflammatory arthritides, as it is also found frequently in the latter. The majority of patients will experience constitutional symptoms such as weakness, easy fatigability, anorexia, and weight loss. Although fever to 40°C occurs on occasion, temperature elevation in excess of 38°C is unusual and suggests the presence of an intercurrent problem such as infection.

Clinically, synovial inflammation causes swelling, tenderness, and limitation of motion. Warmth is usually evident on examination, especially of large joints such as the knee, but erythema is infrequent. Pain originates predominantly from the joint capsule, which is abundantly supplied with pain fibers and is markedly sensitive to stretching or distention. Joint swelling results from accumulation of synovial fluid, hypertrophy of the synovium, and thickening of the joint capsule. Initially, motion is limited by pain. The inflamed joint is usually held in flexion to maximize joint volume and minimize distention of the capsule. Later, fibrous or bony ankylosis or soft tissue contractures lead to fixed deformities.

Although inflammation can affect any diarthrodial joint,RA most often causes symmetric arthritis with characteristic involvement of certain specific joints such as the proximal interphalangeal and metacarpophalangeal joints. The distal interphalangeal joints are rarely involved. Synovitis of the wrist joints is a nearly uniform feature of RA and may lead to limitation of motion, deformity, and median nerve entrapment (carpal tunnel syndrome). Synovitis of the elbow joint often leads to flexion contractures that may develop early in the disease. The knee joint is commonly involved with synovial hypertrophy, chronic effusion, and frequently ligamentous laxity. Pain and swelling behind the knee may be caused by extension of inflamed synovium into the popliteal space (Baker's cyst). Arthritis in the forefoot, ankles, and subtalar joints can produce severe pain with ambulation as well as a number of deformities. Axial involvement is usually limited to the upper cervical spine. Involvement of the lumbar spine is not seen, and lower back pain cannot be ascribed to rheumatoid inflammation. On occasion, inflammation from the synovial joints and bursae of the upper cervical spine leads to atlantoaxial subluxation. This usually presents as pain in the occiput but on rare occasions may lead to compression of the spinal cord.

With persistent inflammation, a variety of characteristic joint changes develop. These can be attributed to a number of pathologic events, including laxity of supporting soft tissue structures; damage or weakening of ligaments, tendons, and the joint capsule; cartilage degradation; muscle imbalance; and unopposed physical forces associated with the use of affected joints. Characteristic changes of the hand include (1) radial deviation at the wrist with ulnar deviation of the digits, often with palmar subluxation of the proximal phalanges ("Z" deformity); (2) hyperextension of the proximal interphalangeal joints, with compensatory flexion of the distal interphalangeal joints (swan-neck deformity); (3) flexion contracture of the proximal interphalangeal joints and extension of the distal interphalangeal joints (boutonniere deformity); and (4) hyperextension of the first interphalangeal joint and flexion of the first metacarpophalangeal joint with a consequent loss of thumb mobility and pinch. Typical joint changes may also develop in the feet, including eversion at the hindfoot (subtalar joint), plantar subluxation of the metatarsal heads, widening of the forefoot, hallux valgus, and lateral deviation and dorsal subluxation of the toes.

Extraarticular Manifestations RA is a systemic disease with a variety of extraarticular manifestations. Although these occur frequently, not all of them have clinical significance. However, on occasion, they may be the major evidence of disease activity and source of morbidity and require management per se. As a rule, these manifestations occur in individuals with high titers of autoantibodies to the Fc component of immunoglobulin G (rheumatoid factors).

Rheumatoid nodules develop in 20 to 30% of persons withRA. They are usually found on periarticular structures, extensor surfaces, or other areas subjected to mechanical pressure, but they can develop elsewhere, including the pleura and meninges. Common locations include the olecranon bursa, the proximal ulna, the Achilles tendon, and the occiput. Nodules vary in size and consistency and are rarely symptomatic, but on occasion they break down as a result of trauma or become infected. They are found almost invariably in individuals with circulating rheumatoid factor. Histologically, rheumatoid nodules consist of a central zone of necrotic material including collagen fibrils, noncollagenous filaments, and cellular debris; a midzone of palisading macrophages that express HLA-DR antigens; and an outer zone of granulation tissue. Examination of early nodules has suggested that the initial event may be a focal vasculitis. In some patients, treatment with methotrexate can increase the number of nodules dramatically.

Clinical weakness and atrophy of skeletal muscle are common. Muscle atrophy may be evident within weeks of the onset ofRA and is usually most apparent in musculature approximating affected joints. Muscle biopsy may show type II fiber atrophy and muscle fiber necrosis with or without a mononuclear cell infiltrate.

Rheumatoid vasculitis (Chap. 317), which can affect nearly any organ system, is seen in patients with severeRA and high titers of circulating rheumatoid factor. Rheumatoid vasculitis is very uncommon in African Americans. In its most aggressive form, rheumatoid vasculitis can cause polyneuropathy and mononeuritis multiplex, cutaneous ulceration and dermal necrosis, digital gangrene, and visceral infarction. While such widespread vasculitis is very rare, more limited forms are not uncommon, especially in white patients with high titers of rheumatoid factor. Neurovascular disease presenting either as a mild distal sensory neuropathy or as mononeuritis multiplex may be the only sign of vasculitis. Cutaneous vasculitis usually presents as crops of small brown spots in the nail beds, nail folds, and digital pulp. Larger ischemic ulcers, especially in the lower extremity, may also develop. Myocardial infarction secondary to rheumatoid vasculitis has been reported, as has vasculitic involvement of lungs, bowel, liver, spleen, pancreas, lymph nodes, and testes. Renal vasculitis is rare.

Pleuropulmonary manifestations, which are more commonly observed in men, include pleural disease, interstitial fibrosis, pleuropulmonary nodules, pneumonitis, and arteritis. Evidence of pleuritis is found commonly at autopsy, but symptomatic disease during life is infrequent. Typically, the pleural fluid contains very low levels of glucose in the absence of infection. Pleural fluid complement is also low compared with the serum level when these are related to the total protein concentration. Pulmonary fibrosis can produce impairment of the diffusing capacity of the lung. Pulmonary nodules may appear singly or in clusters. When they appear in individuals with pneumoconiosis, a diffuse nodular fibrotic process (Caplan's syndrome) may develop. On occasion, pulmonary nodules may cavitate and produce a pneumothorax or bronchopleural fistula. Rarely, pulmonary hypertension secondary to obliteration of the pulmonary vasculature occurs. In addition to pleuropulmonary disease, upper airway obstruction from cricoarytenoid arthritis or laryngeal nodules may develop.

Clinically apparent heart disease attributed to the rheumatoid process is rare, but evidence of asymptomatic pericarditis is found at autopsy in 50% of cases. Pericardial fluid has a low glucose level and is frequently associated with the occurrence of pleural effusion. Although pericarditis is usually asymptomatic, on rare occasions death has occurred from tamponade. Chronic constrictive pericarditis may also occur.

RAtends to spare the central nervous system directly, although vasculitis can cause peripheral neuropathy. *Neurologic manifestations* may also result from atlantoaxial or midcervical spine subluxations. Nerve entrapment secondary to proliferative synovitis or joint deformities may produce neuropathies of median, ulnar, radial (interosseous branch), or anterior tibial nerves.

The rheumatoid process involves the *eye* in fewer than 1% of patients. Affected individuals usually have long-standing disease and nodules. The two principal manifestations are episcleritis, which is usually mild and transient, and scleritis, which involves the deeper layers of the eye and is a more serious inflammatory condition. Histologically, the lesion is similar to a rheumatoid nodule and may result in thinning and perforation of the globe (scleromalacia perforans). From 15 to 20% of persons withRA may develop Sjogren's syndrome with attendant keratoconjunctivitis sicca.

Felty's syndrome consists of chronicRA, splenomegaly, neutropenia, and, on occasion, anemia and thrombocytopenia. It is most common in individuals with long-standing disease. These patients frequently have high titers of rheumatoid factor, subcutaneous nodules, and other manifestations of systemic rheumatoid disease. Felty's syndrome is very uncommon in African Americans. It may develop after joint inflammation has regressed. Circulating immune complexes are often present, and evidence of complement consumption may be seen. The leukopenia is a selective neutropenia with polymorphonuclear leukocyte counts of <1500 cells per microliter and sometimes <1000 cells per microliter. Bone marrow examination usually reveals moderate hypercellularity with a paucity of mature neutrophils. However, the bone marrow may be normal, hyperactive, or hypoactive; maturation arrest may be seen. Hypersplenism has been proposed as one of the causes of leukopenia, but splenomegaly is not invariably found and splenectomy does not always correct the abnormality. Excessive margination of granulocytes caused by antibodies to these cells, complement activation, or binding of immune complexes may contribute to granulocytopenia. Patients with Felty's syndrome have increased frequency of infections usually associated with neutropenia. The cause of the increased susceptibility to infection is related to the defective function of polymorphonuclear leukocytes as well as the decreased number of cells.

Osteoporosis secondary to rheumatoid involvement is common and may be aggravated by glucocorticoid therapy. Glucocorticoid treatment may cause significant loss of bone mass, especially early in the course of therapy, even when low doses are employed. Osteopenia inRAinvolves both juxtaarticular bone and long bones distant from involved joints. RA is associated with a modest decrease in mean bone mass and a moderate increase in the risk of fracture. Bone mass appears to be adversely affected by functional impairment and active inflammation, especially early in the course of the disease.

RA in the Elderly The incidence ofRAcontinues to increase past age 60. It has been suggested that elderly-onset RA might have a poorer prognosis, as manifested by more persistent disease activity, more frequent radiographically evident deterioration, more frequent systemic involvement, and more rapid functional decline. Aggressive disease is largely restricted to those patients with high titers of rheumatoid factor. By contrast, elderly patients who develop RA without elevated titers of rheumatoid factor (seronegative disease) generally have less severe, often self-limited disease.

LABORATORY FINDINGS

No tests are specific for diagnosingRA. However, rheumatoid factors, which are autoantibodies reactive with the Fc portion of IgG, are found in more than two-thirds of adults with the disease. Widely utilized tests largely detect IgM rheumatoid factors. The presence of rheumatoid factor is not specific for RA. Rheumatoid factor is found in 5% of healthy persons. The frequency of rheumatoid factor in the general population increases with age, and 10 to 20% of individuals over 65 years old have a positive test. In addition, a number of conditions besides RA are associated with the presence of rheumatoid factor. These include systemic lupus erythematosus, Sjogren's syndrome, chronic liver disease, sarcoidosis, interstitial pulmonary fibrosis, infectious mononucleosis, hepatitis B, tuberculosis, leprosy, syphilis, subacute bacterial endocarditis, visceral leishmaniasis, schistosomiasis, and malaria. In addition, rheumatoid factor may appear transiently in normal individuals after vaccination or transfusion and may also be found in relatives of individuals with RA.

The presence of rheumatoid factor does not establish the diagnosis ofRA as the predictive value of the presence of rheumatoid factor in determining a diagnosis of RA is poor. Thus fewer than one-third of unselected patients with a positive test for rheumatoid factor will be found to have RA. Therefore, the rheumatoid factor test is not useful as a screening procedure. However, the presence of rheumatoid factor can be of prognostic significance because patients with high titers tend to have more severe and progressive disease with extraarticular manifestations. Rheumatoid factor is uniformly found in patients with nodules or vasculitis. In summary, a test for the presence of rheumatoid factor can be employed to confirm a diagnosis in individuals with a suggestive clinical presentation and, if present in high titer, to designate patients at risk for severe systemic disease. A number of additional autoantibodies may be found in patients with RA, including antibodies to filaggrin, citrulline, calpastatin, components of the spliceosome (RA-33), and an unknown antigen, Sa. Some of these may be useful in diagnosis in that they may occur early in the disease before rheumatoid factor is present or may be associated with aggressive disease.

Normochromic, normocytic anemia is frequently present in activeRA. It is thought to reflect ineffective erythropoiesis; large stores of iron are found in the bone marrow. In general, anemia and thrombocytosis correlate with disease activity. The white blood cell count is usually normal, but a mild leukocytosis may be present. Leukopenia may also exist without the full-blown picture of Felty's syndrome. Eosinophilia, when present, usually reflects severe systemic disease.

The erythrocyte sedimentation rate is increased in nearly all patients with activeRA. The levels of a variety of other acute-phase reactants including ceruloplasmin and C-reactive protein are also elevated, and generally such elevations correlate with disease activity and the likelihood of progressive joint damage.

Synovial fluid analysis confirms the presence of inflammatory arthritis, although none of the findings is specific. The fluid is usually turbid, with reduced viscosity, increased protein content, and a slightly decreased or normal glucose concentration. The white cell count varies between 5 and 50,000/uL; polymorphonuclear leukocytes predominate. A synovial fluid white blood cell count >2000/uL with more than 75% polymorphonuclear leukocytes is highly characteristic of inflammatory arthritis, although not diagnostic ofRA. Total hemolytic complement, C3, and C4 are markedly diminished in synovial fluid relative to total protein concentration as a result of activation of the classic complement pathway by locally produced immune complexes.

RADIOGRAPHIC EVALUATION

Early in the disease, roentgenograms of the affected joints are usually not helpful in establishing a diagnosis. They reveal only that which is apparent from physical examination, namely, evidence of soft tissue swelling and joint effusion. As the disease progresses, abnormalities become more pronounced, but none of the radiographic findings is diagnostic ofRA. The diagnosis, however, is supported by a characteristic pattern of abnormalities, including the tendency toward symmetric involvement. Juxtaarticular osteopenia may become apparent within weeks of onset. Loss of articular cartilage and bone erosions develop after months of sustained activity. The primary value of radiography is to determine the extent of cartilage destruction and bone erosion produced by the disease, particularly when one is monitoring the impact of therapy with disease-modifying drugs or surgical intervention. Other means of imaging bones and joints, including^{99m}Tc bisphosphonate bone scanning and magnetic resonance imaging, may be capable of detecting early inflammatory changes that are not apparent from standard radiography but are rarely necessary in the routine evaluation of patients with RA.

CLINICAL COURSE AND PROGNOSIS

The course ofRA is quite variable and difficult to predict in an individual patient. Most patients experience persistent but fluctuating disease activity, accompanied by a variable degree of joint abnormalities and functional impairment. After 10 to 12 years, fewer than 20% of patients will have no evidence of disability or joint abnormalities. Within 10 years, approximately 50% of patients will have work disability. A number of features are correlated with a greater likelihood of developing joint abnormalities or disabilities. These include the presence of more than 20 inflamed joints, a markedly elevated erythrocyte sedimentation rate, radiographic evidence of bone erosions, the presence of rheumatoid nodules, high titers of serum rheumatoid factor, the presence of functional disability, persistent inflammation, advanced age at onset, the presence of comorbid conditions, low socioeconomic status or educational level, or the presence of HLA-DRB1*0401 or -DRB*0404. The presence of one or more of these implies the presence of more aggressive disease with a greater likelihood of developing progressive joint abnormalities and disability. Persistent elevation of the erythrocyte sedimentation rate, disability, and pain on longitudinal follow-up are good predictors of work disability. Patients who lack these features have more indolent disease with a slower progression to joint abnormalities and disability. The pattern of disease onset does not appear to predict the development of disabilities. Approximately 15% of patients with RA will have a short-lived inflammatory process that remits without major disability. These individuals tend to lack the aforementioned features associated with more aggressive disease.

Several features of patients withRAappear to have prognostic significance. Remissions of disease activity are most likely to occur during the first year. White females tend to have more persistent synovitis and more progressively erosive disease than males. Persons who present with high titers of rheumatoid factor, C-reactive protein, and haptoglobin also have a worse prognosis, as do individuals with subcutaneous nodules or radiographic evidence of erosions at the time of initial evaluation. Sustained disease activity of more than 1 year's duration portends a poor outcome, and persistent elevation of acute-phase reactants appears to correlate strongly with radiographic progression. A large proportion of inflamed joints manifest erosions within 2 years, whereas the subsequent course of erosions is highly variable; however, in general, radiographic damage appears to progress at a constant rate in patients with RA. Foot joints are affected more frequently than hand joints. Despite the decrease in the rate of progressive joint damage with time, functional disability, which develops early in the course of the disease, continues to worsen at the same rate, although the most rapid rate of functional loss occurs within the first 2 years of disease.

The median life expectancy of persons with RAis shortened by 3 to 7 years. Of the 2.5-fold increase in mortality rate, RA itself is a contributing feature in 15 to 30%. The increased mortality rate seems to be limited to patients with more severe articular disease and can be attributed largely to infection and gastrointestinal bleeding. Drug therapy may also play a role in the increased mortality rate seen in these individuals. Factors correlated with early death include disability, disease duration or severity, glucocorticoid use, age at onset, and low socioeconomic or educational status.

DIAGNOSIS

The mean delay from disease onset to diagnosis is 9 months. This is often related to the nonspecific nature of initial symptoms. The diagnosis ofRA is easily made in persons with typical established disease. In a majority of patients, the disease assumes its characteristic clinical features within 1 to 2 years of onset. The typical picture of bilateral symmetric inflammatory polyarthritis involving small and large joints in both the upper and lower extremities with sparing of the axial skeleton except the cervical spine suggests the diagnosis. Constitutional features indicative of the inflammatory nature of the disease, such as morning stiffness, support the diagnosis. Demonstration of subcutaneous nodules is a helpful diagnostic feature. Additionally, the presence of rheumatoid factor, inflammatory synovial fluid with increased numbers of polymorphonuclear leukocytes, and radiographic findings of juxtaarticular bone demineralization and erosions of the affected joints substantiate the diagnosis.

The diagnosis is somewhat more difficult early in the course when only constitutional symptoms or intermittent arthralgias or arthritis in an asymmetric distribution may be present. A period of observation may be necessary before the diagnosis can be established. A definitive diagnosis ofRAdepends predominantly on characteristic clinical features and the exclusion of other inflammatory processes. The isolated finding of a positive test for rheumatoid factor or an elevated erythrocyte sedimentation rate, especially in an older person with joint pains, should not itself be used as evidence of RA.

In 1987, the American College of Rheumatology developed revised criteria for the classification ofRA(Table 312-1). These criteria demonstrate a sensitivity of 91 to 94% and a specificity of 89% when used to classify patients with RA compared with control subjects with rheumatic diseases other than RA. Although these criteria were developed as a means of disease classification for epidemiologic purposes, they can be useful as guidelines for establishing the diagnosis. Failure to meet these criteria, however, especially during the early stages of the disease, does not exclude the diagnosis. Moreover, in patients with early arthritis, the criteria do not discriminate effectively between patients who subsequently develop persistent, disabling, or erosive disease and those who do not.

TREATMENT

General Principles The goals of therapy ofRA are (1) relief of pain, (2) reduction of inflammation, (3) protection of articular structures, (4) maintenance of function, and (5) control of systemic involvement. Since the etiology of RA is unknown, the pathogenesis is not completely delineated, and the mechanisms of action of many of the therapeutic agents employed are uncertain, therapy remains largely empirical. None of the therapeutic interventions is curative, and therefore all must be viewed as palliative, aimed at relieving the signs and symptoms of the disease. The various therapies employed are directed at nonspecific suppression of the inflammatory or immunologic process in the hope of ameliorating symptoms and preventing progressive damage to articular structures.

Management of patients withRAinvolves an interdisciplinary approach, which attempts to deal with the various problems that these individuals encounter with functional as well as psychosocial interactions. A variety of physical therapy modalities may be useful in decreasing the symptoms of RA. Rest ameliorates symptoms and can be an important component of the total therapeutic program. In addition, splinting to reduce unwanted motion of inflamed joints may be useful. Exercise directed at maintaining muscle strength and joint mobility without exacerbating joint inflammation is also an important aspect of the therapeutic regimen. A variety of orthotic and assistive devices can be helpful in supporting and aligning deformed joints to reduce pain and improve function. Education of the patient and family is an important component of the therapeutic plan to help those involved become aware of the potential impact of the disease and make appropriate accommodations in life-style to maximize satisfaction and minimize stress on joints.

Medical management ofRAinvolves five general approaches. The first is the use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) and simple analgesics to control the symptoms and signs of the local inflammatory process. These agents are rapidly effective at mitigating signs and symptoms, but they appear to exert minimal effect on the progression of the disease. Recently, specific inhibitors of the isoform of cyclooxygenase (Cox) that is upregulated at inflammatory sites (Cox-2) have been developed. Cox-2-specific inhibitors (CSIs) have been shown to be as effective as classic NSAIDs, which inhibit both isoforms of Cox, but to cause significantly less gastroduodenal ulceration. The second line of therapy involves use of low-dose oral glucocorticoids. Although low-dose glucocorticoids have been widely used to suppress signs and symptoms of inflammation, recent evidence suggests that they may also retard the development and progression of bone erosions. Intraarticular glucocorticoids can often provide transient symptomatic relief when systemic medical therapy has failed to resolve inflammation. The third line of agents includes a variety of agents that have been classified as the disease-modifying or slow-acting antirheumatic drugs. These agents appear to have the capacity to decrease elevated levels of acute-phase reactants in treated patients and, therefore, are thought to modify the inflammatory component of RA and thus its destructive capacity. Recently, combinations of disease-modifying antirheumatic drugs (DMARDs) have shown promise in controlling the signs and symptoms of RA. A fourth group of agents are theTNF-aneutralizing agents, which have been shown to have a major impact on the signs and symptoms of RA. A fifth group of agents are the immunosuppressive and cytotoxic drugs that have been shown to ameliorate the disease process in some patients. Additional approaches have been employed in an attempt to control the signs and symptoms of RA. Substituting omega-3 fatty acids such as eicosapentaenoic acid found in certain fish oils for dietary omega-6 essential fatty acids found in meat has also been shown to provide symptomatic improvement in patients with RA. A variety of nontraditional approaches have also been claimed to be effective in treating RA, including diets, plant and animal extracts, vaccines, hormones, and topical preparations of various sorts. Many of these are costly, and none has been shown to be effective. However, belief in their efficacy ensures their continued use by some patients.

Drugs

Nonsteroidal Anti-Inflammatory Drugs Besides aspirin, manyNSAIDsare available to treatRA. As a result of the capacity of these agents to block the activity of theCoxenzymes and therefore the production of prostaglandins, prostacyclin, and thromboxanes, they have analgesic, anti-inflammatory, and antipyretic properties. In addition, the agents may exert other anti-inflammatory effects. These agents are all associated with a wide spectrum of toxic side effects. Some, such as gastric irritation, azotemia, platelet dysfunction, and exacerbation of allergic rhinitis and asthma, are related to the inhibition of cyclooxygenase activity, whereas a variety of others, such as rash, liver function abnormalities, and bone marrow depression, may not be. None of the NSAIDs has been shown to be more effective than aspirin in the treatment of RA. However, these nonaspirin drugs are associated with a lower incidence of gastrointestinal intolerance. None of the newer NSAIDs appears to show significant therapeutic advantages over the other available agents. In addition, there is no consistent advantage of any of these newer agents over the others with respect to the incidence or severity of toxic manifestations. Recent evidence indicates that two separate enzymes, Cox-1 and -2, are responsible for the initial metabolism of arachidonic acid into various inflammatory mediators. The former is constitutively present in many cells and tissues, including the stomach and the platelet, whereas the latter is specifically induced in response to inflammatory stimuli. Inhibition of Cox-2 accounts for the anti-inflammatory effects of NSAIDs, whereas inhibition of Cox-1 induces much of the mechanism-based toxicity. As the currently available NSAIDs inhibit both enzymes, therapeutic benefit and toxicity are intertwined.CSIshave now been approved for the treatment of RA. Clinical trials have shown that CSIs suppress the signs and symptoms of RA as effectively as classic Cox-nonspecific NSAIDs but are associated with a significantly reduced incidence of gastroduodenal ulceration. This suggests that CSIs might be considered instead of classic Cox-nonspecific NSAIDs, especially in persons with increased risk of NSAID-induced major upper gastrointestinal side effects, including persons over 65, those with a history of peptic ulcer disease, persons receiving glucocorticoids or anticoagulants, or those requiring high doses of NSAIDs.

Disease-Modifying Antirheumatic Drugs Clinical experience has delineated a number of agents that appear to have the capacity to alter the course ofRA. This group of agents includes methotrexate, gold compounds, D-penicillamine, the antimalarials, and sulfasalazine. Despite having no chemical or pharmacologic similarities, in practice these agents share a number of characteristics. They exert minimal direct nonspecific anti-inflammatory or analgesic effects, and thereforeNSAIDsmust be continued during their administration, except in a few cases when true remissions are induced with them. The appearance of benefit fromDMARDtherapy is usually delayed for weeks or months. As many as two-thirds of patients develop some clinical improvement as a result of therapy with any of these agents, although the induction of true remissions is unusual. In addition to clinical improvement, there is frequently an improvement in serologic evidence of disease activity, and titers of rheumatoid factor and C-reactive protein and the erythrocyte sedimentation rate frequently decline. Moreover, emerging evidence suggests that DMARDs actually retard the development of bone erosions or facilitate their healing. Furthermore, developing evidence suggests that early aggressive treatment with DMARDs may be effective at slowing the appearance of bone erosions.

WhichDMARDshould be the drug of first choice remains controversial, and trials have failed to demonstrate a consistent advantage of one over the other. Despite this, methotrexate has emerged as the DMARD of choice because of its relatively rapid onset of action, its capacity to effect sustained improvement with ongoing therapy, and the high level of patient retention on therapy. Each of the DMARDs is associated with considerable toxicity, and therefore careful patient monitoring is necessary. Toxicity of the various agents also becomes important in determining the drug of first choice. Of note, failure to respond or development of toxicity to one DMARD does not preclude responsiveness to another. Thus, a similar percentage ofRApatients who have failed to respond to one DMARD will respond to another when it is given as the second disease-modifying drug.

No characteristic features of patients have emerged that predict responsiveness to aDMARD. Moreover, the indications for the initiation of therapy with one of these agents are not well defined, although recently the trend has been to begin DMARD therapy early in the course of the disease, and data have begun to emerge to support the conclusion that this approach may slow the development of bone erosions, although this remains controversial.

The folic acid antagonist methotrexate, given in an intermittent low dose (7.5 to 30 mg once weekly), is currently a frequently utilizedDMARD. Most rheumatologists recommend use of methotrexate as the initial DMARD, especially in individuals with evidence of aggressiveRA. Recent trials have documented the efficacy of methotrexate and have indicated that its onset of action is more rapid than other DMARDs, and patients tend to remain on therapy with methotrexate longer than they remain on other DMARDs because of better clinical responses and less toxicity. Long-term trials have indicated that methotrexate does not induce remission but rather suppresses symptoms while it is being administered. Maximal improvement is observed after 6 months of therapy, with little additional improvement thereafter. Major toxicity includes gastrointestinal upset, oral ulceration, and liver function abnormalities that appear to be dose-related and reversible and hepatic fibrosis that can be quite insidious, requiring liver biopsy for detection in its early stages. Drug-induced pneumonitis has also been reported. Liver biopsy is recommended for individuals with persistent or repetitive liver function abnormalities. Concurrent administration of folic acid or folinic acid may diminish the frequency of some side effects without diminishing effectiveness.

Glucocorticoid Therapy Systemic glucocorticoid therapy can provide effective symptomatic therapy in patients withRA. Low-dose (<7.5 mg/d) prednisone has been advocated as useful additive therapy to control symptoms. Moreover, recent evidence suggests that low-dose glucocorticoid therapy may retard the progression of bone erosions. Monthly pulses with high-dose glucocorticoids may be useful in some patients and may hasten the response when therapy with aDMARD is initiated.

TNF-aneutralizing agents Recently, biologic agents that bind and neutralizeTNF-ahave become available. One of these is a TNF-a type II receptor fused to IgG1 (etanercept), and the second is a chimeric mouse/human monoclonal antibody to TNF-a (infliximab). Clinical trials have shown that parenteral administration of either TNF-a neutralizing agent is remarkably effective at controlling signs and symptoms ofRA in patients who have failedDMARDtherapy. Repetitive therapy with these agents is effective with or without concomitant methotrexate. Although these agents are notably effective in persistently controlling signs and symptoms of RA in a majority of patients, their impact on the progression of bone erosions has not been proven. Side effects include the potential for an increased risk of serious infections and the development of anti-DNA antibodies, but with no associated evidence of signs and symptoms of systemic lupus erythematosus. Although these side effects are uncommon, their occurrence mandates that TNF-aneutralizing therapy be supervised by physicians with experience in their use.

Immunosuppressive Therapy The immunosuppressive drugs azathioprine, leflunomide, cyclosporine, and cyclophosphamide have been shown to be effective in the treatment ofRA and to exert therapeutic effects similar to those of theDMARDs. However, these agents appear to be no more effective than the DMARDs. Moreover, they cause a variety of toxic side effects, and cyclophosphamide appears to predispose the patient to the development of malignant neoplasms. Therefore, these drugs have been reserved for patients who have clearly failed therapy with DMARDs. On occasion, extraarticular disease such as rheumatoid vasculitis may require cytotoxic immunosuppressive therapy.

Surgery Surgery plays a role in the management of patients with severely damaged joints. Although arthroplasties and total joint replacements can be done on a number of joints, the most successful procedures are carried out on hips, knees, and shoulders. Realistic goals of these procedures are relief of pain and reduction of disability. Reconstructive hand surgery may lead to cosmetic improvement and some functional benefit. Open or arthroscopic synovectomy may be useful in some patients with persistent monarthritis, especially of the knee. Although synovectomy may offer short-term relief of symptoms, it does not appear to retard bone destruction or alter the natural history of the disease. In addition, early tenosynovectomy of the wrist may prevent tendon rupture.

Approach to the Patient

An approach to the medical management of patients withRA is depicted in Fig. 312-3.

The principles underlying care of these patients reflect the variability of the disease, the frequent persistent nature of the inflammation and its potential to cause disability, the relationship between sustained inflammation and bone erosions, and the need to reevaluate the patient frequently for symptomatic response to therapy, progression of disability and joint damage, and side effects of treatment. At the onset of disease it is difficult to predict the natural history of an individual patient's illness. Therefore, the usual approach is to attempt to alleviate the patient's symptoms withNSAIDsor CSIs. Some patients may have mild disease that requires no additional therapy.

At some time during most patients' course, the possibility of initiatingDMARDtherapy and/or low-dose oral glucocorticoids is entertained. With aggressive disease this might occur sooner, often within 1 to 3 months of diagnosis, whereas in patients with more indolent disease, smoldering activity may not require such therapy for many years. The development of bone erosions or radiographic evidence of cartilage loss is clear-cut evidence of the destructive potential of the inflammatory process and indicates the need for DMARD therapy. The other indications as outlined above, including persistent pain, joint swelling, or functional impairment, are much more subjective, however. As persistent inflammation, involvement of multiple joints, elevated levels of acute-phase reactants, and rheumatoid factor titers correlate with the development of disability and/or bony erosions, some have advocated the use of these prognostic indicators of aggressive disease in the decision to employ DMARDs early in the course ofRA. The decision to begin use of a DMARD and/or low-dose oral glucocorticoids requires experience and clinical judgment as well as the ability to assess joint swelling and functional activity and the patient's pain tolerance and expectation of therapy accurately. In this setting, the fully informed patient must play an active role in the decision to begin DMARD and/or low-dose oral glucocorticoid therapy, after careful review of the therapeutic and toxic potential of the various drugs.

If a patient responds to aDMARD, therapy is continued with careful monitoring to avoid toxicity. All DMARDs provide a suppressive effect and therefore require prolonged administration. Even with successful therapy, local injection of glucocorticoids may be necessary to diminish inflammation that may persist in a limited number of joints. In addition,NSAIDsor CSIsmay be necessary to mitigate symptoms. Even after inflammation has totally resolved, symptoms from loss of cartilage and supervening degenerative joint disease or altered joint function may require additional treatment. Surgery may also be necessary to relieve pain or diminish the functional impairment secondary to alterations in joint function. Recently an alternative approach to treat patients withRA has been suggested. This involves the initiation of therapy with multiple agents early in the course of disease in an attempt to control inflammation, followed by maintenance on one or more agents as necessary to control disease activity. The effectiveness of this therapeutic alternative has not been proved.

(Bibliography omitted in Palm version)

Back to Table of Contents

313. SYSTEMIC SCLEROSIS (SCLERODERMA) - *Bruce C. Gilliland*

DEFINITION

Systemic sclerosis (SSc) is a chronic multisystem disorder of unknown etiology characterized clinically by thickening of the skin caused by accumulation of connective tissue and by involvement of visceral organs, including the gastrointestinal tract, lungs, heart, and kidneys. Classification of SSc and sclerodermal-like disorders is shown in (Table 313-1). Vascular abnormalities, especially of the microvasculature are a prominent feature of SSc. The degree and rate of skin and internal organ involvement vary among patients. Two subsets, however, can be identified, even though there is some overlap (Table 313-2). One subset is referred to as *diffuse cutaneous scleroderma* and is characterized by the rapid development of symmetric skin thickening of proximal and distal extremities, face, and trunk. These patients are at greater risk for developing kidney and other visceral disease early in their course. The other subset is *limited cutaneous scleroderma*, which is defined by symmetric skin thickening limited to distal extremities and face. This subset frequently has features of the *CREST syndrome*, standing for *c*alcinosis, *R*aynaud's phenomenon, *e*sophageal dysmotility, *s*clerodactyly, and *t*elangiectasia. The prognosis in limited cutaneous scleroderma is better except for those patients who, after many years, develop pulmonary arterial hypertension or biliary cirrhosis. Involvement of visceral organs may also occur in the absence of any skin involvement, which is referred to as *systemic sclerosis sine scleroderma*. Survival is determined by the severity of visceral disease, especially involving the lungs, heart, and/or kidneys.

Preliminary criteria for the classification of systemic sclerosis were developed by the American Rheumatism Association (now called the American College of Rheumatology) for the purpose of uniformity in clinical studies. A major criterion was the presence of sclerodermatous skin changes of the fingers of both hands plus involvement at any location proximal to the metacarpal phalangeal joints, entire extremity, face, neck, chest, and abdomen. Minor criteria were sclerodactyly, digital pitting scars or tissue loss of the volar pads of the fingertips, and bibasilar pulmonary fibrosis. The diagnosis ofSSc was based on the presence of the major criterion or two or more minor criteria. The sensitivity of these criteria was 97%, and the specificity 98%. These criteria are not, however, applicable to clinical practice as many patients have SSc who do not meet these criteria. Scleroderma can also occur in a localized form limited to the skin, subcutaneous tissue, and muscle and without systemic involvement. Localized scleroderma occurs most often in children and young women but can affect any age group. The two localized forms are *morphea*, which occurs as single or multiple plaques of skin induration, and *linear scleroderma*, which involves an extremity or face. Linear scleroderma of one side of the forehead and scalp produces a disfiguration referred to as *en coup de sabre* because it resembles a wound from a sword. It may be associated with hemiatrophy of the same side of the face.

SScalso occurs in association with features of other connective tissue diseases. The term *overlap syndrome* has been used to describe such patients. Undifferentiated connective tissue disease has been suggested as a designation for patients who do not have diagnostic criteria for any one connective tissue disease. *Mixed connective tissue disease* (MCTD), a syndrome involving features of systemic lupus erythematosus (SLE), SSc, polymyositis, and rheumatoid arthritis and very high titers of circulating antibody to nuclear ribonucleoprotein (RNP) antigen, will be discussed later in the chapter. *Eosinophilic fasciitis* and the *eosinophilia-myalgia syndrome* (EMS) associated with contaminated L-tryptophan ingestion (Chap. 382) are scleroderma-like illnesses and will also be discussed in this chapter.

EPIDEMIOLOGY

SSchas a worldwide distribution and affects all races. The onset of disease is unusual in childhood and young men. The incidence increases with age, peaking in the third to fifth decade. Women overall are affected approximately three times as often as men and even more often during the late childbearing years (³8:1). SSc is more frequent and severe in young black women. The annual incidence has been estimated to be 19 cases per million population. The reported prevalence of SSc is between 19 and 75 per 100,000 persons. An exceptionally high prevalence of SSc (472 per 100,000 persons) has been noted in the Choctaw Native Americans in Oklahoma -- the highest found to date in any ethnic group. Both incidence and prevalence may be underestimated because patients with early and atypical disease may be overlooked in surveys. The role of heredity has not been clarified. Several examples of familial SSc have been reported, and the finding of other connective tissue diseases and antinuclear antibodies in relatives of involved patients suggests a hereditary predisposition. However, spouses of SSc patients also have an increased incidence of antinuclear antibodies, suggesting environmental factors. Discordance of SSc in identical twins speaks against a significant genetic predisposition to disease. Immunogenetic studies have not shown strong associations between the major histocompatibility complex and susceptibility to SSc. Some studies have shown an association of SSc with HLA-A1, -B8, -DR3, or with DR3/DR52. C4A null alleles (C4AQ0) and HLA-DQA2 have been reported by some investigators to be markers for disease susceptibility. A more consistent relationship has been found between certain HLA types and the occurrence of specific autoantibodies in SSc patients. Anticentromere antibodies have been shown to be associated with HLA-DQB1*05 allele and, less often, with -DQB1*0301, -*0401, or -*0402. Antitopoisomerase 1 antibodies, on the other hand, are associated most frequently with HLA-DQB1*0301 in several populations, including Caucasians and African Americans.

Several environmental factors have been associated with the development ofSSc and scleroderma-like illnesses. SSc appears to be more common in coal and gold miners, especially in those with more extensive exposure, suggesting that silica dust may be a predisposing factor. Workers exposed to polyvinyl chloride may develop Raynaud's phenomenon, acroosteolysis, scleroderma-like skin lesions, pulmonary fibrosis, and nail fold capillary abnormalities similar to those observed in SSc. These workers may also develop hepatic fibrosis and angiosarcoma. The observation that individuals exposed to similar amounts of vinyl chloride do not develop the same degree of disease suggests that a genetic factor may determine susceptibility and disease severity. The development of scleroderma has also been associated with exposure to epoxy resins and aromatic hydrocarbons such as benzine, toluene, and trichloroethylene. In 1981, in Spain, a multisystem disease resembling scleroderma occurred following the ingestion of aniline-adulterated cooking oil (rapeseed oil). Approximately 20,000 people were affected. The patients initially developed interstitial pneumonitis, eosinophilia, arthralgias, arthritis, and myositis, followed subsequently by joint contractures, skin thickening, Raynaud's phenomenon, pulmonary hypertension, sicca syndrome, and resorption of the distal fingertips. Extensive sclerosis of the dermis and subcutaneous tissue has been noted in patients receiving pentazocine, a nonnarcotic analgesic agent. Bleomycin, an anticancer agent, produces fibrotic skin nodules, linear hyperpigmentation, alopecia, Raynaud's phenomenon, gangrene of fingers, and pulmonary fibrosis affecting mainly the lower lobes. Scleroderma and other connective tissue diseases have been reported in women who have had silicone breast implants. Recent studies have not shown that women with these implants carry an increased risk for developing scleroderma or other connective tissue diseases. Localized fibrosis, however, can occur around the implant. While environmental factors or undefined infectious agents may be of etiologic significance, the cause of SSc remains unknown.

PATHOGENESIS

The outstanding feature ofSSc is overproduction and accumulation of collagen and other extracellular matrix proteins, including fibronectin, tenascin, and glycosaminoglycans, in skin and other organs. The disease process involves immunologic mechanisms, vascular endothelial cell activation and/or injury, and activation of fibroblasts resulting in production of excessive collagen.

An early event inSSc that precedes fibrosis is vascular injury involving small arteries, arterioles, and capillaries in the skin, gastrointestinal tract, kidneys, heart, and lungs. Raynaud's phenomenon, the initial symptom of SSc in the majority of patients, is a clinical expression of the abnormal regulation of blood flow resulting from vascular injury. Injury to endothelial cells and basal lamina occurs early and is followed by thickening of the intima, narrowing of the lumen, and eventual obliteration of the vessel. As vascular damage progresses, the microvascular bed in the skin and other sites is diminished, producing a state of chronic ischemia. Vascular abnormalities can be observed in the nail folds by wide-field microscopy, which shows drop-out of capillaries with dilatation and tortuosity of remaining ones. In the skin, remaining capillaries may proliferate and dilate to become visible telangiectasia. Endothelial cell damage is reflected in elevated levels of factor VIII/von Willebrand factor in the sera of some patients with SSc.

Several mechanisms for endothelial injury or activation have been proposed inSSc. Any or all of these mechanisms may be involved in a given patient; some evidence for each exists. A cytotoxic factor for endothelium has been identified in some patients that degrades the basal lamina, releasing fragments of type IV collagen and laminin. This factor, a type IV collagenase, is secreted by activated T cells and is referred to as *granzyme 1* because of its location in cytolytic T cells. Type IV collagen and laminin fragments may stimulate an immune response to the basal lamina. Both antibodies and cell-mediated immunity to type IV collagen and laminin have been observed in some SSc patients and may be involved in endothelial injury or may be an epiphenomenon. Anti-endothelial cell antibodies (AECA) may be another mechanism for microvascular damage. In 25% of SSc patients, AECA have been shown to mediate antibody-dependent cell cytotoxicity against human endothelial cells. Circulating AECA in general have been reported in the sera of SSc patients in amounts ranging from 21 to 85%. This wide variation reflects patient selection, type of assay, and the source of endothelium. These antibodies are not specific for SSc and are found in other connective tissue diseases. The frequency of AECA is higher in patients with diffuse cutaneous SSc. They have also been shown to be associated with digital infarcts, pulmonary hypertension, and impaired alveolocapillary diffusion. Studies have show that AECA initiate programmed cell death (apoptosis), which may be an important event in the pathogenesis of SSc. These antibodies also induce expression of vascular cell adhesion molecule-1 (VCAM-1), intercellular adhesion molecule-1 (ICAM-1), E-selectin, and P-selectin on endothelial cells in SSc and stimulate the production of chemoattractants [interleukin (IL) 1, IL-8, monocyte chemotactic protein], leading to the binding of lymphocytes to the endothelium and their migration into the perivascular tissue. Elevated serum levels of VCAM-1, ICAM-1, and P-selectin are observed in early stages of SSc.

The injury to the endothelium leads to a state favoring vasoconstriction and ischemia. The damaged endothelium produces decreased amounts of prostacylin, which is an important vasodilator and inhibitor of platelet aggregation. Platelets are activated on binding to the damaged endothelium and release thromboxane, a potent vasoconstrictor. Activated platelets also release platelet-derived growth factor (PDGF), which is chemotactic and mitogenic for both smooth-muscle cells and fibroblasts, and transforming growth factor (TGF)b, which stimulates fibroblast collagen synthesis. These and other cytokines stimulate intimal fibrosis and, with their passage through the injured endothelium, may produce adventitial and perivascular fibrosis. Endothelin-1, a vasoconstricting factor released from endothelial cells on cold exposure, is also increased in SSc patients. In addition, it stimulates fibroblasts and smooth-muscle cells. The vasoconstriction action of endothelin-1 is normally opposed by endothelium-derived relaxation factor (EDRF, nitric oxide), also secreted by endothelial cells. The normal compensatory increase in EDRF is not seen in some patients withSSc, suggesting impairment of its synthesis. A deficiency of vasodilatory neuropeptides resulting from sensory system nerve damage may also produce a condition favoring vasoconstriction. Vasoconstriction itself also contributes to endothelial damage through a mechanism of reperfusion injury, resulting in vascular occlusion and fibrosis.

Existing evidence indicates that cell-mediated immunity plays a central role in the development of fibrosis inSSc. T cells, macrophages, endothelial cells, and other cells along with cytokines and growth factors interact in a complex manner to stimulate fibrosis. The vascular endothelium has been proposed as a target for cell-mediated immunity. Laminin and type IV collagen, components of the subendothelial basement membrane, induce in vitro transformation of lymphocytes from SSc patients. In the early stages of SSc, a mononuclear cell infiltrate consisting predominantly of activated helper-inducer T cells surrounds small blood vessels in the dermis. Subsequently, mononuclear cell infiltrates are found in macroscopically normal-appearing skin adjacent to areas of fibrosis. T cell hyperactivity is reflected by increased serum levels of CD4+ T cells. The ratio of CD4+ to CD8+ T cells is also increased. Elevated circulating levels ofIL-2, a product of activated T cells, and IL-2 receptors have been shown to be associated with active fibrosis. In addition, serum levels of IL-4 are increased in SSc patients. IL-4, a product of activated T cells, stimulates fibroblast chemotaxis and proliferation and collagen production. In a recent study, CD8+ T cells isolated from bronchoalveolar lavage fluid from SSc patients made IL-4 and/or IL-5 mRNA. SSc patients with these type 2 cytokines were more likely to have alveolitis and a lower forced vital capacity. Although larger studies are needed, the findings suggest that these cytokines are involved in the pathogenesis of interstitial pulmonary fibrosis. Another cytokine, interferong, is produced by activated T cells and stimulates macrophages but inhibits collagen synthesis by fibroblasts. Reduced serum levels of interferon g are found in some SSc patients. In vitro stimulation of T cells from SSc patients did not show an increased production of interferon g compared to normal individuals, suggesting an inability in SSc patients to suppress fibrosis normally.

Macrophages are present in increased numbers in the infiltrates ofSSclesions, including the pulmonary alveoli. Activated macrophages secrete several important products involved in the pathogenesis of SSc includingIL-1, IL-6, tumor necrosis factor (TNF) a,TGF-b, andPDGF. IL-1 has been shown to stimulate fibroblast proliferation and collagen synthesis. The important role for IL-6 may be in stimulating the local release of tissue inhibitor of metalloproteinase (TIMP) by fibroblasts and thereby limiting the breakdown of collagen. TNF-a, in conjunction with interferon g, can cause endothelial cell cytolysis and also induces the expression of endothelial cell adhesion molecules (see above), which are responsible for the binding of T cells and subsequent vascular injury. The role of TGF-b and PDGF secreted by macrophages and other cells is discussed below. In addition to the above cytokines, macrophages secrete *fibronectin*, a large matrix protein that is increased in SSc lesions. Fibronectin is also secreted by fibroblasts. Fibronectin interacts with collagen in the SSc lesions where it binds fibroblasts and mononuclear cells through receptors called *integrins*. Fibronectin functions as a chemoattractant and mitogen for fibroblasts.

Additional support for involvement of cell-mediated immunity in the pathogenesis ofSSc is the appearance of scleroderma-like lesions in patients with graft-versus-host disease (GVHD) after bone marrow transplantation and in a murine model of chronic GVHD, conditions known to be associated with activated T cells. GVHD and SSc are both associated with progressive skin induration, joint contractures, and gastrointestinal and pulmonary involvement and are frequently accompanied by Sjogren's syndrome. Antinuclear antibodies are present in both diseases. Raynaud's phenomenon and kidney involvement are infrequent in GVHD.

Mast cells may also be involved in the development of fibrosis. Increased numbers of mast cells are found in the dermis in both involved and uninvolved skin. Mast cell degranulation has been noted in skin that subsequently became fibrosed. Interaction with T cells may be one mechanism for mast cell degranulation resulting in release of products that stimulate fibroblast collagen synthesis. Release of histamine from mast cells may also contribute to edema observed in early disease.

Fibroblast growth and synthesis of collagen, fibronectin, and glycosaminoglycans are increased inSSc. Fibroblasts from SSc appear to have aberrant regulation of growth compared with fibroblasts from normal persons. When fibroblasts from affected SSc skin are removed and cultured in vitro, they continue to produce excessive quantities of collagen. The collagen is biochemically normal, and the proportion of type I to type III is the same as in normal skin. Fibroblasts from SSc patients appear to be in a state of permanent activation, most likely as a result of stimulation by cytokines. These activated cells are thought to represent an expanded subpopulation of fibroblasts that inherently express increased matrix genes. Studies have revealed a subpopulation of SSc fibroblasts that produces two to three times more collagen than other cells from the same tissue. Fibroblasts expressing elevated levels of mRNA for types I and III collagen have been demonstrated by in situ hybridization, particularly around dermal blood vessels in affected SSc skin. Collagen deposition is also initially perivascular in other organs including myocardium, muscle, and kidney. A small number of fibroblasts express increased levels of mRNA for types VI and VII collagen. Type VII collagen is normally found at the dermal-epidermal basement membrane zone and is the major component of anchoring fibrils that act to stabilize the attachment of the basement membrane to the underlying dermis. In SSc patients, type VII collagen is found throughout the dermis and may account for the indurated, tightly bound skin in this disease.PDGFreceptors are expressed on SSc fibroblasts not only from affected areas but also from macroscopically normal-appearing skin. Fibroblasts from normal persons lack expression of these receptors.TGF-bhas been shown to upregulate the expression of these receptors in SSc fibroblasts but not in normal cells and, in conjunction with PDGF, stimulates SSc fibroblast proliferation. Macrophages and fibroblasts are capable of secreting PDGF and TGF-b, and activated T cells release TGF-b. TGF-b also induces the autocrine production of PDGF-related peptides, referred to as connective tissue growth factor (CTGF), by fibroblasts. TGF-b interacts with CTGF on fibroblasts to stimulate fibroblast proliferation and collagen synthesis. Serum levels of CTGF have been found to be elevated in SSc and correlate with the degree of dermal and pulmonary fibrosis.

Fibroblasts may activate T cells to release cytokines that stimulate fibrosis. Fibroblasts inSScpatients have been shown to have increased expression of an adhesion molecule,ICAM-1, which binds to specific integrins on T cells. This binding allows interaction between T cell antigen receptor and class II molecules and antigen on fibroblasts, resulting in T cell activation and cytokine release. T cells may also be activated by their interaction with extracellular matrix molecules including collagen, fibronectin, and laminin.

Recent studies have suggested that microchimerism may be involved in the pathogenesis ofSSc. Microchimerism in SSc is of interest because of the clinical similarities between SSc andGVHDafter allogeneic bone marrow transplantation. Also relevant are the predilection for women in SSc and the increased incidence of SSc in women after the childbearing years. Fetal progenitor cells can persist in the serum of normal women for many years after childbirth. Compared to normal controls, both the quantity and frequency of fetal cells have been found to be increased in the serum of SSc patients. Microchimerism can also occur in nulligravid women and in men with SSc as non-host cells may come from blood transfusion, engraftment of cells from a twin, or from maternal cells in utero. Two-directional traffic of cells occurs during pregnancy. The mechanism by which microchimerism is involved in the pathogenesis is not known, but it is conceivable that these small numbers of non-host cells interfere with immune regulation, leading to autoimmunity.

Chromosomal abnormalities have been noted in >90% ofSScpatients. These acquired abnormalities include chromatid breaks, acentric fragments, and ring chromosomes and are found in ~30% of mitotic cells. A chromosomal breakage factor has been found in the serum of SSc patients and their first-degree relatives. The significance of these chromosomal abnormalities is unknown.

PATHOLOGY

Skin In the skin, a thin epidermis overlies compact bundles of collagen that lie parallel to the epidermis. Fingerlike projections of collagen extend from the dermis into the subcutaneous tissue and bind the skin to the underlying tissue. Dermal appendages are atrophied, and rete pegs are lost. In early stages of disease, a mononuclear cell infiltrate of predominantly T cells surrounds small dermal blood vessels. Increased numbers of T cells, monocytes, plasma cells, and mast cells are found, particularly in the lower dermis of involved skin.

Gastrointestinal Tract In the lower two-thirds of the esophagus, the histologic findings consist of a thin mucosa and increased collagen in the lamina propria, submucosa, and serosa. The degree of fibrosis is less than in the skin. Atrophy of the muscularis in the esophagus and throughout the involved portions of the gastrointestinal tract is more prominent than the amount of fibrotic replacement of muscle. Ulceration of the mucosa is often present and may be due to eitherSSc or superimposed peptic esophagitis. Chronic esophageal reflux can lead to metaplasia of the lower esophagus (Barrett's esophagus), which is a premalignant lesion. Striated muscles in the upper third of the esophagus are relatively spared. Similar changes may be found throughout the gastrointestinal tract, especially in the second and third portions of the duodenum, in the jejunum, and in the large intestine. Atrophy of the muscularis of the large intestine may lead to the development of large-mouth diverticula. In the later stages of the disease, the involved portions of the gastrointestinal tract become dilated. Infiltration of lymphocytes and plasma cells in the lamina propria is also present.

Lung With pulmonary involvement, diffuse interstitial fibrosis, thickening of the alveolar membrane, and peribronchial and pleural fibrosis are observed. Bronchiolar epithelial proliferation accompanies the pulmonary fibrosis. Rupture of septa produces small cysts and areas of bullous emphysema. Small pulmonary arteries and arterioles show intimal thickening, fragmentation of the elastica, and muscular hypertrophy; this may occur without interstitial pulmonary fibrosis and produce pulmonary hypertension, particularly in a subset of patients with limited cutaneousSSc.

Musculoskeletal System The synovium in patients with arthritis is similar to that seen in early rheumatoid arthritis and shows edema with infiltration of lymphocytes and plasma cells. A characteristic finding is a thick layer of fibrin overlying and within the synovium. Later in the disease the synovium may become fibrotic. Fibrinous deposits appear on the surfaces of tendon sheaths and in the overlying fascia and may lead to audible creaking over moving tendons.

Histologic features of primary myopathy consist of interstitial and perivascular lymphocytic infiltrations, degeneration of muscle fibers, and interstitial fibrosis. Arterioles may be thickened, and capillaries may be decreased in number. Pathologic and electrophysiologic findings of polymyositis in proximal muscles are present in the few patients who are considered to have the overlap syndrome ofSSc and polymyositis.

Heart Cardiac involvement consists of degeneration of myocardial fibers and irregular areas of interstitial fibrosis that are most prominent around blood vessels. Intermittent spasm of blood vessels may result in contraction band necrosis, similar to change observed in myocardial infarction in patients with atherosclerotic coronary artery disease. Fibrosis also involves the conduction system, leading to atrioventricular conduction defects and arrhythmias. The wall of smaller coronary arteries may be thickened. Fibrinous pericarditis and pericardial effusions are found in some patients.

Kidney Renal involvement is found in over half the patients and consists of intimal hyperplasia of the interlobular arteries; fibrinoid necrosis of the afferent arterioles, including the glomerular tuft; and thickening of the glomerular basement membrane. Small cortical infarctions and glomerulosclerosis may be present. The renal pathologic change is often indistinguishable from that observed in malignant hypertension. Renal vascular lesions, however, may be present in the absence of hypertension. Immunofluorescence studies of kidney have shown IgM, complement components, and fibrinogen in the walls of affected vessels. Angiographic renal studies in patients withSSc may show constriction of the intralobular arteries, a finding that simulates the vasospasm of the digital arteries observed in Raynaud's phenomenon. Cold-induced Raynaud's phenomenon has been shown to decrease renal blood flow.

Other Organs Primary liver involvement is not common. Primary biliary cirrhosis occurs in some patients, particularly in those with the limited cutaneous form ofSSc. Fibrosis of the thyroid gland may develop in the presence or absence of autoimmune thyroiditis.

Thickening of the periodontal membrane with replacement of the lamina dura is demonstrated radiographically as widening of the periodontal space and may cause gingivitis and loosening of the teeth. The decreased oral aperture and mucosal dryness make eating and oral hygiene difficult.

CLINICAL MANIFESTATIONS (See Table 313-3)

Raynaud's Phenomenon SScusually begins insidiously; the first symptoms are frequently Raynaud's phenomenon and puffy fingers. Some 95% of patients will experience Raynaud's phenomenon, which is defined as episodic vasoconstriction of small arteries and arterioles of fingers, toes, and sometimes the tip of the nose and earlobes. Episodes are brought on by cold exposure, vibration, or emotional stress. Patients experience pallor and/or cyanosis followed by rubor on rewarming. Pallor and/or cyanosis are usually associated with coldness and numbness of fingers and/or toes, and rubor with pain and tingling. Not all patients appreciate the three color phases. A history of digit pallor appears to be the most reliable symptom for the presence of Raynaud's phenomenon. Raynaud's phenomenon may precede skin changes by several months or even years in those patients who subsequently develop the limited cutaneous form of SSc. In diffuse cutaneous SSc, skin changes are seen typically within a year of the onset of Raynaud's phenomenon. After 2 or more years of Raynaud's phenomenon, few patients who have this as their only symptom will subsequently develop SSc.

Skin Features In early disease, fingers and hands are swollen. Swelling may also involve forearms, feet, lower legs, and face. However, lower extremities are relatively spared. This edematous phase may last for a few weeks, months, or even longer. The edema may be pitting or nonpitting and accompanied by erythema. The skin changes begin distally in the extremities and advance proximally. The skin gradually becomes firm, thickened, and eventually tightly bound to underlying subcutaneous tissue (indurative phase). In patients with diffuse cutaneous scleroderma, skin changes will become generalized, involving initially the extremities, followed by the face and trunk over a period of time, varying from months to a few years. In some patients, the skin changes may develop gradually over several years. Rapid progression of these changes over a 1- to 3-year period is associated with a greater risk of visceral disease, particularly of the lungs, heart, or kidneys. Also in diffuse cutaneousSSc, the skin changes usually peak in 3 to 5 years and then slowly improve. On the other hand, patients with limited cutaneous scleroderma will usually have a more gradual progression of skin changes, which are restricted to fingers or distal extremity and face and may continue to worsen. In both subsets of SSc, skin thickening is usually greater in the distal extremity. After many years of disease, the skin may soften and return to normal thickness or become thin and atrophic.

In the extremities, the taut skin over fingers gradually limits full extension, and flexion contractures develop. Ulcers may appear on the volar pads of the fingertips and over bony prominences such as elbows, malleoli, and the extensor surface of the proximal interphalangeal joints of the hands. These ulcers may become secondarily infected. The volar pads of the fingertips develop pitting scars and lose soft tissue. In some instances, resorption of the terminal phalanges occurs. Skin over the extremities, face, and trunk may become darkly pigmented, even without exposure to the sun. Hyperpigmentation of the skin may occur over superficial blood vessels and tendons. Areas of hypopigmentation may also develop, similar to vitiligo, involving the eyebrows, scalp, and trunk. The sparing of pigment around hair follicles gives the skin a "salt-and-pepper" appearance. Other patients may develop a diffuse tanning of the skin. The skin loses hair, oil, and sweat glands and so becomes dry and coarse. Vaginal dryness occurs and may cause dyspareunia.

In some patients, particularly those with the limited cutaneous form of disease, calcific deposits develop in intracutaneous and subcutaneous tissue. The sites commonly involved are periarticular tissue, digital pads, olecranon and prepatellar bursae, and skin along the extensor surface of the forearms. The overlying skin may break down, with drainage of calcific material. Involvement of the face results in thinning of the lips, loss of skin wrinkles and facial expression, as well as microstomia, which may make eating and dental hygiene difficult. The nose takes on a pinched or beaklike appearance. Wrinkles appear around the mouth perpendicular to the lips. Small telangiectatic mats may appear on the fingers, face, lips, tongue, and buccal mucosa after several years. They are seen more frequently in patients with limited cutaneousSSc but are also observed in patients with long-standing diffuse cutaneous SSc. The capillary beds of nail folds of the fingers may show enlargement of capillaries with little or no capillary loss, usually indicative of limited cutaneous scleroderma. In diffuse cutaneous scleroderma, there is disorganization of the capillary beds with dilated capillaries interspersed with areas where capillaries have disappeared. These capillary changes, which are observed by wide-angle microscopy or with an ophthalmoscope used as a magnifier, are not found in patients who have only Raynaud's phenomenon.

Musculoskeletal Features More than half the patients withSSccomplain of pain, swelling, and stiffness of the fingers and knees. A symmetric polyarthritis resembling rheumatoid arthritis may be seen. In more advanced stages of the disease, leathery crepitation can be palpated over moving joints, especially the knee. Extensive fibrotic thickening of the tendon sheaths in the wrist can produce a carpal tunnel syndrome. Muscle weakness is usually present in patients with severe skin involvement and, in most cases, is due to disuse atrophy. There is a distinctive histologic myopathy that accompanies SSc that is not associated with muscle enzyme abnormalities. A few patients develop a myositis characterized by proximal muscle weakness and muscle enzyme elevations that are identical to polymyositis (overlap syndrome). In addition to terminal phalanges, resorption of bone may involve ribs, clavicle, and angle of mandible.

Gastrointestinal Features The majority of patients from both subsets ofSSc have gastrointestinal involvement. Symptoms attributable to esophageal involvement are present in >50% of patients and include epigastric fullness, burning pain in the epigastric or retrosternal regions, and regurgitation of gastric contents. These symptoms, most noticeable when the patient is lying flat or bending over, are due to the reduced tone of the gastroesophageal sphincter and to dilatation of the distal esophagus. Peptic esophagitis frequently occurs and may lead to strictures and narrowing of the lower esophagus. However, it seldom results in bleeding. Barrett's metaplasia may develop, but transition to adenocarcinoma is uncommon. Dysphagia, particularly of solid foods, may occur independent of other esophageal symptoms and is caused by loss of esophageal motility due to neuromuscular dysfunction. Manometry or cineradiography reveals decreased amplitude or disappearance of peristaltic waves in the lower two-thirds of the esophagus. Raynaud's phenomenon in the absence of a connective tissue disease is also associated with esophageal dysmotility. Later in the course of the illness, dilatation and atony of the lower portion of the esophagus as well as reflux are seen. With gastric involvement, barium studies show dilatation, atony, and delayed gastric emptying. Patients may complain of early satiety. Gastric outlet obstruction can also occur.

Hypomotility of the small intestine produces symptoms of bloating and abdominal pain and may suggest an intestinal obstruction or paralytic ileus (pseudoobstruction). Malabsorption syndrome with weight loss, diarrhea, and anemia is due to bacterial overgrowth in the atonic intestine or possibly to obliteration of lymphatics by fibrosis. Roentgenographic features of the second and third portions of the duodenum and of the jejunum include dilatation, loss of the usual feathery pattern, and delayed disappearance of barium. Pneumatosis intestinalis occasionally occurs and appears as radiolucent cysts or linear streaks within the wall of the small intestine. Benign pneumoperitoneum may result from the rupture of these cysts. Involvement of the large intestine may cause chronic constipation and fecal impaction with episodes of bowel obstruction. A segment of atonic bowel may act as a fulcrum for intussuception to occur. Barium studies of the large intestine may show dilatation, atony, and large-mouth diverticula. Laxity of the anal sphincter may cause incontinence or rarely anal prolapse. Some patients may have gastrointestinal features ofSSc with little or no cutaneous or other organ involvement, referred to as *SSc sine scleroderma*. Vascular ectasia may develop in the stomach and intestine and can be the source of gastrointestinal bleeding. These dilated submucosal capillaries in the stomach appear on endoscopy as broad stripes -- hence the term "watermelon stomach."

Pulmonary Features Pulmonary involvement occurs in at least two-thirds ofSScpatients and is now the leading cause of death in SSc, replacing renal disease, which can usually be treated effectively. The most common symptom is exertional dyspnea, often accompanied by a dry, nonproductive cough. Bilateral basilar rales may be present. In the majority of patients, symptoms usually correlate with radiologic evidence of pulmonary fibrosis and with restrictive lung disease on pulmonary function tests.

Pulmonary function tests are frequently abnormal and show a reduction in vital capacity and decreased lung compliance. Impairment of gas exchange is reflected by a low diffusing capacity and low P_{O2} with exercise. These abnormalities may be present even when the chest radiograph is normal. Chest film may show a pattern of linear densities, mottling, and honeycombing involving most prominently the lower two-thirds of the lung. Early interstitial pulmonary disease can be detected by high-resolution computed tomography (HRCT) and bronchoalveolar lavage (BAL). Active inflammatory alveolitis gives a "ground glass" appearance on HRCT. The recovery by BAL of increased numbers of cells, mostly alveolar macrophages accompanied by neutrophils or eosinophils, is evidence for alveolitis.

Both interstitial fibrosis and vascular lesions are found in the lungs of patients withSSc. Interstitial pulmonary fibrosis may be the predominant lesion in patients with diffuse or limited cutaneous SSc. Patients with diffuse cutaneous involvement who have antitopoisomerase 1 antibodies are particularly at risk of developing severe pulmonary fibrosis. In the absence of significant interstitial fibrosis, a severe form of pulmonary arterial hypertension may develop after many years of disease in a subset of patients with limited cutaneous SSc. Fewer than 10% of patients will develop this complication, which is caused by narrowing and obliteration of pulmonary arteries and arterioles by intimal fibrosis and medial hypertrophy. Pulmonary hypertension is manifested initially by exertional dyspnea and eventually by the appearance of right-sided heart failure. Pulmonary artery pressure can be measured noninvasively by two-dimensional echocardiography. The prognosis is extremely poor with the development of pulmonary hypertension; the mean duration of survival is approximately 2 years.

A less common pulmonary problem is aspiration pneumonia resulting from gastric reflux due to lower esophageal atony. Restriction of chest movement caused by extensive fibrotic skin involvement of the thorax rarely occurs. Superimposed bacterial or viral infection can be a serious complication in patients with pulmonary fibrosis. An increased frequency of alveolar cell and bronchogenic carcinoma is seen in patients with pulmonary fibrosis.

Cardiac Features Primary cardiac involvement inSScincludes pericarditis with or without effusions, heart failure, and varying degrees of heart block or arrhythmias. The majority of patients with diffuse cutaneous SSc have cardiac abnormalities. Cardiomyopathy attributable to myocardial fibrosis appears in <10% of patients and involves primarily those patients with diffuse cutaneous scleroderma. Radionuclide studies have shown abnormalities of left ventricular function due to myocardial fibrosis. Cold-induced vasospasm of the hands produces defects in myocardial thallium perfusion. The characteristic pathologic feature of contraction band necrosis results from cardiac muscle damage caused by intermittent vasospasm of coronary vessels. Patients may experience angina pectoris even though coronary angiograms are normal. Patients can also develop left ventricular failure secondary to systemic hypertension or cor pulmonale secondary to pulmonary arterial hypertension.

Renal Features Renal failure was the leading cause of death inSScuntil the advent of effective treatment. Significant renal disease occurs mostly in those patients with diffuse cutaneous scleroderma. A high risk of renal crisis is present in those patients who have rapidly progressive widespread skin thickening in their first 2 to 3 years of disease. Renal crisis is characterized by malignant hypertension, which can progress rapidly to renal failure. These patients manifest hypertensive encephalopathy, severe headache, retinopathy, seizures, and left ventricular failure. Hematuria and proteinuria are followed by oliguria and renal failure. The mechanism for the hypertensive crisis is activation of the renin-angiotensin system. Before the advent of effective antihypertensive drugs, the majority of these patients died within 6 months. A small number of patients may develop renal crises in the absence of hypertension. Renal failure can also develop insidiously later in the course of disease in the setting of mild to moderate hypertension and proteinuria. In these patients or those with clinically unrecognized renal disease, reduction of renal plasma flow secondary to heart failure or volume depletion resulting from overdiuresis may precipitate renal crisis. An indicator of impending renal failure is microangiopathic anemia, which may occur in a normotensive patient. The presence of a large chronic pericardial effusion may also herald subsequent renal failure.

Other Features Symptoms of dry eyes and/or dry mouth are frequently present in patients withSSc. Lip biopsy may show lymphocytic infiltration of minor salivary glands characteristic of Sjogren's syndrome or intraglandular or periglandular fibrosis secondary to SSc. Antibodies to SS-A (Ro) and/or SS-B (La) are found in those patients with lip biopsies consistent with Sjogren's syndrome (overlap syndrome-SSc and Sjogren's syndrome) and not in those with salivary gland fibrosis.

Hypothyroidism occurs in a significant number of patients and may be associated with high levels of antithyroid antibodies. Fibrosis of the thyroid gland may be present but also occurs in the absence of autoimmune thyroiditis. Other manifestations ofSScinclude trigeminal neuralgia and male impotence secondary to decreased penile tumescence. These men have normal serum levels of testosterone and gonadotropins. Pathogenesis of this abnormality has been considered to be caused by vascular and/or autonomic nervous system abnormalities. Biliary cirrhosis is occasionally observed in patients with limited cutaneous SSc.

LABORATORY FINDINGS

The erythrocyte sedimentation rate may be elevated. Hypoproliferative anemia related to chronic inflammation is the most common cause of anemia inSSc. Anemia may also be caused by iron deficiency secondary to gastrointestinal bleeding. Bacterial overgrowth due to atony of the small bowel may lead to vitamin $B_{12}$and/or folic acid-deficiency anemia. Microangiopathic hemolytic anemia is most often associated with renal involvement and is caused by the presence of intravascular fibrin in renal arterioles. Polyclonal hypergammaglobulinemia, consisting mostly of IgG, is found in approximately half the patients. Rheumatoid factor, in low titer, is present in 25% of patients. Cryoglobulins may be present in the serum. Antinuclear antibodies detected by using a cultured human laryngeal carcinoma cell line (HEp-2) substrate are present in 95% of patients (Table 313-4). Antinuclear antibodies that have a high specificity for

SSc are antitopoisomerase 1 (Scl-70), antinucleolar, and anticentromere. Antitopoisomerase 1, originally called anti-Scl-70, recognizes the nuclear enzyme DNA topoisomerase 1, a nuclear enzyme involved in the unwinding of DNA for replication and RNA transcription. These antibodies are found in ~20% of all SSc patients and in ~40% of those with diffuse cutaneous SSc. They are associated with diffuse cutaneous involvement, interstitial pulmonary disease, and renal and other visceral organ involvement. A very high frequency of these antibodies has been reported in Choctaw Native Americans in association with diffuse cutaneous SSc. They are seldom present in other disorders or in conjunction with anticentromere antibodies. Anticentromere antibodies react with protein antigens located in the kinetochore region of chromosomes and are present in 40 to 80% of patients with limited cutaneous scleroderma or CREST syndrome. Anticentromere antibodies are found in only about 2 to 5% of patients with diffuse cutaneous scleroderma and rarely in other connective tissue diseases. They are found occasionally in patients with only Raynaud's phenomenon and may indicate subsequent development of limited cutaneous disease. Antinucleolar antibodies are relatively specific for SSc and are present in ~20 to 30% of patients. Several antinucleolar antibodies have been associated with SSc: Anti-RNA polymerases I, II, and III are found in patients with diffuse cutaneous SSc who have a higher prevalence of renal and cardiac involvement. Anti-ThRNP has been found in patients with limited cutaneous SSc, and anti-PM-Scl, formerly referred to as anti-PM1, along with anti-Ku, may be found in a subset of patients with overlapping features of limited cutaneous SSc and polymyositis. Anti-U$_3$RNP (anti-fibrillarin) is also highly specific for SSc and may be associated with skeletal muscle disease, bowel involvement, and pulmonary arterial hypertension. Anti-U$_1$RNP is found in ~5 to 10% of SSc patients and in 95 to 100% of those patients with the overlap syndrome ofMCTD. The titers in MCTD are usually high (see below). Anti-SS-A (Ro) and/or anti-SS-B (La) are present in those patients with overlap syndrome of SSc and Sjogren's syndrome.

DIAGNOSIS

The diagnosis ofSScpresents no difficulty in the presence of Raynaud's phenomenon, with typical skin lesions and visceral involvement. Although Raynaud's phenomenon may be the first symptom of SSc, most patients with Raynaud's phenomenon alone do not develop a connective tissue disease. Other causes of Raynaud's phenomenon include thoracic outlet (scalenus anticus and cervical rib) syndromes, shoulder-hand syndrome, trauma (jackhammer or vibratory machine operators), previous cold injury, vinyl chloride exposure, and circulating cryoglobulins or cold agglutinins. Linear scleroderma and morphea are localized forms of scleroderma that can usually be distinguished clinically. In early disease, SSc may initially be confused with rheumatoid arthritis,SLE, or polymyositis when articular or muscle involvement is prominent. SSc without cutaneous involvement should be considered in patients with unexplained pulmonary fibrosis, pulmonary hypertension, cardiomyopathies, heart block, dysphagia, or malabsorption syndrome. Several conditions have scleroderma-like features but lack the visceral involvement. Scleredema (scleredema adultorum of Buschke) occurs predominantly in children and is characterized by painless edematous induration involving the face, scalp, neck, trunk, and proximal portions of the extremities. Involvement of the hands and feet usually does not occur. Scleredema may be associated with previous streptococcal infection and is usually self-limited, resolving in 6 to 12 months. Histology reveals accumulation of mucopolysaccharides in the dermis and skeletal muscle. A rare entity, scleromyxedema is manifested by yellowish or pale red papules in association with diffuse skin thickening that may involve the face and hands. Acid mucopolysaccharide deposits are found in the dermis. Monoclonal IgG may be detected in some of these patients. Patients with insulin-dependent diabetes mellitus may develop digital sclerosis and contractures (prayer hand deformity). Primary amyloidosis and amyloidosis associated with multiple myeloma may involve the skin of the extremities and face diffusely to give the appearance of scleroderma. Biopsy will clearly differentiate these entities.

COURSE AND PROGNOSIS

The course ofSSc is quite variable. Until the disease differentiates into recognizable subsets, prognosis in early disease is difficult to predict. Patients with limited cutaneous scleroderma, especially those with anticentromere antibodies, have a good prognosis, with the notable exception of those few patients,<10%, who after ³10 to 20 years develop pulmonary arterial hypertension. Malabsorption syndrome and primary biliary cirrhosis are the causes of morbidity and mortality in some patients with limited cutaneous disease. On the other hand, the prognosis is generally worse in patients with diffuse cutaneous disease, particularly when the onset occurs at an older age. In addition, males have a worse prognosis. Renal and other visceral organ disease may develop early in the course of those patients with rapidly progressive generalized skin thickening. Death occurs most often from pulmonary, cardiac, and renal involvement. With the advent of effective therapy for renal crisis along with renal dialysis for those patients with renal failure, survival has greatly improved. In patients with diffuse cutaneous disease, the 5-year cumulative survival rate is ~70% and the 10-year is ~55%. In limited cutaneous disease the 5-year is ~90% and the 10-year is ~75%.

Skin may spontaneously soften after years of disease. Softening occurs in the reverse order of original skin involvement, beginning with the trunk and followed by the proximal and then the distal extremities. Sclerodactyly and flexion contractures may persist. Skin thickness may eventually approach normal; however, the skin may be atrophic.

TREATMENT

Even thoughSSccannot be cured, treatment of involved organ systems can relieve symptoms and improve function. The doctor-patient relationship is extremely important in caring for patients with this chronic debilitating illness. Once the diagnosis of SSc has been made, the patient and family should be instructed about this disorder. The patient will need repeated explanations and reassurances throughout his or her illness. Depending on the severity of illness, the patient will require monitoring of blood pressure, blood counts, urinalysis, and monitoring of renal and pulmonary function on a regular basis.

Effectiveness of drug therapy inSSc is difficult to evaluate because of the variable course and severity of the disease. Many drugs have been used in the treatment of SSc without any consistent or prolonged benefit. In uncontrolled studies, D-penicillamine has been reported to reduce skin thickening and prevent development of significant organ involvement when compared to similar historic controls. Five-year cumulative survival rates of 80% have been reported in D-penicillamine-treated patients. This drug interferes with inter- and intramolecular cross-linking of collagen and is also immunosuppressive. Its immunosuppressive activity may also lead to decreased collagen production. Penicillamine is better tolerated when started at a low dose, usually 250 mg/d, and then increased at 1- to 3-month intervals up to 1.5 g/d as tolerated. Although a few patients can tolerate higher doses, most patients are maintained on a dose between 0.5 and 1 g/d. For optimal absorption, it is important to give this drug 1 h before or 2 h after a meal. This drug can be quite toxic; its more serious complications include glomerulonephritis with nephrotic syndrome, aplastic anemia, leukopenia, thrombocytopenia, and myasthenia gravis. Other side effects are fever, rash, anorexia, nausea, and loss of taste. Patients should have monthly complete blood counts (including platelet count) and urinalyses. The results of a 2-year double-blind randomized study comparing high-dose D-penicillamine (750 to 1000 mg/d) with low-dose D-penicillamine (125 mg every other day) in patients with early diffuse cutaneous SSc were recently reported. The degree of skin thickening and the occurrence of renal crises and other organ involvement as well as mortality were not significantly different between the high- and low-dose treated groups. This study suggested that there was no advantage in using doses >125 mg every other day. Azathioprine, methotrexate, cyclophosphamide, and other immunosuppressives have also been used in SSc and should be reserved for those patients with rapidly progressive disease. Control studies are lacking. Trials of treatment with recombinant interferon g, 5-fluorouracil, and extracorporeal photochemotherapy have shown improvement in some disease parameters. No therapy, however, has been clearly demonstrated in a controlled, prospective study to suppress or reverse the disease process of SSc. Because of the poor prognosis in SSc patients who have a rapid onset of diffuse cutaneous disease and early visceral organ involvement (pulmonary, cardiac, or renal), clinical trials are under way using high-dose immunosuppressive therapy followed by autologous stem cell transplantation. The rationale is that high doses of an immunosuppressive drug such as cyclophosphamide may reverse or modify the disease course. The autologous stem cell transplantation permits the rapid reconstitution of hematopoiesis.

Antiplatelet therapy may play a role in the treatment ofSSc, since the biologic products of platelets affect blood vessels. Low doses of aspirin block the formation of thromboxane A_2, a powerful vasoconstrictor and platelet aggregator. In addition, dipyridamole, 200 to 400 mg in divided daily doses, also decreases platelet adhesion to damaged vessel walls. While these drugs have a reasonable therapeutic rationale, a 2-year double-blind study did not show any benefit from their use. Reports of beneficial effects of colchicine or chlorambucil have not been documented in controlled studies.

Glucocorticoids are indicated in those patients with inflammatory myositis or pericarditis. The initial dose is 40 to 60 mg/d and is tapered based on clinical improvement (see below). They should not be used for the indolent primary form of muscle disease of SSc. Prednisone in the range of 20 to 40 mg/d may decrease edema associated with the edematous phase of early skin involvement. Glucocorticoids are not otherwise indicated in the long-term treatment of SSc. High doses of glucocorticoids may play a role in precipitating acute renal failure. A retrospective case-control study in patients with early diffuse cutaneous SSc showed a significant association between prior high-dose glucocorticoids (prednisone^3 15 mg/d) and the development of scleroderma renal crisis. Based on these observations, immunosuppressive drugs (e.g. methotrexate, azathioprine, or cyclophosphamide) should be considered in treating the inflammatory myositis, pericarditis, or early inflammatory skin disease.

The management of Raynaud's phenomenon is directed at control of vasospasm. Patients should be advised to dress warmly and wear mittens and socks, not to smoke, to remove causes of external stress, and to avoid drugs such as amphetamine and ergotamine. Cold drafts should be avoided. Air-conditioned rooms in warm climates can also be a problem for patients with Raynaud's phenomenon. Beta-blocking drugs may make Raynaud's phenomenon worse. Warmth of the central body induces peripheral vasodilatation. Drugs that block sympathetic vasoconstriction, such as reserpine,a-methyldopa, phenoxybenzamine, and prazosin, may be useful in the treatment of Raynaud's phenomenon, but their side effects often curtail extended use. The calcium channel blockers nifedipine, diltiazem, and the longer acting amlodipine can be effective in alleviating Raynaud's phenomenon, but side effects of light-headedness and palpitations may limit their use. The sustained-release form of nifedipine is better tolerated; the dose is 30 mg/d up to 60 or 90 mg/d as required to control symptoms. Nitroglycerine paste, applied to an affected digit, may improve local blood flow. In a 12-week pilot study, losartan, a specific nonpeptide angiotensen II type 1 receptor antagonist, reduced the severity and frequency of Raynaud's phenomenon episodes. Ketanserin, an oral serotonin antagonist, also has been shown to be effective. Selective serotonin reuptake inhibitors (e.g., fluoxetine) may be beneficial in some patients. These drugs decrease platelet 5-hydroxytryptamine, which is thought to play a role in the pathogenesis of Raynaud's phenomenon. Studies with intravenous iloprost, a prostacyclin analogue, have shown a decrease in frequency and severity of Raynaud's phenomenon and healing of digital ulcers in some patients. Iloprost is still not available in the United States for general use. Intravenous alprostadil, a prostaglandin, can be effective in treating severe Raynaud's phenomenon with digital ulcers. Epoprostenol (prostacyclin), used in the treatment of pulmonary hypertension, also improves Raynaud's phenomenon. Pentoxifylline may also improve perfusion by increasing the deformability of the red cell plasma membranes. Techniques of biofeedback have also been used with variable success for teaching patients to control the temperature of their hands. Stellate ganglion blockage may be useful in temporarily alleviating severe ischemic pain in the fingers. Surgical sympathectomy usually provides only temporary improvement, and it, along with other forms of therapy, does not prevent progression of the vascular lesion. Digital sympathectomy can be effective in some patients. The response to any therapy for Raynaud's phenomenon is limited by the degree of existing structural narrowing of digital arteries. In patients with severe Raynaud's phenomenon and refractory digital ulcers, distal ulnar artery occlusion should be considered. A positive Allen test is suggestive, and the diagnosis is confirmed by angiography. When ulnar artery occlusion is present, revascularization and a digital sympathectomy may be beneficial. Gangrene of distal digits may occur and require surgical amputation.

Numerous drugs have been claimed to soften the hidebound skin, but documentation in controlled studies is lacking. These drugs include D-penicillamine, colchicine, *p*-aminobenzoic acid, and vitamin E. In a recent randomized, double-blind, placebo-controlled trial, recombinant human relaxin given by continuous subcutaneous infusion for 24 weeks was associated with reduced skin thickening and improved mobility in patients with moderate to severe diffuse cutaneous scleroderma. Relaxin, a hormone associated with pregnancy, has been shown to have antifibrotic properties.

Dryness of the skin may be reduced by avoiding frequent use of detergent soaps and by regularly applying hydrophilic ointments and bath oils. Regular exercise helps to maintain flexibility of extremities and pliability of skin. Massaging the skin several times a day may also be beneficial. Fingertip ulcerations can be protected by applying a guard or cage over the end of the finger. The use of an occlusive dressing, such as the hydrocolloid duo-DERM or other membranes, over a noninfected ulcer may promote healing and protect the finger. Skin ulcers should be kept clean by soaking or by surgical or chemical debridement. Sympatholytic drugs or local nitroglycerine paste applied to or adjacent to the ulcer may be beneficial in promoting healing. Infected ulcers can usually be treated with topical antibiotics but may require systemic antibiotics, especially when there is a question of underlying osteomyelitis. The development of calcinosis cannot be prevented, nor can deposits be dissolved. Warfarin has been reported to reduce calcinosis in a few patients.

In patients experiencing dry mouth, frequent sips of water help to relieve symptoms. Pilocarpine hydrochloride tablets may increase salivary secretions in some patients. Patients with dry eyes should use artificial tears regularly.

Patients with reflux esophagitis are treated with small, frequent meals, antacids between meals, and elevation of the head of the bed. Patients should be advised not to lie down for a few hours after a meal and to avoid coffee, tea, alcohol, peppermint, and chocolate, which reduce the pressure of the lower esophageal sphincter. Fatty foods and late-evening snacks should be avoided. Cimetidine, ranitidine or other newer H_2blockers may be beneficial. Gastric acid (proton) pump inhibitors are more effective in treating erosive esophagitis than are H_2blockers. Metoclopramide and cisapride increase gastrointestinal motility but do not significantly improve esophageal motility. They both increase lower esophageal sphincter tone and can be of help in some patients. Cisapride is no longer available (as of July 2000) because it caused life-threatening arrhythmias. Nifedipine and, to a lesser extent, diltiazem reduce lower esophageal sphincter tone resulting in esophageal reflux. Patients with dysphagia should be instructed to chew their food thoroughly and wash it down with fluids. Malabsorption syndrome due to duodenal hypomotility and bacterial overgrowth causes bloating and diarrhea, which may improve with intermittent use of appropriate antibiotics. Antibiotics are rotated every 2 weeks. Commonly used antibiotics are metronidazole, vancomycin, erythromycin, ciprofloxacin, neomycin, and tetracycline. Patients with severe debilitating malabsorption may benefit from parenteral hyperalimentation. Patients with chronic intestinal pseudoobstruction might respond to octreotide. Stool softeners and mild laxatives are usually adequate for treating constipation caused by hypomotility of the colon.

Articular symptoms are treated with nonsteroidal anti-inflammatory agents. Low-dose prednisone (£10 mg/d) may improve symptoms in those not responding to these agents. Physical therapy may help to reduce the loss of joint mobility that occurs inSSc.

In patients with diffuse cutaneousSSc, the early recognition of alveolitis as previously described (see "Pulmonary Features") may allow treatment that might slow or prevent the development of pulmonary fibrosis. Cyclophosphamide has been reported in uncontrolled studies to be beneficial, and a controlled study is presently being done. The role of glucocorticoids in preventing progression of interstitial lung disease is not clear but may be of benefit in early disease. Pulmonary fibrosis is not reversible, and therefore treatment is directed at symptoms or complications. Pulmonary infection requires prompt treatment with antibiotics. Hypoxia necessitates giving low concentrations of oxygen. Patients should receive polyvalent pneumoccal vaccine (Pneumovax) and yearly influenza immunizations.

For patients with limited cutaneousSSc who develop isolated pulmonary arterial hypertension, treatment is limited. The usual treatment is supplemental oxygen, anticoagulation, and the administration of a vasodilator. A calcium channel blocker such as nifedipine lowers pulmonary arterial resistance and improves cardiac function, but in most patients this is only for a short period of time. Few patients survive more than 5 years. Heart-lung or single-lung transplantation may be a therapeutic option only in those patients without other significant systemic involvement. Current reports of intravenous epoprostenol (prostacyclin) in the treatment of SSc-associated pulmonary hypertension have been encouraging. Epoprostenol is infused continuously via a central line with a portable pump. Improvement in symptoms of right heart failure and exercise tolerance occurred. Also hemodynamic tests showed a decrease in the pulmonary vascular resistance and pulmonary artery pressure both in the short term and in a few patients after 1 or 2 years.

Recognition of early signs of renal hypertensive crisis is important in order to preserve renal function and prevent hypertensive encephalopathy. Renal involvement is often accompanied by hypertension and mild to moderate proteinuria. An occasional patient may be normotensive. Antihypertensive agents are often effective in lowering blood pressure and stabilizing or reversing renal failure. These drugs include propranolol, clonidine, and minoxidil. Particularly effective are the angiotensin-converting enzyme inhibitors, which include captopril, enalapril, and lisinopril. Dialysis may be required in patients with progressive renal failure. Some patients, however, have a slow return of renal function after several months and may no longer require dialysis. Patients are usually not candidates for kidney transplantation because of the other systemic manifestations ofSSc.

Patients with cardiac failure require careful monitoring of digitalis and diuretic administration. Noninflammatory pericardial effusions may also improve with diuretics. Care should be taken to avoid overdiuresis, which may lead to decreased renal blood flow, decreased cardiac output, and renal failure.

MIXED CONNECTIVE TISSUE DISEASE

MCTDis an overlap syndrome characterized by combinations of clinical features ofSLE(Chap. 311),SSc, polymyositis (Chap. 382), and rheumatoid arthritis (Chap. 312) and the presence of very high titers of circulating autoantibodies to nuclearRNPantigen. This antibody in high titer, now referred to as *anti-U_1RNP*, has been a justification for considering MCTD as a distinct clinical entity. MCTD has been challenged as a distinct disorder by those who consider it as a subset of SLE or scleroderma. Others prefer to classify MCTD as an undifferentiated connective tissue disease. MCTD occurs worldwide and in all races. The peak onset of disease is in the second and third decades, but MCTD is seen in children and the elderly. Women are predominantly affected. The pathogenic mechanisms in MCTD reflect the disorders making up this syndrome.

Clinical Features The presenting symptoms ofMCTD are most often Raynaud's phenomenon, puffy hands, arthralgias, myalgias, and fatigue. Occasionally, patients may present with the acute onset of high fever, polymyositis, arthritis, and neurologic features such as trigeminal neuralgia and aseptic meningitis. The various features of the connective tissue disorders making up MCTD develop over months and years.

The fingers as well as the entire hand may be puffy, followed later by sclerodactyly. Sclerodermal changes are usually limited to the distal extremities and sometimes the face but spare the trunk. Telangiectasia and calcinosis may develop. Some patients have mucocutaneous features ofSLEincluding a classic malar rash, photosensitivity, discoid lesions, alopecia, and painful oral ulcerations. An erythematous rash over the knuckles, elbows, and knees and heliotropic eyelids, typical of dermatomyositis, are uncommon.

Joint pain, stiffness, and swelling involving the peripheral joints occur frequently. Deformities of the hands similar to those of rheumatoid arthritis may develop but usually without bony erosions. A destructive polyarthritis is occasionally observed. Myalgias are a frequent symptom. Some patients develop typical symptoms of polymyositis with proximal muscle weakness, abnormal electromyographic findings, elevated levels of muscle enzymes, and inflammatory changes on muscle biopsy.

Approximately 85% of patients have pulmonary involvement, which is often asymptomatic. Diffusing capacity for carbon monoxide may be the only abnormality. Pleurisy commonly occurs but is seldom associated with large pleural effusions. Some patients develop interstitial lung disease. Pulmonary arterial hypertension is the most common cause of death inMCTD.

Approximately 25% of patients develop renal disease. Membranous glomerulonephritis is most common and usually mild but can cause nephrotic syndrome. Diffuse proliferative glomerulonephritis is unusual inMCTD, perhaps because of the protective role believed to be played by the high titers of anti-U$_1$RNP. Renal crisis secondary to malignant renovasculature hypertension, as occurs in scleroderma, is seen in a few patients.

Gastrointestinal involvement is seen in ~70% of patients. The most common manifestations are esophageal dysmotility, lower esophageal sphincter laxity, and gastroesophageal reflux. Bowel manifestations mimic those of scleroderma bowel disease.

Pericarditis occurs in 30% of patients. Other cardiac features include myocarditis, arrhythmia, conduction disturbances, and mitral valve prolapse. Other clinical features ofMCTDinclude trigeminal neuropathy, peripheral neuropathy, aseptic meningitis, lymphadenopathy, and Sjogren's syndrome. The majority of patients have developed, or will develop within 5 years of presentation, diagnostic clinical criteria for one of the overlapping connective tissue diseases, most oftenSLE orSSc.

Laboratory Findings Anemia of chronic inflammation is seen in the majority of patients.

A positive direct Coombs' test is found in many patients, but hemolytic anemia is unusual. Leukopenia, thrombocytopenia, or both are present in some patients. Hypergammaglobulinemia is common, and rheumatoid factor is present in 50% of patients.

All patients, by definition ofMCTD, have antibodies to U_1RNP. The specificity of this antibody is to the 70-kDa protein complexed to small nuclear RNA. The anti-U_1RNP antibodies are associated with HLA-DR4 but not with -DR2 and -DR3 as found in SLE. Molecular mimicry has been demonstrated between U_1 RNP and retroviral antigens by some laboratories.

TREATMENT

The treatment ofMCTD is essentially the same as would be indicated for the respective connective tissue diseases defining this syndrome. More than half the patients have a favorable course. The 10-year survival rate overall is approximately 80% but varies depending on the connective tissue disease that may eventually develop.

EOSINOPHILIC FASCIITIS

Eosinophilic fasciitis is a scleroderma-like syndrome of unknown cause characterized by inflammation followed later by sclerosis of the dermis, subcutis, and deep fascia. The disease affects adults and often occurs after strenuous physical activity. Patients do not have Raynaud's phenomenon or internal organ involvement. Several immunologic abnormalities have been associated with eosinophilic fasciitis and include aplastic anemia, myelodysplastic syndrome, and thrombocytopenia. Patients usually have the abrupt onset of symmetric tenderness and swelling of the extremities, rapidly followed by induration of the skin and subcutaneous tissue. The skin takes on a cobblestone or puckered appearance. Carpal tunnel syndrome appears early in the course, and flexion contractures develop later. A low-grade myositis is often present, but creatinine kinase levels are usually normal. A marked eosinophilia is found in the early stage of disease and subsequently decreases. Increased levels of polyclonal IgG and immune complexes are often present in the serum. A full-thickness biopsy consisting of skin, fascia, and superficial muscle shows perivascular infiltration of histiocytes, eosinophils, lymphocytes, and plasma cells. Biopsies later in the course show sclerosis. Spontaneous improvement and occasionally complete remission may occur after 2 to 5 years of disease. Some patients have persistent disease, while others are left with flexion contractures. Administration of glucocorticoids may provide symptomatic improvement and will decrease the eosinophilia. Improvement has been reported with the use of the H_2blocker cimetidine.

EOSINOPHILIA-MYALGIA SYNDROME

In 1989, reports of patients with scleroderma-like skin changes, myalgias, and eosinophilia dramatically increased. Most, but not all, of these cases were associated with ingestion of L-tryptophan manufactured by a single Japanese company. Batches of L-tryptophan implicated inEMS were found to contain trace amounts of a contaminant identified as a dimer of L-tryptophan that appeared in 1988 after changes were made in the method of manufacturing this drug. It is not clear whether this chemical contaminant is the etiologic agent or whether another unidentified substance is responsible. *L-Tryptophan products were taken off the market in 1990*. The onset of EMS can be either abrupt or insidious. In the early phases of the disease, clinical manifestations include low-grade fever, fatigue, dyspnea, cough, arthralgias/arthritis, evanescent erythematous rashes, muscle cramping, and severe myalgias. Pulmonary infiltrates may be present. Over the next 2 to 3 months, scleroderma-like skin changes appear. Some patients develop a peripheral neuropathy, which may persist. An ascending polyneuropathy may lead to paralysis and respiratory failure requiring ventilatory assistance. Cognitive dysfunction with impairment of memory and concentration has been recognized in this syndrome. Myocarditis and cardiac arrhythmias occur in some patients, and a few patients develop pulmonary hypertension. Approximately a third of patients have features of eosinophilic fasciitis. EMS most closely resembles toxic oil syndrome; however, Raynaud's phenomenon does not occur, and there is a lower prevalence of pulmonary hypertension and thromboembolic disease. The peripheral eosinophil count is >1000/uL in most patients. The histologic findings on biopsy of skin, fascia, and superficial muscle are similar to those found in eosinophilic fasciitis. The clinical features of EMS may persist after L-tryptophan has been discontinued. EMS may run a chronic course, and response to therapy has been variable. Treatment has included glucocorticoids, antimalarial drugs, immunosuppressive drugs, and plasmapheresis. Prednisone was beneficial during the acute inflammatory phase of the disease in the majority of patients and resulted in resolution of pulmonary infiltrates, peripheral edema, and eosinophilia. In the later phase of the illness, no treatment was found to be of particular value. The pathogenesis of this disease is not known. A follow-up of patients 2 years after their onset of illness showed that most symptoms and physical findings had resolved or improved except for cognitive dysfunction, which became worse in approximately one-third of the patients, and peripheral neuropathy, which remained unchanged (Chap. 382).

(Bibliography omitted in Palm version)

Back to Table of Contents

314. SJOGREN'S SYNDROME - *Haralampos M. Moutsopoulos*

DEFINITION

Sjogren's syndrome is a chronic, slowly progressive autoimmune disease characterized by lymphocytic infiltration of the exocrine glands resulting in xerostomia and dry eyes. Approximately one-third of patients present with systemic manifestations. A small but significant number of the patients may develop malignant lymphoma. The disease can be seen alone (primary Sjogren's syndrome) or in association with other autoimmune rheumatic diseases (secondary Sjogren's syndrome) (Table 314-1).

INCIDENCE AND PREVALENCE

The disease affects predominantly middle-aged women (female-to-male ratio 9:1), although it occurs in all ages, including childhood. The prevalence of primary Sjogren's syndrome is approximately 0.5 to 1.0%. In addition, 30% of patients with autoimmune rheumatic diseases suffer from secondary Sjogren's syndrome.

PATHOGENESIS

Sjogren's syndrome is characterized by lymphocytic infiltration of the exocrine glands and B lymphocyte hyperreactivity, as illustrated by circulating autoantibodies. The latter is accompanied by an oligomonoclonal B cell process, which is characterized by serum and urine monoclonal light chains and cryoprecipitable monoclonal immunoglobulins.

Sera of patients with Sjogren's syndrome often contain a number of autoantibodies directed against non-organ-specific antigens such as immunoglobulins (rheumatoid factors) and extractable nuclear and cytoplasmic antigens (Ro/SS-A, La/SS-B). Ro/SS-A autoantigen consists of three polypeptide chains (52, 54, and 60 kDa) in conjunction with RNAs, whereas the 48-kDa La/SS-B protein is bound to RNA III polymerase transcripts. The presence of autoantibodies to Ro/SS-A and La/SS-B antigens in Sjogren's syndrome is associated with earlier disease onset, longer disease duration, salivary gland enlargement, severity of lymphocytic infiltration of minor salivary glands, and certain extraglandular manifestations such as lymphadenopathy, purpura, and vasculitis. Antibodies toa-fordin (120 kDa), a salivary gland-specific protein, have recently been found in sera of patients with Sjogren's syndrome but not in sera of patients with other connective tissue diseases.

Phenotypic and functional studies have shown that the predominant cell infiltrating the affected exocrine glands is the helper/inducer T cell with characteristics of memory cells. Both B and T infiltrating lymphocytes are activated, as illustrated by production of immunoglobulins with autoantibody activity, spontaneous release of interleukin 2, and expression on the T cell surface of activation markers such as class II HLA as well as costimulatory molecules and lymphocyte function-associated antigen 1. Macrophages and natural killer cells are rarely detected in infiltrates, while epithelial cells of the affected glands inappropriately express class II molecules and possess messages for c-*myc* protooncogene and proinflammatory cytokines. All these phenomena suggest that the epithelial cell of the exocrine glands in Sjogren's syndrome may act as an antigen-presenting cell. In contrast to infiltrating lymphocytes, these cells undergo apoptotic death, resulting in exocrine gland dysfunction.

Immunogenetic studies have demonstrated that HLA-B8, -DR3, and -DRw52 are prevalent in patients with primary Sjogren's syndrome as compared with the normal control population. Molecular analysis of HLA class II genes has revealed that patients with Sjogren's syndrome, regardless of their ethnic origin, are highly associated with the HLA DQA1*0501 allele.

CLINICAL MANIFESTATIONS

The majority of the patients with Sjogren's syndrome have symptoms related to diminished lacrimal and salivary gland function. In most patients, the primary syndrome runs a slow and benign course. The initial manifestations can be mucosal dryness or nonspecific, and 8 to 10 years elapse from the initial symptoms to full-blown development of the disease.

The principal oral symptom of Sjogren's syndrome is dryness (xerostomia). Patients complain of difficulty in swallowing dry food, inability to speak continuously, a burning sensation, increase in dental caries, and problems in wearing complete dentures. Physical examination shows a dry, erythematous, sticky oral mucosa. There is atrophy of the filiform papillae on the dorsum of the tongue, and saliva from the major glands is either not expressible or is cloudy. Enlargement of the parotid or other major salivary glands occurs in two-thirds of patients with primary Sjogren's syndrome but is uncommon in those with the secondary syndrome. Diagnostic tests include sialometry, sialography, and scintigraphy. The labial minor salivary gland biopsy permits histopathologic confirmation of the focal lymphocytic infiltrates.

Ocular involvement is the other major manifestation of Sjogren's syndrome. Patients usually complain of dry eyes, with a sandy or gritty feeling under the eyelids. Other symptoms include burning, accumulation of thick strands at the inner canthi, decreased tearing, redness, itching, eye fatigue, and increased photosensitivity. These symptoms are attributed to the destruction of corneal and bulbar conjunctival epithelium, defined as keratoconjunctivitis sicca. Diagnostic evaluation of keratoconjunctivitis sicca includes measurement of tear flow by Schirmer's I test and tear composition as assessed by the tear breakup time or tear lysozyme content. Slit-lamp examination of the cornea and conjunctiva after rose Bengal staining reveals punctate corneal ulcerations and attached filaments of corneal epithelium.

Involvement of other exocrine glands occurs less frequently and includes a decrease in mucous gland secretions of the upper and lower respiratory tree, resulting in dry nose, throat, and trachea (xerotrachea), and diminished secretion of the exocrine glands of the gastrointestinal tract, leading to esophageal mucosal atrophy, atrophic gastritis, and subclinical pancreatitis. Dyspareunia due to dryness of the external genitalia and dry skin also may occur.

Extraglandular (systemic) manifestations are seen in one-third of patients with Sjogren's syndrome (Table 314-2), while they are very rare in patients with Sjogren's syndrome associated with rheumatoid arthritis. These patients complain more often of easy fatigability, low-grade fever, Raynaud's phenomenon, myalgias, and arthralgias. Most patients with primary Sjogren's syndrome experience at least one episode of nonerosive arthritis during the course of their disease. Manifestations of pulmonary involvement are frequent but rarely important clinically. Dry cough is the major manifestation that is attributed to small airway disease. Renal involvement includes interstitial nephritis, clinically manifested by hypostenuria and renal tubular dysfunction with or without acidosis. Untreated acidosis may lead to nephrocalcinosis. Glomerulonephritis is a rare finding that occurs in patients with systemic vasculitis, cryoglobulinemia, or systemic lupus erythematosus overlapping with Sjogren's syndrome. Vasculitis affects small and medium-sized vessels. The most common clinical features are purpura, recurrent urticaria, skin ulcerations, glomerulonephritis, and mononeuritis multiplex. Sensorineural hearing loss was found in one-half of patients with Sjogren's syndrome and correlated with the presence of anticardiolipin antibodies.

It has been suggested that primary Sjogren's syndrome with vasculitis also may present with multifocal, recurrent, and progressive nervous system disease, such as hemiparesis, transverse myelopathy, hemisensory deficits, seizures, and movement disorders. Aseptic meningitis and multiple sclerosis also have been reported in these patients.

Lymphoma is a well-known manifestation of Sjogren's syndrome that usually presents later in the illness. Persistent parotid gland enlargement, lymphadenopathy, cutaneous vasculitis, peripheral neuropathy, lymphopenia, and cryoglobulinemia are manifestations suggesting the development of lymphoma. Most lymphomas are extranodal, marginal zone B cell, and low grade. Salivary glands are the most common site of involvement.

Routine laboratory tests reveal mild normochromic, normocytic anemia. An elevated erythrocyte sedimentation rate is found in approximately 70% of patients.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

A European multicenter study has developed diagnostic criteria of Sjogren's syndrome (Table 314-3), which have been validated and present high specificity and sensitivity. A diagnostic algorithm is depicted inFig. 314-1.

The differential diagnosis of Sjogren's syndrome includes other conditions that may cause dry mouth or eyes or parotid salivary gland enlargement (Table 314-4). Infections with HIV and hepatitis C virus (Chap. 309) and sarcoidosis (Chap. 318) appear to produce a clinical picture indistinguishable from that of Sjogren's syndrome (Table 314-5).

TREATMENT

Sjogren's syndrome remains fundamentally an incurable disease. Hence treatment is aimed at symptomatic relief and limiting the damaging local effects of chronic xerostomia and keratoconjunctivitis sicca by substitution of the missing secretions.

The sicca complex is treated with fluid replacement supplied as often as necessary. To replace deficient tears, there are several readily available ophthalmic preparations (Tearisol; Liquifilm; 0.5% methylcellulose; Hypo Tears). It may be necessary for severely affected patients to use these preparations as often as every 30 min. If corneal ulceration is present, eye patching and boric acid ointments are recommended. Certain drugs that may increase lacrimal and salivary hypofunction such as diuretics, antihypertensive drugs, and antidepressants should be avoided. Propionic acid gels may be used to treat vaginal dryness.

Pilocarpine (5 mg thrice daily) given orally appears to improve sicca manifestations. Hydroxychloroquine (200 mg/day) is helpful for arthralgias. Glucocorticoids (1 mg/kg per day) or other immunosuppressive agents (i.e., cyclophosphamide) are indicated for the treatment of extraglandular manifestations, particularly when renal or severe pulmonary involvement and systemic vasculitis have been documented.

(Bibliography omitted in Palm version)

Back to Table of Contents

315. ANKYLOSING SPONDYLITIS, REACTIVE ARTHRITIS, AND UNDIFFERENTIATED SPONDYLOARTHROPATHY - *Joel D. Taurog*, *Peter E. Lipsky*

The spondyloarthropathies are a group of disorders that share certain clinical features and an association with the HLA-B27 allele. These disorders include ankylosing spondylitis, Reiter's syndrome, reactive arthritis, psoriatic arthritis and spondylitis, enteropathic arthritis and spondylitis, juvenile-onset spondyloarthropathy, and undifferentiated spondyloarthropathy. The similarities in clinical manifestations and genetic predisposition suggest that these disorders share pathogenic mechanisms. Specific definitions and diagnostic criteria for the individual conditions will be provided in subsequent sections of this chapter.

ANKYLOSING SPONDYLITIS

Ankylosing spondylitis (AS) is an inflammatory disorder of unknown cause that primarily affects the axial skeleton; peripheral joints and extraarticular structures may also be involved. The disease usually begins in the second or third decade; the prevalence in men is approximately three times that in women. It is considered the prototype of the spondyloarthropathies. Older names include *Marie-Strumpell disease* or *Bechterew's disease*.

EPIDEMIOLOGY

ASshows a striking correlation with the histocompatibility antigen HLA-B27 and occurs worldwide roughly in proportion to the prevalence of this antigen (Chap. 306). In North American Caucasians, the general prevalence of B27 is 7%, whereas over 90% of patients with AS have inherited this antigen. The association with B27 is independent of disease severity.

In population surveys, 1 to 6% of adults inheriting B27 have been found to haveAS. In contrast, in families of patients with AS, the prevalence is 10 to 30% among adult first-degree relatives inheriting B27. The concordance rate in identical twins is estimated to exceed 65%. It is currently believed that susceptibility to AS is determined almost entirely by genetic factors, with as yet unidentified allelic genes in addition to B27 comprising about two-thirds of the genetic component and B27 itself comprising about one-third. AS is strongly associated with inflammatory bowel disease (IBD), including both ulcerative colitis and Crohn's disease. IBD is a risk factor for AS independent of HLA-B27, although 50 to 75% of patients with both AS and IBD are B27 positive.**See also Chap. 287.*

PATHOLOGY

The *enthesis*, the site of ligamentous attachment to bone, is thought to be the primary site of pathology inAS, particularly in the lesions around the pelvis and spine. Enthesitis is associated with prominent edema of the adjacent bone marrow and is often characterized by erosive lesions that eventually undergo ossification.

Sacroiliitis is usually one of the earliest manifestations ofAS, with features of both enthesitis and synovitis. The early lesions consist of subchondral granulation tissue containing lymphocytes, plasma cells, mast cells, macrophages, and chondrocytes; infiltrates of lymphocytes and macrophages in ligamentous and periosteal zones; and subchondral bone marrow edema. Synovitis follows and may progress to pannus formation. Islands of new bone formation can be found within the inflammatory infiltrates. Usually, the thinner iliac cartilage is eroded before the thicker sacral cartilage. The irregularly eroded, sclerotic margins of the joint are gradually replaced by fibrocartilage regeneration and then by ossification. Ultimately, the joint may be totally obliterated. This progression is evident by imaging techniques (see below).

In the spine, early in the process there is inflammatory granulation tissue at the junction of the annulus fibrosus of the disk cartilage and the margin of vertebral bone. The outer annular fibers are eroded and eventually replaced by bone, forming the beginning of a bony excrescence called a *syndesmophyte*, which then grows by continued enchondral ossification, ultimately bridging the adjacent vertebral bodies. Ascending progression of this process leads to the "bamboo spine" observed radiographically. Other lesions in the spine include diffuse osteoporosis, erosion of vertebral bodies at the disk margin, "squaring" of vertebrae, and inflammation and destruction of the disk-bone border. Inflammatory arthritis of the apophyseal joints is common, with erosion of cartilage by pannus, often followed by bony ankylosis.

Bone mineral density is significantly diminished in the spine and proximal femur early in the course of the disease, before the advent of significant immobilization. The mechanism for this is not known.

Peripheral arthritis inAS can show synovial hyperplasia, lymphoid infiltration, and pannus formation, but the process lacks the exuberant synovial villi, fibrin deposits, ulcers, and accumulations of plasma cells seen in rheumatoid arthritis (Chap. 312). Central cartilaginous erosions caused by proliferation of subchondral granulation tissue are common in AS but rare in rheumatoid arthritis.

Acute anterior uveitis (iritis) occurs in at least 20% of patients withAS. Few cases have been studied histologically, none at an early stage. After recurrent attacks, the iris shows nonspecific inflammatory changes, scarring, increased vascularity, and pigment-laden macrophages. Pupillary synechiae and cataract formation are common sequelae.

Aortic insufficiency develops in a small percentage of cases. There is thickening of the aortic valve cusps and the aorta near the sinuses of Valsalva, with dense adventitial scar tissue and intimal fibrous proliferation. The scar tissue can extend into the ventricular septum with resultant heart block.

Microscopic inflammatory lesions of the colon and ileocecal valve have been found in 25 to 50% of patients withAS, even in those lacking any clinical evidence ofIBD. IgA nephropathy has been reported with increased frequency.

PATHOGENESIS

The pathogenesis ofAS is incompletely understood. A number of features of the disease implicate immune-mediated mechanisms, including elevated serum levels of IgA and acute-phase reactants, inflammatory histology, and close association with HLA-B27. The inflamed sacroiliac joint is infiltrated with CD4+ and CD8+ T cells and macrophages and shows high levels of tumor necrosis factora. Transforming growth factorb is detectable near the sites of new bone formation. No specific event or exogenous agent that triggers the onset of disease has been identified, although overlapping features with reactive arthritis andIBDsuggest that enteric bacteria may play a role. Elevated serum titers of antibodies to certain enteric bacteria, particularly *Klebsiella pneumoniae*, are common in AS patients, but no role for these antibodies in the pathogenesis of AS has been identified. Evidence that B27 plays a direct role is provided by the finding that rats transgenic for B27 spontaneously develop spondylitis, along with colitis, peripheral arthritis, and other lesions characteristic of the spondyloarthropathies (see below).

Some evidence has accumulated for autoimmunity to the cartilage proteoglycan aggrecan, and particularly its G1 globulin domain and link protein.ASpatients have been found to have cellular immunity to these molecules, and mice immunized with the G1 domain develop spondylitis and discitis. Sharing of proteoglycan antigenic epitopes among the pathologic sites in the skeleton, uveal tract, and aorta in AS suggests a possible explanation for the distribution of pathologic sites in AS.

CLINICAL MANIFESTATIONS

The symptoms of the disease are usually first noticed in late adolescence or early adulthood; the median age in western countries is 23 in both genders. In 5% of patients, symptoms begin after age 40. The initial symptom is usually dull pain, insidious in onset, felt deep in the lower lumbar or gluteal region, accompanied by low-back morning stiffness of up to a few hours' duration that improves with activity and returns following periods of inactivity. Within a few months of onset, the pain has usually become persistent and bilateral. Nocturnal exacerbation of pain that forces the patient to rise and move around may be frequent.

In some patients bony tenderness may accompany back pain or stiffness, while in others it may be the predominant complaint. Common sites include the costosternal junctions, spinous processes, iliac crests, greater trochanters, ischial tuberosities, tibial tubercles, and heels. Occasionally, bony chest pain is the presenting complaint. Arthritis in the hips and shoulders ("root" joints) occurs in 25 to 35% of patients, in many cases early in the disease course. Arthritis of peripheral joints other than the hips and shoulders, usually asymmetric, occurs in up to 30% of patients and can occur at any stage of the disease. Neck pain and stiffness from involvement of the cervical spine are usually relatively late manifestations. Occasional patients, particularly in the older age group, present with predominantly constitutional symptoms such as fatigue, anorexia, fever, weight loss, or night sweats.

ASoften has a juvenile onset in developing countries. In these individuals, peripheral arthritis and enthesitis usually predominate, with axial symptoms supervening in late adolescence.

The most common extraarticular manifestation is acute anterior uveitis, which can antedate the spondylitis. Attacks are typically unilateral, causing pain, photophobia, and increased lacrimation. These tend to recur, often in the opposite eye. Cataracts and secondary glaucoma are not uncommon sequelae. Aortic insufficiency, sometimes producing symptoms of congestive heart failure, occurs in a few percent of patients, occasionally early in the course of the spinal disease. Third-degree heart block may occur alone or together with aortic insufficiency. The block is in the atrioventricular node in 95% of cases. Up to half the patients have inflammation in the colon or ileum. This is usually asymptomatic, but in 5 to 10% of patients with AS, frankIBD will develop.

Initially, physical findings mirror the inflammatory process. The most specific findings involve loss of spinal mobility, with limitation of anterior and lateral flexion and extension of the lumbar spine and of chest expansion. Limitation of motion is usually out of proportion to the degree of bony ankylosis, reflecting muscle spasm secondary to pain and inflammation. Pain in the sacroiliac joints may be elicited either with direct pressure or with maneuvers that stress the joints, but these techniques are unreliable in discriminating inflammatory sacroiliitis. In addition, there is commonly tenderness upon palpation at the sites of symptomatic bony tenderness and paraspinous muscle spasm.

The Schober test is a useful measure of flexion of the lumbar spine. The patient stands erect, with heels together, and marks are made directly over the spine 5 cm below and 10 cm above the lumbosacral junction (identified by a horizontal line between the posterosuperior iliac spines.) The patient then bends forward maximally, and the distance between the two marks is measured. The distance between the two marks increases 5 cm or more in the case of normal mobility and less than 4 cm in the case of decreased mobility. Chest expansion is measured as the difference between maximal inspiration and maximal forced expiration in the fourth intercostal space in males or just below the breasts in females. Normal chest expansion is 5 cm or greater.

Limitation or pain with motion of the hips or shoulders is usually present if either of these joints is involved. Careful examination is also necessary to detect inflammatory disease of peripheral joints. It should be emphasized that early in the course of mild cases, symptoms may be subtle and nonspecific, and the physical examination may be completely normal.

The course of the disease is extremely variable, ranging from the individual with mild stiffness and radiographically equivocal sacroiliitis to the patient with a totally fused spine and severe bilateral hip arthritis, possibly accompanied by severe peripheral arthritis and extraarticular manifestations. Pain tends to be persistent early in the disease and then to become intermittent, with alternating exacerbations and quiescent periods. In a typical severe untreated case with progression of the spondylitis to syndesmophyte formation, the patient's posture undergoes characteristic changes. The lumbar lordosis is obliterated with accompanying atrophy of the buttocks. The thoracic kyphosis is accentuated. If the cervical spine is involved, there may be a forward stoop of the neck. Hip involvement with ankylosis may lead to flexion contractures, compensated by flexion at the knees. The progression of the disease may be followed by measuring the patient's height, chest expansion, Schober test, and occiput-to-wall distance when the patient stands erect with the heels and back flat against the wall. Occasional individuals are encountered with advanced physical findings suggestive of long-standingAS who report having never had significant symptoms.

In some but not all studies, onset of the disease in adolescence correlates with a worse prognosis, but there is general agreement that early severe hip involvement is an indication of progressive disease. The disease in women tends to progress less frequently to total spinal ankylosis, although there is some evidence for an increased prevalence of isolated cervical ankylosis and peripheral arthritis in women. In industrialized countries, peripheral arthritis (distal to hips and shoulders) occurs overall in about 25% of patients, usually as a late manifestation, whereas in developing countries, the prevalence is much higher, with onset typically early in the disease course. Pregnancy has no consistent effect onAS, with symptoms improving, remaining the same, or deteriorating in about one-third of pregnant patients, respectively.

The most serious complication of the spinal disease is spinal fracture, which can occur with even minor trauma to the rigid, osteoporotic spine. The cervical spine is most commonly involved. These fractures are often displaced and cause spinal cord injury. Cauda equina syndrome and slowly progressive upper pulmonary lobe fibrosis are rare complications of long-standingAS. The prevalence of aortic insufficiency and of cardiac conduction disturbances, including third-degree heart block, increases with prolonged disease. Subclinical pulmonary lesions and cardiac dysfunction may be relatively common. Prostatitis has been reported to have an increased prevalence in men with AS. Amyloidosis is only rarely associated (Chap. 319).

Several validated measures of disease activity and functional outcome have recently been developed forAS. Despite the persistence of the disease, most patients remain gainfully employed. The effect of AS on survival is controversial. Some, but not all, studies have suggested that AS shortens life span, compared with the general population. Mortality attributable to AS is largely the result of spinal trauma, aortic insufficiency, respiratory failure, amyloid nephropathy, or complications of therapy such as upper gastrointestinal hemorrhage.

LABORATORY FINDINGS

No laboratory test is diagnostic of AS. In most ethnic groups, the HLA-B27 gene is present in approximately 90% of patients with AS. Most, but not all, patients with active disease have an elevated erythrocyte sedimentation rate and an elevated level of C-reactive protein. A mild normochromic, normocytic anemia may be present. Patients with severe disease may show an elevated alkaline phosphatase level. Elevated serum IgA levels are common. Rheumatoid factor and antinuclear antibodies are largely absent unless caused by a coexistent disease. Synovial fluid from inflamed peripheral joints in AS is not distinctly different from that of other inflammatory joint diseases. In cases with restriction of chest wall motion, decreased vital capacity and increased functional residual capacity are common, but airflow measurements are normal and ventilatory function is usually well maintained.

RADIOGRAPHIC FINDINGS

Radiographically demonstrable sacroiliitis is usually present inAS. The earliest changes in the sacroiliac joints demonstrable by standard radiography are blurring of the cortical margins of the subchondral bone, followed by erosions and sclerosis. Progression of the erosions leads to "pseudowidening" of the joint space; as fibrous and then bony ankylosis supervene, the joints may become obliterated radiographically. The changes and progression of the lesions are usually symmetric.

Roentgenographic abnormalities generally appear in the sacroiliac joints before appearing elsewhere in the spine. In the lumbar spine, progression of the disease leads to straightening, caused by loss of lordosis, and reactive sclerosis, caused by osteitis of the anterior corners of the vertebral bodies with subsequent erosion, leading to "squaring" of the vertebral bodies. Progressive ossification of the superficial layers of the annulus fibrosus leads to eventual formation of marginal syndesmophytes, visible on plain films as bony bridges connecting successive vertebral bodies anteriorly and laterally.

In mild cases, years may elapse before unequivocal sacroiliac abnormalities are evident on plain radiographs. Computed tomography (CT) and magnetic resonance imaging (MRI) can detect abnormalities reliably at an earlier stage than plain radiography. MRI has emerged as a highly sensitive and specific technique for identifying early intraarticular inflammation, cartilage changes, and underlying bone marrow edema in sacroiliitis (Fig. 315-1). In suspected cases in which conventional radiography does not reveal definite sacroiliac abnormalities or is undesirable (e.g., in young women or children), dynamic MRI is the procedure of choice for establishing a diagnosis of sacroiliitis.

Reduced bone mineral density can be detected by dual-energy x-ray absorptiometry of the femoral neck and the lumbar spine. Falsely elevated readings related to spinal ossification can be avoided by using a lateral projection of the L3 vertebral body.

DIAGNOSIS

The diagnosis of earlyASbefore the development of irreversible deformity can be difficult to establish. Currently, modified New York criteria (1984) are widely used for diagnosis. These consist of the following: (1) a history of inflammatory back pain (see below); (2) limitation of motion of the lumbar spine in both the sagittal and frontal planes; (3) limited chest expansion, relative to standard values for age and sex; and (4) definite radiographic sacroiliitis. Using these criteria, the presence of radiographic sacroiliitis plus any one of the other three criteria is sufficient for a diagnosis of definite AS. These criteria may need to be further modified to include sacroiliitis demonstrated byMRI to increase their sensitivity.

The presence of B27 is neither necessary nor sufficient for the diagnosis, but the B27 test can be helpful in patients with suggestive clinical findings who have not yet developed radiographic sacroiliitis. Moreover, the absence of B27 in a typical case ofASsignificantly increases the probability of coexistentIBD.

ASmust be differentiated from numerous other causes of low-back pain, some of which are far more common than AS. The inflammatory back pain of AS is usually distinguished by the following five features: (1) age of onset below 40, (2) insidious onset, (3) duration greater than 3 months before medical attention is sought, (4) morning stiffness, and (5) improvement with exercise or activity. The most common causes of back pain other than AS are primarily mechanical or degenerative rather than inflammatory and do not show these features. Less common metabolic, infectious, and malignant causes of back pain also must be differentiated from AS. Ochronosis can produce a phenotype that is clinically and radiographically similar to AS.

Marked calcification and ossification of paraspinous ligaments occur in *diffuse idiopathic skeletal hyperostosis* (DISH). Although DISH is often categorized as a variant of osteoarthritis, diarthrodial joints are not involved. Ligamentous calcification and ossification are usually most prominent in the anterior spinal ligament and give the appearance of "flowing wax" on the anterior bodies of the vertebrae. However, a radiolucency may be seen between the newly deposited bone and the vertebral body, differentiating DISH from the marginal osteophytes in spondylosis. Intervertebral disk spaces are preserved, and sacroiliac and apophyseal joints appear normal, helping to differentiate DISH from spondylosis and fromAS, respectively.

DISHoccurs in the middle-aged and the elderly and is more common in men than in women. Patients are frequently asymptomatic but may have stiffness. Radiographic changes are generally much more severe than might be predicted from the mild symptoms caused by DISH.

TREATMENT

There is no definitive treatment forAS. The principal goal of management is the conscientious participation by the patient in an exercise program designed to maintain functional posture and to preserve range of motion. There is evidence that exercise increases mobility and improves function. The proportion of patients wtih severe deformity has decreased markedly in recent decades, probably because of earlier diagnosis and widespread use of physical therapy. Smoking has been associated with a poor outcome and should be emphatically discouraged. Most patients require anti-inflammatory agents to achieve sufficient symptomatic relief to be able to remain functional and carry out the exercise program. It is not known whether drug treatment alone can alter the progression of the disease.

Several nonsteroidal anti-inflammatory drugs (NSAIDs) have proved effective in reducing the pain and stiffness ofAS and are commonly used. Indomethacin is particularly effective as a 75-mg slow-release preparation taken once or twice daily. Although phenylbutazone, at doses of 200 to 400 mg/d, has been considered the most effective anti-inflammatory agent in AS, because of its greater potential for serious side effects such as aplastic anemia and agranulocytosis, its use in the United States is confined to patients with very severe disease whose symptoms do not respond at all to other agents. Recent controlled trials suggest that sulfasalazine, in doses of 2 to 3 g/d, is useful in reducing peripheral joint symptoms as well as reversing laboratory evidence of inflammation. Some studies have not shown it to benefit axial arthritis, and its effect on natural progression of the disease is unproven. The peripheral arthritis may also respond to the folic acid antagonist methotrexate.[1]No therapeutic role for gold, penicillamine, immunosuppressive drugs, or oral glucocorticoids has been documented in AS. Occasionally, intralesional or intraarticular glucocorticoid injections may be beneficial in patients with persistent enthesopathy or synovitis unresponsive to anti-inflammatory agents. Recent studies have suggested that symptomatic benefit can be achieved fromCT-guided glucocorticoid injections into the sacroiliac joints, but the effects are not sustained. Anecdotal benefit has been reported for diverse agents such as pamidronate, thalidomide, pulse intravenous methylprednisolone, and tumor necrosis factor a antagonists. Controlled trials of these and other agents are needed, since for many patients current therapy is inadequate even for control of pain and stiffness.

The most common indication for surgery in patients withAS is severe hip joint arthritis, the pain and stiffness of which are usually dramatically relieved by total hip arthroplasty. A small number of patients may benefit from surgical correction of extreme flexion deformities of the spine or of atlantoaxial subluxation.

Attacks of iritis are usually effectively managed with local glucocorticoid administration in conjunction with mydriatic agents, although systemic glucocorticoids or even immunosuppressive drugs may be required in some cases. Coexistent cardiac disease may require pacemaker implantation and/or aortic valve replacement.

[1] Azathioprine, methotrexate, and sulfasalazine have not been approved for this purpose by the U.S. Food and Drug Administration at the time of publication.

REACTIVE ARTHRITIS AND UNDIFFERENTIATED SPONDYLOARTHROPATHY

Reactive arthritis (ReA) refers to acute nonpurulent arthritis complicating an infection elsewhere in the body. In recent years, the term has been used primarily to refer to spondyloarthropathies following enteric or urogenital infections and occurring predominantly in individuals with the histocompatibility antigen HLA-B27. Included in this category is the constellation of clinical findings formerly commonly called *Reiter's syndrome*.**Other forms of reactive and infection-related arthritis not associated with B27 and showing a different spectrum of clinical features, such as rheumatic fever or Lyme disease, are discussed in Chaps. 235 and 176.*

HISTORIC BACKGROUND

The association of acute arthritis with episodes of diarrhea or urethritis has been recognized for centuries. A large number of cases during World Wars I and II focused attention on the triad of arthritis, urethritis, and conjunctivitis, which became known as Reiter's syndrome, often occurring with additional mucocutaneous lesions.

The identification of bacterial species capable of triggering the clinical syndrome and the finding that up to 85% of the patients possess the B27 antigen have led to the unifying concept ofReA as a clinical syndrome triggered by specific etiologic agents in a genetically susceptible host. A similar spectrum of clinical manifestations can be triggered by enteric infection with any of several *Shigella*, *Salmonella*, *Yersinia*, and *Campylobacter* species, by genital infection with *Chlamydia trachomatis*; and possibly by other agents as well. Although Reiter's syndrome can be said to represent one part of the spectrum of the clinical manifestations ofReA, particularly that induced by *Shigella* or *Chlamydia*, the term is now largely of historic interest only. Since most patients with spondyloarthropathy do not have the classic features of Reiter's syndrome, it has become customary to employ the term *reactive arthritis*, regardless of whether or not there is evidence for a triggering infection. For the purposes of this chapter, the use of ReA will be restricted to those cases of spondyloarthropathy in which there is at least presumptive evidence for a related antecedent infection. Patients with clinical features of ReA who lack both evidence of an antecedent infection and the classic findings of Reiter's syndrome (urethritis, arthritis, conjunctivitis) will be considered to have *undifferentiated spondyloarthropathy*, which is discussed at the end of this chapter.

EPIDEMIOLOGY

LikeAS,ReAoccurs predominantly in individuals who have inherited the B27 gene; in most series, 60 to 85% of patients are B27 positive. In epidemics of arthritogenic bacterial infection, e.g., *S. flexneri*, it has been estimated that ReA develops in ~20% of exposed B27-positive individuals. In families with multiple cases of AS or ReA, the two conditions have been said to "breed true," i.e., to be uncommonly found together within an individual family. Whether this is caused by genetic or environmental factors is not known. The disease is most common in individuals 18 to 40 years of age, but it can occur both in children over 5 years of age and in older adults.

The sex ratio inReAfollowing enteric infection is nearly 1:1, whereas venereally acquired ReA is predominantly confined to men. The overall prevalence and incidence of ReA are difficult to assess because of the variable prevalence of the triggering infections and genetic susceptibility factors in different populations. For example, in Olmsted County, MN, the incidence was estimated as 3.5 cases per 100,000 population per year. In contrast, in a population with a high rate of genitourinary and/or gastrointestinal infections such as urban homosexual and bisexual men, the prevalence may approach 1 per 1000.

A particularly severe form of peripheral spondyloarthropathy has been described in patients with AIDS (Chap. 309). Most of these patients are HLA-B27 positive, but HIV infection per se is not an independent risk factor for spondyloarthropathy.

PATHOLOGY

Synovial histology is similar to that of other inflammatory arthropathies. Enthesitis is a common clinical finding inReA; the histology of this lesion resembles that ofAS. Microscopic histopathologic evidence of inflammation has occasionally been noted in the colon and ileum of patients with postvenereal ReA, but much less commonly than in postenteric ReA. The skin lesions of keratoderma blenorrhagica, which is associated mainly with venereally acquired ReA, are histologically indistinguishable from psoriatic lesions.

ETIOLOGY AND PATHOGENESIS

The first bacterial infection noted to be causally related toReA was *S. flexneri*. An outbreak of shigellosis among Finnish troops in 1944 resulted in numerous cases of ReA. Of the four species *S. sonnei*, *S. boydii*, *S. flexneri*, and *S. dysenteriae*, *S. flexneri* has most often been implicated in cases of ReA, both sporadic and epidemic. *S. sonnei*, although responsible for the majority of cases of shigellosis in the United States, has only rarely been implicated in cases of ReA.

Other bacteria that have been definitively identified as triggers ofReAinclude several

Salmonella spp., *Y. enterocolitica*, *C. jejuni*, and *C. trachomatis*. There is suggestive evidence implicating several other microorganisms, including *Y. pseudotuberculosis*, *Clostridium difficile*, and *Ureaplasma urealyticum*. *Chlamydia pneumoniae*, a respiratory pathogen, has also recently been implicated in triggering ReA. There are also numerous isolated reports of acute arthritis preceded by other bacterial, viral, or parasitic infections, but whether the microorganisms involved are actual triggers of ReA remains to be determined.

It has not been determined whetherReAoccurs by the same pathogenic mechanism following infection with each of these microorganisms, nor has the mechanism been fully elucidated in the case of any one of the known bacterial triggers. Most, if not all, of the triggering organisms produce lipopolysaccharide (LPS) and share a capacity to attack mucosal surfaces, to invade host cells, and survive intracellularly. Antigens from *Chlamydia*, *Yersinia*, *Salmonella*, and *Shigella* have been shown to be present in the synovium and/or synovial fluid leukocytes of patients with ReA for long periods following the acute attack. In ReA triggered by *Y. enterocolitica*, bacterial LPS and heat shock protein antigens have been found in peripheral blood cells years after the triggering infection. In the case of *C. trachomatis*, synovial persistence of microbial DNA and RNA suggests the presence of viable organisms, despite uniform failure to culture the organism from these specimens. There is thus evidence that ReA, at least in some cases, may be a form of chronic infection, rather than solely "reactive." T cells that specifically respond to antigens of the inciting organism are typically found in inflamed synovium but not in peripheral blood of patients with ReA. These T cells are predominantly CD4+, but CD8+ B27-restricted bacteria-specific cytolytic T cells have also been isolated in *Yersinia*- and *C. trachomatis*-induced ReA. Specific peptide antigens from these organisms have been identified as dominant T cells epitopes. Unlike the synovial CD4 T cells in rheumatoid arthritis, which are predominantly of the T_H1 phenotype, those in ReA also show a T_H2 phenotype. It is likely that antigen-specific T cells play an important role in the pathogenesis of ReA, but the precise mechanisms remain to be determined.

The role of HLA-B27 inReA also remains to be determined. Transgenic rats with high expression of B27 spontaneously develop a multiple organ system inflammatory disease affecting the gut, peripheral and axial joints, male genital tract, and skin that resembles these human conditions clinically and histologically. When raised in a germ-free environment, the B27 rats do not develop gut or joint inflammation, but the skin and genital lesions are not prevented. These findings suggest that bacteria are necessary, and normal gut bacteria are sufficient, to induce B27-related joint inflammation. In both the rat and human diseases, it remains to be determined whether the primary process is an autoimmune response against host tissues or an immune response against antigens of the triggering organism that have disseminated to the target tissues, and the specific role of B27 itself remains to be determined. A potentially very informative converse observation, in which humans develop a disease process resembling one first described in rats, is the recent finding that 0.4 to 0.8% of individuals treated with intravesicular bacillus Calmette-Guerin for bladder cancer develop reactive arthritis, and 60% of these patients are B27 positive. The process closely mimics adjuvant-induced arthritis in rats given complete Freund's adjuvant, first described over 40 years ago, which is currently thought to be mediated by CD4+ T cells specific for mycobacterial heat shock protein.

An intriguing in vitro finding indicates that the presence of HLA-B27 significantly prolongs the intracellular survival of *Y. enterocolitica*, and *S. enteritides* in human and mouse cell lines. A unifying hypothesis suggests that prolonged intracellular bacterial survival, promoted by B27, other factors, or both, permits trafficking of infected leukocytes from the site of primary infection to joints, where a T cell response to persistent bacterial antigens promotes arthritis. Evidence exists supporting each step of this scheme.

CLINICAL FEATURES

The clinical manifestations ofReAconstitute a spectrum that ranges from an isolated, transient monarthritis to severe multisystem disease. In the majority of cases, a careful history will elicit some evidence of an antecedent infection 1 to 4 weeks before the onset of symptoms of the reactive disease. However, in a sizable minority, no clinical or laboratory evidence of an antecedent infection can be found. In many cases of presumed venereally acquired reactive disease, there is a history of a recent new sexual partner, even in the absence of laboratory evidence of infection.

Constitutional symptoms are common, including fatigue, malaise, fever, and weight loss. The musculoskeletal symptoms are usually acute in onset. Arthritis is usually asymmetric and additive, with involvement of new joints occurring over a period of a few days to 1 to 2 weeks. The joints of the lower extremities, especially the knee, ankle, and subtalar, metatarsophalangeal, and toe interphalangeal joints, are the most common sites of involvement, but the wrist and fingers can be involved as well. The arthritis is usually quite painful, and tense joint effusions are not uncommon, especially in the knee. Dactylitis, or "sausage digit," a diffuse swelling of a solitary finger or toe, is a distinctive feature of bothReA and psoriatic arthritis (Chap. 324). It is not specific, however, in that it is also seen in polyarticular gout and sarcoidosis. Tendinitis and fasciitis are particularly characteristic lesions, producing pain at multiple insertion sites, especially the Achilles insertion, the plantar fascia, and sites along the axial skeleton. Spinal and low-back pain are quite common and may be caused by insertional inflammation, muscle spasm, acute sacroiliitis, or, presumably, arthritis in intervertebral articulations.

Urogenital lesions may occur throughout the course of the disease. In males, urethritis may be marked or relatively asymptomatic and may be either an accompaniment of the triggering infection or a result of the reactive phase of the disease. Prostatitis is also common. Similarly, in females, cervicitis or salpingitis may be caused either by the infectious trigger or by the sterile reactive process.

Ocular disease is common, ranging from transient, asymptomatic conjunctivitis to an aggressive anterior uveitis that occasionally proves refractory to treatment and may result in blindness.

Mucocutaneous lesions are frequent. Oral ulcers tend to be superficial, transient, and often asymptomatic. The characteristic skin lesions, *keratoderma blenorrhagica*, consist of vesicles that become hyperkeratotic, ultimately forming a crust before disappearing. They are most common on the palms and soles but may occur elsewhere as well. In patients with HIV infection, these lesions are often extremely severe and extensive, to the point of dominating the clinical picture (Chap. 309). Lesions on the glans penis, termed *circinate balanitis*, are common; these consist of vesicles that quickly rupture to form painless superficial erosions, which in circumcised individuals can form crusts similar to those of keratoderma blenorrhagica. Nail changes are common and consist of onycholysis, distal yellowish discoloration, and/or heaped-up hyperkeratosis.

Less frequent or rare manifestations ofReAinclude cardiac conduction defects, aortic insufficiency, central or peripheral nervous system lesions, and pleuropulmonary infiltrates.

Long-term follow-up studies suggest that some joint symptoms persist in 30 to 60% of patients withReA. Recurrences of the acute syndrome are common, and as many as 25% of patients either become unable to work or are forced to change occupations because of persistent joint symptoms. Chronic heel pain is often a particularly distressing symptom. Some aspects of ankylosing spondylitis are also common sequelae (see below). In some but not all studies, HLA-B27-positive patients have shown a worse outcome than B27-negative patients. The extent to which the long-term prognosis varies with different inciting agents is not known. However, patients with *Yersinia*-induced arthritis appear to have less chronic disease than those whose initial episode follows epidemic shigellosis.

LABORATORY AND RADIOGRAPHIC FINDINGS

The erythrocyte sedimentation rate is usually elevated during the acute phase of the disease. Mild anemia may be present, and acute-phase reactants tend to be increased. Synovial fluid is nonspecifically inflammatory, showing an elevated white cell count with a predominance of neutrophils. In most ethnic groups, 50 to 75% of the patients are B27 positive. It is unusual for the triggering infection to persist at the site of primary mucosal infection through the time of onset of the reactive disease, but it may occasionally be possible to culture the organism, e.g., in the case of *Yersinia*- or *Chlamydia*-induced disease. Serologic evidence of a recent infection may be present, such as a marked elevation of antibodies to *Yersinia*, *Salmonella*, or *Chlamydia*.

In early or mild disease, radiographic changes may be absent or confined to juxtaarticular osteoporosis. With long-standing persistent disease, marginal erosions and loss of joint space can be seen in affected joints. Periostitis with reactive new bone formation is characteristic of the disease, as it is with all the spondyloarthropathies. Spurs at the insertion of the plantar fascia are common.

Sacroiliitis and spondylitis may be seen as late sequelae. The sacroiliitis is more commonly asymmetric than inAS, and the spondylitis, rather than ascending symmetrically from the lower lumbar segments, can begin anywhere along the lumbar spine. The syndesmophytes may be coarse and nonmarginal, arising from the middle of a vertebral body, a pattern rarely seen in primary AS. Progression to spinal fusion as a sequela ofReA is uncommon.

DIAGNOSIS

ReAis a clinical diagnosis, there being no definitively diagnostic laboratory test or radiographic finding. The diagnosis should be entertained in any patient with an acute inflammatory, asymmetric, additive arthritis or tendinitis. The evaluation of such a patient should include careful questioning regarding possible antecedent triggering events such as an episode of diarrhea or dysuria. On physical examination, careful attention must be paid to the distribution of the joint and tendon involvement and to possible sites of extraarticular involvement, such as the eyes, mucous membranes, skin, nails, and genitalia. Synovial fluid aspiration and analysis may be helpful in excluding septic or crystal-induced arthritis. Culture or serology may help to identify a triggering infection. The role of molecular methods of microbial detection has not been established (see below).

Although typing for B27 is not needed to secure the diagnosis in clear-cut cases, it may have prognostic significance in terms of severity, chronicity, and the propensity for spondylitis and uveitis. Furthermore, it can be helpful diagnostically in atypical cases, a positive test increasing and a negative test decreasing the probability ofReA.

It is particularly important to differentiateReA from disseminated gonococcal disease, both of which can be venereally acquired and associated with urethritis (Chap. 147). Gonococcal arthritis and tenosynovitis tend to involve both upper and lower extremities equally, whereas in ReA lower extremity symptoms usually predominate. Back pain is common in ReA but is not a feature of gonococcal disease, whereas the vesicular skin lesions characteristic of disseminated gonococcal disease are not found in ReA. A positive gonococcal culture from the urethra or cervix does not exclude a diagnosis of ReA; however, culturing gonococci from blood, skin lesion, or synovium establishes the diagnosis of disseminated gonococcal disease. Polymerase chain reaction (PCR) technology has recently been used in the diagnosis of infections with *Neisseria gonorrheae* and with *C. trachomatis*. Occasionally, the only definitive way to distinguish the two is through a therapeutic trial of antibiotics.

ReAshares many features in common with psoriatic arthropathy, including the asymmetry of the arthritis, a propensity for "sausage digits" and nail involvement, an association with uveitis, and skin lesions of similar histology (Chap. 324). However, psoriatic arthritis is usually gradual in onset, the arthritis tends to affect primarily the upper extremities, and there is less associated periarthritis. Psoriatic arthritis is not associated with mouth ulcers or urethritis, or, usually, with bowel symptoms. Although psoriatic arthropathy shows some distinctive radiographic features that are not found in ReA, these occur only late in the disease and are of little help diagnostically. Only psoriatic spondylitis, not the peripheral arthritis, is associated with B27, about 50% of patients being positive. Occasional patients, usually B27 positive, following what appears to be a typical episode of ReA, will develop typical psoriasis and persistent arthritis such that the two entities become indistinguishable.

Undifferentiated spondyloarthropathy, or simply "spondyloarthropathy," is diagnosed in patients who lack evidence of an antecedent infection that might triggerReA and who do not meet criteria forAS but who show clinical features of these disorders.

TREATMENT

Most patients withReA are benefitted to some degree byNSAIDs, although rarely are symptoms of the acute arthritis completely ameliorated, and some patients fail to respond at all. Indomethacin, 75 to 150 mg/d in divided doses, is the initial treatment of choice. Other NSAIDs may be tried, with phenylbutazone, 100 mg tid or qid, being the NSAID of last resort, to be used only in severe, refractory cases because of its potentially serious side effects.

It is unclear whether antibiotics have a role in the therapy ofReA. One controlled study suggested that prolonged administration of a long-acting tetracycline may accelerate recovery from *Chlamydia*-induced ReA, but subsequent results have been less encouraging, and therapy for other bacterial triggers of ReA has shown little or no benefit. However, there is evidence that prompt, appropriate antibiotic treatment of acute chlamydial urethritis may prevent subsequent ReA. Currently, expert opinion supports the use of antibiotic therapy in established urogenital ReA but not in gastrointestinal ReA.

Two recent multicenter trials have suggested that sulfasalazine, up to 3 g/d in divided doses, may be beneficial to patients with persistentReA.1Patients with debilitating symptoms refractory toNSAIDand sulfasalazine therapy may respond to immunosuppressive agents such as azathioprine, 1 to 2 mg/kg per day, or to methotrexate, 7.5 to 15 mg per week. Systemic glucocorticoids are not generally recommended but in rare instances may be helpful in mobilizing a severely affected bedridden patient. Antimalarials, gold, and penicillamine are not useful in the treatment of ReA. Trials of new agents proven useful in rheumatoid arthritis, such as COX-2 inhibitors, leflunomide, and tumor necrosis factor a inhibitors, remain to be implemented.

Tendinitis and other enthesitic lesions occasionally may benefit from intralesional glucocorticoids. Uveitis may require aggressive treatment with glucocorticoids to prevent serious sequelae. Skin lesions ordinarily require only symptomatic treatment. In patients with HIV infection andReA, many of whom have severe skin lesions, the skin lesions in particular appear to respond to systemic treatment with anti-retroviral agents (Chap. 309). Cardiac complications are managed conventionally; management of neurologic complications is symptomatic.

Patients need to be educated about the nature of the disease and the factors that predispose to its recurrence. Comprehensive management includes counseling of patients in the avoidance of sexually transmitted disease and exposure to enteropathogens, as well as appropriate use of physical therapy, vocational counseling, and continued surveillance for long-term complications such as ankylosing spondylitis.

UNDIFFERENTIATED AND JUVENILE-ONSET SPONDYLOARTHROPATHY

It is not uncommon for clinicians to encounter patients, usually young adults, who do not haveIBD or psoriasis, lack evidence of an antecedent triggering infection, and do not have the classic triad of Reiter's syndrome or meet criteria for ankylosing spondylitis, who nonetheless present with some features of one or more of the spondyloarthropathies discussed above. For example, a patient may present with inflammatory synovitis of one knee, Achilles tendinitis, and dactylitis of one digit ("sausage digit"), or sacroiliitis in the absence of other criteria forAS. It is now common to consider such patients as having *undifferentiated spondyloarthropathy*, or simply *spondyloarthropathy*. Other terms for this condition have included *seronegative oligoarthritis*, *undifferentiated oligoarthritis*, and the now-outmoded *incomplete Reiter's syndrome*. There is strong evidence that some, perhaps most, of these patients haveReA in which the triggering infection remains clinically silent. In some other cases, the patient subsequently develops IBD or psoriasis or the process eventually meets criteria for ankylosing spondylitis. Approximately half the patients with undifferentiated spondyloarthropathy are HLA-B27 positive, and thus the absence of B27 is not useful in establishing or excluding the diagnosis.

In *juvenile-onset spondyloarthropathy*, which begins most commonly in boys (60 to 80%) between ages 7 and 16, an asymmetric, predominantly lower extremity oligoarthritis and enthesitis without extraarticular features is the typical mode of presentation. The prevalence of B27 in this condition, which has been termed the *SEA syndrome* (*s*eronegative, *e*nthesopathy, *a*rthropathy), is approximately 80%. Many, but not all, of these patients go on to develop typical ankylosing spondylitis in late adolescence or adulthood.

Management of undifferentiated spondyloarthropathy is similar to that of the other spondyloarthropathies, withNSAIDsand physical therapy forming the mainstays of treatment. Textbooks of pediatrics should be consulted for information on management of juvenile-onset spondyloarthropathy. An algorithm for the diagnosis of the spondyloarthropathies in adults is presented in Fig. 315-2.

(Bibliography omitted in Palm version)

Back to Table of Contents

316. BEHCET'S SYNDROME - *Haralampos M. Moutsopoulos*

DEFINITION

Behcet's syndrome is a multisystem disorder presenting with recurrent oral and genital ulcerations as well as ocular involvement. Internationally agreed diagnostic criteria have been proposed (Table 316-1).

PREVALENCE, PATHOGENESIS, AND PATHOLOGY

The disease has a worldwide distribution. The prevalence of Behcet's syndrome ranges from 1:10,000 in Japan to 1:500,000 in North America and Europe. It affects mainly young adults, with males having more severe disease than females.

The etiology and pathogenesis of this syndrome remain obscure; vasculitis is the main pathologic lesion with a tendency to venous thrombus formation, and circulating autoantibodies to human oral mucous membrane are found in approximately 50% of the patients. Familial occurrence has been sporadically reported; and in patients from eastern Mediterranean countries and Japan, the disease appears to be linked to HLA-B5 (B51) alloantigens.

CLINICAL FEATURES

The recurrent aphthous ulcerations are a sine qua non for the diagnosis. The ulcers are usually painful, shallow or deep with a central yellowish necrotic base, appear singly or in crops, and are located anywhere in the oral cavity. The ulcers persist for 1 to 2 weeks and subside without leaving scars. The genital ulcers resemble the oral ones.

Skin involvement includes folliculitis, erythema nodosum, an acne-like exanthem, and infrequently vasculitis. Nonspecific skin inflammatory reactivity to any scratches or intradermal saline injection (pathergy test) is a common and specific manifestation.

Eye involvement is the most dreaded complication, since it occasionally progresses rapidly to blindness. The eye disease is usually present at the onset but also may develop within the first few years. In addition to iritis, posterior uveitis, retinal vessel occlusions, and optic neuritis can be seen in some patients with the syndrome. Hypopyon uveitis, which is considered the hallmark of Behcet's syndrome, is in fact a rare manifestation.

The arthritis of Behcet's syndrome is not deforming and affects the knees and ankles.

Superficial or deep peripheral vein thrombosis is seen in one-fourth of the patients. Pulmonary emboli are a rare complication. The superior vena cava is obstructed occasionally, producing a dramatic clinical picture. Arterial involvement occurs infrequently and presents with aortitis or peripheral arterial aneurysm and arterial thrombosis. Pulmonary artery vasculitis presenting with dyspnea, cough, chest pain, hemoptysis, and infiltrates on chest roentgenograms has been reported recently in 5% of patients.

Central nervous system involvement is found more frequently in patients from northern Europe and the United States. The most common lesions are benign intracranial hypertension, a multiple sclerosis-like picture, pyramidal involvement, and psychiatric disturbances.

Gastrointestinal involvement is reported in patients from Japan and includes mucosal ulcerations of the gut.

Laboratory findings are mainly nonspecific indices of inflammation such as leukocytosis and elevated erythrocyte sedimentation rate as well as C-reactive protein levels; antibodies to human oral mucosa are also found.

TREATMENT

The severity of the syndrome usually abates with time. Apart from the patients with neurologic complications, the life expectancy seems to be normal, and the only serious complication is blindness.

Mucous membrane involvement may respond to topical glucocorticoids in the form of mouthwash or paste. In more serious cases thalidomide (100 mg/d) is effective. Thrombophlebitis is treated with aspirin, 325 mg/d. Colchicine or interferona can be beneficial for the arthritis of the syndrome. Uveitis and central nervous system involvement require systemic glucocorticoid therapy (prednisone, 1 mg/kg per day) and azathioprine, 2 to 3 mg/kg per day, or cyclosporine, 5 to 10 mg/kg per day. Early initiation of azathioprine tends to favorably affect the long-term prognosis of Behcet's syndrome.

(Bibliography omitted in Palm version)

Back to Table of Contents

317. THE VASCULITIS SYNDROMES - *Anthony S. Fauci*

DEFINITION

Vasculitis is a clinicopathologic process characterized by inflammation of and damage to blood vessels. The vessel lumen is usually compromised, and this is associated with ischemia of the tissues supplied by the involved vessel. A broad and heterogeneous group of syndromes may result from this process, since any type, size, and location of blood vessel may be involved. Vasculitis and its consequences may be the primary or sole manifestation of a disease; alternatively, vasculitis may be a secondary component of another primary disease. Vasculitis may be confined to a single organ such as the skin, or it may simultaneously involve several organ systems.

CLASSIFICATION OF VASCULITIC SYNDROMES

A major feature of the vasculitic syndromes as a group is the fact that there is a great deal of heterogeneity at the same time as there is considerable overlap among them. This has led to both difficulty and confusion with regard to the categorization of these diseases. The classification scheme listed inTable 317-1 takes into account this heterogeneity and overlap and will serve as a matrix to emphasize the fact that certain syndromes are predominantly systemic in nature and almost invariably lead to irreversible organ system dysfunction and even death if untreated, while others are usually localized to the skin and rarely result in irreversible dysfunction of vital organs. The distinguishing and overlapping features of the diseases listed inTable 317-1, which justify this classification scheme, will be discussed below.

PATHOPHYSIOLOGY AND PATHOGENESIS

Generally, most of the vasculitic syndromes are assumed to be mediated at least in part by immunopathogenic mechanisms (Table 317-2). However, evidence to this effect is for the most part indirect and may reflect epiphenomena as opposed to true causality.

Pathogenic Immune-Complex Formation Vasculitis is generally considered within the broader category of *immune-complex diseases* that include serum sickness and certain of the connective tissue diseases of which systemic lupus erythematosus (Chap. 311) is the prototype. Although deposition of immune complexes in vessel walls is the most widely accepted pathogenic mechanism of vasculitis, the causal role of immune complexes has not been clearly established in most of the vasculitic syndromes. Circulating immune complexes need not result in deposition of the complexes in blood vessels with ensuing vasculitis, and many patients with active vasculitis do not have demonstrable circulating or deposited immune complexes. The actual antigen contained in the immune complex has only rarely been identified in vasculitic syndromes. In this regard, hepatitis B antigen has been identified in both the circulating and deposited immune complexes in a subset of patients with systemic vasculitis, most notably within the polyarteritis nodosa group (see below). Essential mixed cryoglobulinemia has been associated with hepatitis C virus infection; hepatitis C virions and hepatitis C virus antigen-antibody complexes have been identified in the cryoprecipitates of these patients. An association between persistent parvovirus B19 infection and certain vasculitides has been reported; however, the pathogenic mechanisms related to this association are unclear.

The mechanisms of tissue damage in immune complex-mediated vasculitis resemble those described for serum sickness. In this model, antigen-antibody complexes are formed in antigen excess and are deposited in vessel walls whose permeability has been increased by vasoactive amines such as histamine, bradykinin, and leukotrienes released from platelets or from mast cells as a result of IgE-triggered mechanisms. The deposition of complexes results in activation of complement components, particularly C5a, which is strongly chemotactic for neutrophils. These cells then infiltrate the vessel wall, phagocytose the immune complexes, and release their intracytoplasmic enzymes, which damage the vessel wall. As the process becomes subacute or chronic, mononuclear cells infiltrate the vessel wall. The common denominator of the resulting syndrome is compromise of the vessel lumen with ischemic changes in the tissues supplied by the involved vessel.

Antineutrophil Cytoplasmic Antibodies (ANCA) ANCA are antibodies directed against certain proteins in the cytoplasm of neutrophils. They are present in a high percentage of patients with systemic vasculitis, particularly Wegener's granulomatosis, as well as in patients with microscopic polyangiitis and in patients with necrotizing and crescentic glomerulonephritis. There are two major categories of ANCA based on different targets for the antibodies. The terminology of *cytoplasmic* (c) *ANCA* refers to the diffuse, granular cytoplasmic staining pattern observed by immunofluorescence microscopy when serum antibodies bind to indicator neutrophils. Proteinase-3, the 29-kDa neutral serine proteinase present in neutrophil azurophilic granules is the major c-ANCA antigen. More than 90% of patients with typical Wegener's granulomatosis and active glomerulonephritis have a positive c-ANCA titer. The terminology of *perinuclear* (p) *ANCA* refers to the more localized perinuclear or nuclear staining pattern of the indicator neutrophils. The major target for p-ANCA is the enzyme myeloperoxidase; other targets of p-ANCA include elastase, cathepsin G, lactoferrin, lysozyme, and bactericidal/permeability-increasing protein. p-ANCA have been reported to occur in variable percentages of patients with microscopic polyangiitis, polyarteritis nodosa, Churg-Strauss syndrome, crescentic glomerulonephritis, and Goodpasture's syndrome as well as in association with nonvasculitic entities such as certain rheumatic and nonrheumatic autoimmune diseases, inflammatory bowel disease, certain drugs, and infections such as endocarditis and bacterial airway infections in patients with cystic fibrosis.

It is unclear why patients with these vasculitis syndromes developANCA, whereas ANCA are rare in other inflammatory diseases. However, once ANCA are present, there are a number of in vitro observations that suggest feasible mechanisms whereby these antibodies can contribute to the pathogenesis of the vasculitis syndromes. When neutrophils are in the resting state, proteinase-3 exists in the azurophilic granules of the cytoplasm, apparently inaccessible to serum antibodies. However, when neutrophils are primed by tumor necrosis factor (TNF)a or interleukin (IL)1, proteinase-3 translocates to the cell membrane where it can interact with extracellular ANCA. The neutrophils then degranulate and produce reactive oxygen species that can cause tissue damage. Endothelial cells also translocate their cytoplasmic proteinase-3 to the cell membrane upon priming with TNF-a, IL-1, or interferon (IFN)g, thus rendering them susceptible to interaction with ANCA and leading possibly to tissue damage due to complement-mediated cytotoxicity or antibody-dependent cellular cytotoxicity. Despite the attractiveness of these in vitro data, there is no conclusive evidence that ANCA are directly involved in the pathogenesis of the vasculitis syndromes, and they may represent merely an epiphenomenon; in fact, a number of clinical and laboratory observations argue against a primary pathogenic linkage. Patients may have vasculitis in the absence of ANCA; the absolute height of the antibody titers does not correlate well with disease activity; and patients with vasculitis, particularly Wegener's granulomatosis, in remission may continue to have high c-ANCA titers for years. Thus, their role in the pathogenesis of systemic vasculitis remains an open question.

Pathogenic T Lymphocyte Responses and Granuloma Formation In addition to the classic immune complex-mediated mechanisms of vasculitis as well asANCA, other immunopathogenic mechanisms may be involved in damage to vessels. The most prominent of these are delayed hypersensitivity and cell-mediated immune injury as reflected in the histopathologic feature of granulomatous vasculitis. However, immune complexes themselves may induce granulomatous responses. Vascular endothelial cells can express HLA class II molecules following activation by cytokines such as IFN-g. This allows these cells to participate in immunologic reactions such as interaction with CD4+ T lymphocytes in a manner similar to antigen-presenting macrophages. Endothelial cells can secreteIL-1 which may activate T lymphocytes and initiate or propagate in situ immunologic processes within the blood vessel. In addition, IL-1 andTNF-aare potent inducers of endothelial-leukocyte adhesion molecule 1 (ELAM-1) and vascular cell adhesion molecule 1 (VCAM-1), which may enhance the adhesion of leukocytes to endothelial cells in the blood vessel wall. Other mechanisms such as direct cellular cytotoxicity or antibody directed against vessel components or antibody-dependent cellular cytotoxicity have been suggested in certain types of vessel damage. However, there is no convincing evidence to support their causal contribution to the pathogenesis of any of the recognized vasculitic syndromes.

It is unknown why certain individuals develop vasculitis in response to certain antigenic stimuli, whereas others do not. However, it is likely that a number of factors are involved in the ultimate expression of a vasculitic syndrome. These include the genetic predisposition and the regulatory mechanisms associated with immune response to certain antigens. When immune complexes are involved in the pathogenic process, the ability of the reticuloendothelial system to clear circulating complexes from the blood, the size and physicochemical properties of immune complexes, the relative degree of turbulence of blood flow, the intravascular hydrostatic pressure in different vessels, and the preexisting integrity of the vessel endothelium likely explain why only certain types of immune complexes cause vasculitis and why the vasculitic process is selective for only certain vessels in individual patients.

Approach to the Patient

Given the heterogeneous nature of the vasculitis syndromes, workup of a patient with suspected vasculitis should follow a series of progressive steps that establish the diagnosis of vasculitis, determine where possible the category of the vasculitis syndrome (Table 317-1), and determine the pattern and extent of disease activity. This information should then be utilized to determine the choice of therapeutic options (Fig. 317-1). This approach is of considerable importance since several of the vasculitis syndromes require aggressive therapy with glucocorticoids and immunosuppressive agents, while other syndromes usually resolve spontaneously and require symptomatic treatment only. Vasculitis is often suspected on clinical and laboratory grounds (see individual syndromes below). Depending on the individual category of vasculitis, measurement ofANCAtiters may be helpful in this regard. However, a diagnosis of a vasculitis syndrome should not be made nor should treatment be initiated on the basis of a positive ANCA titer alone. The definitive diagnosis of vasculitis is made upon biopsy of involved tissue. The yield of "blind" biopsies of organs with no subjective or objective evidence of involvement is very low and should be avoided. When syndromes such as classic polyarteritis nodosa, Takayasu's arteritis, or the polyangiitis overlap syndrome are suspected, angiogram of organs with suspected involvement should be performed. However, angiograms should not be performed routinely when patients present with localized cutaneous vasculitis with no clinical indication of visceral involvement.

The constellation of clinical, laboratory, biopsy, and radiographic findings usually allows proper categorization to a specific syndrome, and therapy where appropriate should be initiated according to this information (see individual syndromes below). If an offending antigen that precipitates the vasculitis is recognized, the antigen should be removed where possible. If the syndrome resolves, no further action should be taken. If disease activity continues, treatment should be initiated. If the vasculitis is associated with an underlying disease such as an infection, neoplasm, or connective tissue disease, the underlying disease should be treated. If the syndrome resolves, no further action should be taken. If the syndrome does not resolve or if there is no recognizable underlying disease and the vasculitis persists, treatment should be initiated according to the category of the vasculitis syndrome. Treatment options will be considered under the individual syndromes, and general principles of therapy will be considered at the end of the chapter.

SYSTEMIC NECROTIZING VASCULITIS

POLYARTERITIS NODOSA AND MICROSCOPIC POLYANGIITIS

Definition *Classic polyarteritis nodosa* (PAN) was described in 1866 by Kussmaul and Maier. It is a multisystem, necrotizing vasculitis of small and medium-sized muscular arteries in which involvement of the renal and visceral arteries is characteristic. Classic PAN does not involve pulmonary arteries, although bronchial vessels may be involved; granulomas, significant eosinophilia, and an allergic diathesis are not part of the classic syndrome. The term *microscopic polyangiitis* (microscopic polyarteritis) was introduced into the literature by Davson in 1948. The Chapel Hill Consensus Conference on the Nomenclature of Systemic Vasculitis held in 1992 officially adopted the term to connote a necrotizing vasculitis with few or no immune complexes (pauci-immune) affecting small vessels (capillaries, venules, or arterioles). Since necrotizing arteritis involving small and medium-sized arteries may also be present, it shares features with classic PAN except that glomerulonephritis is very common in microscopic polyangiitis, and pulmonary capillaritis often occurs.

Incidence and Prevalence It is difficult to establish an accurate incidence of these diseases because of the fact that many reports ofPANactually have included both classic PAN and microscopic polyangiitis as well as other related vasculitides. Both diseases are uncommon, but classic PAN is felt to be more uncommon than microscopic polyangiitis. The mean age of onset of both PAN and microscopic polyangiitis is approximately 50 years of age, and males are slightly more frequently affected than females in both diseases.

Pathophysiology and Pathogenesis The vascular lesion in classicPAN is a necrotizing inflammation of small and medium-sized muscular arteries. The lesions are segmental and tend to involve bifurcations and branchings of arteries. They may spread circumferentially to involve adjacent veins. However, involvement of venules is not seen in classic PAN and, if present, suggests microscopic polyangiitis or the polyangiitis overlap syndrome (see below). In the acute stages of disease, polymorphonuclear neutrophils infiltrate all layers of the vessel wall and perivascular areas, which results in intimal proliferation and degeneration of the vessel wall. Mononuclear cells infiltrate the area as the lesions progress to the subacute and chronic stages. Fibrinoid necrosis of the vessels ensues with compromise of the lumen, thrombosis, infarction of the tissues supplied by the involved vessel, and, in some cases, hemorrhage. As the lesions heal, there is collagen deposition, which may lead to further occlusion of the vessel lumen. Aneurysmal dilatations up to 1 cm in size along the involved arteries are characteristic of classic PAN. Granulomas and substantial eosinophilia with eosinophilic tissue infiltrations are not characteristically found and suggest allergic angiitis and granulomatosis (see below).

Multiple organ systems are involved, and the clinicopathologic findings reflect the degree and location of vessel involvement and the resulting ischemic changes. As mentioned above, pulmonary arteries are not involved in classicPAN, and bronchial artery involvement is uncommon, whereas pulmonary capillaritis occurs frequently in microscopic polyangiitis. The pathology in the kidney in classic PAN is predominantly that of arteritis without glomerulonephritis. In contrast, glomerulonephritis is very common in microscopic polyangiitis. In patients with significant hypertension, typical pathologic features of glomerulosclerosis may be seen alone or superimposed on lesions of glomerulonephritis. In addition, pathologic sequelae of hypertension may be found elsewhere in the body.

The presence of hepatitis B antigenemia in approximately 20 to 30% of patients with systemic vasculitis, particularly of the classicPANtype, together with the isolation of circulating immune complexes composed of hepatitis B antigen and immunoglobulin, and the demonstration by immunofluorescence of hepatitis B antigen, IgM, and complement in the blood vessel walls, strongly suggest the role of immunologic phenomena in the pathogenesis of this disease. Hepatitis C infection has been reported in approximately 5% of patients with PAN; however, its pathogenic role in the vasculitis is unclear at present. Hairy cell leukemia can be associated with classic PAN; the pathogenic mechanisms of this association are unclear.

Clinical and Laboratory Manifestations Nonspecific signs and symptoms are the hallmarks of classicPAN. Fever, weight loss, and malaise are present in over one-half of cases. Patients usually present with vague symptoms such as weakness, malaise, headache, abdominal pain, and myalgias. Specific complaints related to the vascular involvement within a particular organ system may also dominate the presenting clinical picture as well as the entire course of the illness (Table 317-3). In classic PAN, renal involvement most commonly manifests as hypertension, renal insufficiency, or hemorrhage due to microaneurysms. In microscopic polyangiitis acute glomerulonephritis is the characteristic renal lesion.

There are no diagnostic serologic tests for classicPAN. In over 75% of patients, the leukocyte count is elevated with a predominance of neutrophils. Eosinophilia is seen only rarely and, when present at high levels, suggests the diagnosis of allergic angiitis and granulomatosis. The anemia of chronic disease may be seen, and an elevated erythrocyte sedimentation rate (ESR) is almost always present. Other common laboratory findings reflect the particular organ involved. Hypergammaglobulinemia may be present, and up to 30% of patients have a positive test for hepatitis B surface antigen. PositiveANCAtiters (usually of the p-ANCA type) are found in a low percentage (<20%) of patients with classic PAN. Microscopic polyangiitis is strongly associated with ANCA that are usually of the p-ANCA type, but c-ANCA have also been reported. In contrast, the ANCA in Wegener's granulomatosis (see below) are almost always of the c-ANCA type. Arteriograms may demonstrate characteristic abnormalities such as aneurysms in the small and medium-sized muscular arteries of the kidneys and abdominal viscera in classic PAN.

Diagnosis The diagnosis of classicPAN is based on the demonstration of characteristic findings of vasculitis on biopsy material of involved organs. In the absence of easily accessible tissue for biopsy, the angiographic demonstration of involved vessels, particularly in the form of aneurysms of small and medium-sized arteries in the renal, hepatic, and visceral vasculature, is sufficient to make the diagnosis. Aneurysms of vessels are not pathognomonic of classic PAN; furthermore, aneurysms need not always be present, and angiographic findings may be limited to stenotic segments and obliteration of vessels. Biopsy of symptomatic organs such as nodular skin lesions, painful testes, and muscle groups provides the highest diagnostic yields, while blind biopsy of asymptomatic organs is frequently negative. The presence of small vessel vasculitis, particularly in the setting of glomerulonephritis and pulmonary capillaritis distinguishes microscopic polyangiitis from classic PAN. In this regard, biopsy of the kidney or lung may establish the diagnosis of microscopic polyangiitis.

Prognosis The prognosis of untreated classicPAN as well as that of microscopic polyangiitis is extremely poor. The usual clinical course is characterized either by fulminant deterioration or by relentless progression associated with intermittent acute flare-ups. In classic PAN, death usually results from renal failure; from gastrointestinal complications, particularly bowel infarcts and perforation; and from cardiovascular causes. In microscopic polyangiitis, death usually results from renal failure or pulmonary hemorrhage. Intractable hypertension often compounds dysfunction in other organ systems, such as the kidneys, heart, and central nervous system, leading to additional late morbidity and mortality in classic PAN. The 5-year survival rate of untreated patients has been reported to be between 10 and 20% for both diseases; this rate has increased substantially as a result of treatment (see below).

TREATMENT

Extremely favorable therapeutic results have been reported in classicPAN with the combination of prednisone, 1 mg/kg per day, and cyclophosphamide, 2 mg/kg per day (see "Wegener's Granulomatosis" for a detailed description of this therapeutic regimen). This regimen has been reported to result in up to a 90% long-term remission rate even following the discontinuation of therapy. In less severe cases of classic PAN, glucocorticoids alone have resulted in disease remission. In addition, long-term remissions have been reported in PAN associated with hepatitis B virus antigenemia using the antiviral agent vidarabine in combination with plasma exchange with and without glucocorticoids. Favorable results have also been reported in the treatment of PAN related to hepatitis B virus withIFN-aand plasma exchange. Careful attention to the treatment of hypertension can lessen the acute and late morbidity and mortality associated with renal, cardiac, and central nervous system complications of PAN. The treatment regimen for microscopic polyangiitis is similar to that for Wegener's granulomatosis (see below), particularly if glomerulonephritis is present.

ALLERGIC ANGIITIS AND GRANULOMATOSIS (CHURG-STRAUSS DISEASE)

Definition *Allergic angiitis and granulomatosis* was described in 1951 by Churg and Strauss and is a disease characterized by granulomatous vasculitis of multiple organ systems, particularly the lung. It is characterized by vasculitis of blood vessels of various types or sizes (including veins and venules), intra- and extravascular granuloma formation together with eosinophilic tissue infiltration, and a strong association with severe asthma and peripheral eosinophilia.

Incidence and Prevalence Allergic angiitis and granulomatosis is an uncommon disease whose exact incidence, similar to classicPAN, is difficult to determine due to the grouping of multiple types of vasculitic syndromes in many reported series. The disease can occur at any age with the possible exception of infants. The mean age of onset is 44 years, with a male-to-female ratio of 1.3:1.

Pathophysiology and Pathogenesis The vasculitis of allergic angiitis and granulomatosis involves small and medium-sized muscular arteries, capillaries, veins, and venules. The characteristic histopathologic features of allergic angiitis and granulomatosis are granulomatous reactions that may be present in the tissues or even within the walls of the vessels themselves. These are usually associated with infiltration of the tissues with eosinophils. This process can occur in any organ in the body; lung involvement is predominant, with skin, cardiovascular system, kidney, peripheral nervous system, and gastrointestinal tract also commonly involved. Although the precise pathogenesis of this disease is uncertain, its strong association with asthma and its clinicopathologic manifestations including eosinophilia, granulomata, and vasculitis, which strongly suggest hypersensitivity phenomena, point to aberrant immunologic phenomena.

Clinical and Laboratory Manifestations Patients with allergic angiitis and granulomatosis exhibit nonspecific manifestations such as fever, malaise, anorexia, and weight loss, which are characteristic of a multisystem disease. The pulmonary findings in allergic angiitis and granulomatosis clearly dominate the clinical picture with severe asthmatic attacks and the presence of pulmonary infiltrates. Clinically recognizable heart disease occurs in approximately one-third of patients. Heart involvement is seen at autopsy in 62% of cases and is the cause of death in 23% of patients. Skin lesions occur in approximately 70% of patients and include purpura in addition to cutaneous and subcutaneous nodules. The renal disease in allergic angiitis and granulomatosis is less common and generally less severe than that of classicPAN and microscopic polyangiitis.

The characteristic laboratory finding in virtually all patients with allergic angiitis and granulomatosis is a striking eosinophilia, which reaches levels greater than 1000 cells/uL in more than 80% of patients. The other laboratory findings are similar to those of classicPAN and microscopic polyangiitis and reflect the organ systems involved. Allergic angiitis and granulomatosis is associated with p-ANCA.

Diagnosis The diagnosis of allergic angiitis and granulomatosis is made by biopsy in a patient with the characteristic clinical manifestations (see above). Granulomatous vasculitis with eosinophilic tissue involvement together with peripheral eosinophilia are typical.

Prognosis The prognosis of untreated allergic angiitis and granulomatosis is poor, with a reported 5-year survival of 25%. The cause of death is likely to be related to pulmonary and cardiac disease.

TREATMENT

Glucocorticoid therapy has been reported to increase the 5-year survival to more than 50%. In certain patients, the disease may be quite mild and may remit spontaneously or with short courses of glucocorticoids. In glucocorticoid failures or in patients who present with fulminant multisystem disease, the treatment of choice is a combined regimen of cyclophosphamide and alternate-day prednisone, which has resulted in a high rate of complete remission (see "Wegener's Granulomatosis" for a detailed description of this therapeutic regimen).

POLYANGIITIS OVERLAP SYNDROME

Many patients with systemic vasculitis manifest clinicopathologic characteristics that do not fit precisely into any classification but have overlapping features of classicPAN, allergic angiitis and granulomatosis, Wegener's granulomatosis, Takayasu's arteritis, and the hypersensitivity group of vasculitides. This subgroup has been referred to as the *polyangiitis overlap syndrome* and is part of the major grouping of systemic necrotizing vasculitis. This entity has been designated with a distinct classification in order to avoid confusion in attempting to fit such overlap syndromes into one or other of the more classic vasculitic syndromes. This subgroup is truly a systemic vasculitis with the same potential for resulting in irreversible organ system dysfunction as the other systemic necrotizing vasculitides. The diagnostic and therapeutic considerations as well as the prognosis for this subgroup are the same as those for classic PAN, microscopic polyangiitis, and allergic angiitis and granulomatosis.

WEGENER'S GRANULOMATOSIS

DEFINITION

Wegener's granulomatosis is a distinct clinicopathologic entity characterized by granulomatous vasculitis of the upper and lower respiratory tracts together with glomerulonephritis. In addition, variable degrees of disseminated vasculitis involving both small arteries and veins may occur.

INCIDENCE AND PREVALENCE

Wegener's granulomatosis is an uncommon disease whose true incidence is difficult to determine. It is extremely rare in blacks compared with whites; the male-to-female ratio is 1:1. The disease can be seen at any age; approximately 15% of patients are less than 19 years of age, but only rarely does the disease occur before adolescence; the mean age of onset is approximately 40 years.

PATHOPHYSIOLOGY AND PATHOGENESIS

The histopathologic hallmarks of Wegener's granulomatosis are necrotizing vasculitis of small arteries and veins together with granuloma formation, which may be either intravascular or extravascular (Fig. 317-2). Lung involvement typically appears as multiple, bilateral, nodular cavitary infiltrates (Fig. 317-3), which on biopsy almost invariably reveal the typical necrotizing granulomatous vasculitis. Endobronchial disease, either in its active form or as a result of fibrous scarring, may lead to obstruction with atelectasis. Upper airway lesions, particularly those in the sinuses and nasopharynx, typically reveal inflammation, necrosis, and granuloma formation with or without vasculitis.

In its earliest form, renal involvement is characterized by a focal and segmental glomerulitis that may evolve into a rapidly progressive crescentic glomerulonephritis. Granuloma formation is only rarely seen on renal biopsy. In addition to the classic triad of upper and lower respiratory tracts and kidney disease, virtually any organ can be involved with vasculitis, granuloma, or both.

The immunopathogenesis of this disease is unclear, although the involvement of upper airways and lungs with granulomatous vasculitis suggests an aberrant hypersensitivity response to an exogenous or even endogenous antigen that enters through or resides in the upper airway. Chronic nasal carriage of *Staphylococcus aureus* has been reported to be associated with a higher relapse rate of Wegener's granulomatosis; however, there is no evidence for a role of this organism in the pathogenesis of the disease.

Peripheral blood mononuclear cells obtained from patients with Wegener's granulomatosis manifest increased secretion ofIFN-gbut not ofIL-4, IL-5, or IL-10 compared to normal controls. The increased IFN-g production is inhibited by exogenous IL-10. In addition,TNF-aproduction from peripheral blood mononuclear cells and CD4+ T is elevated. Furthermore, monocytes from patients with Wegener's granulomatosis produce increased amounts of IL-12. These findings indicate an unbalanced Th1-type T cell cytokine pattern in this disease that may have pathogenic and perhaps ultimately therapeutic implications.

A high percentage of patients with Wegener's granulomatosis developANCA; c-ANCA are the predominant ANCA in this disease. As with the other categories of vasculitis, there is no clear evidence that ANCA play a primary role in the pathogenesis of Wegener's granulomatosis.

CLINICAL AND LABORATORY MANIFESTATIONS

A typical patient presents with severe upper respiratory tract findings such as paranasal sinus pain and drainage and purulent or bloody nasal discharge with or without nasal mucosal ulceration (Table 317-4). Nasal septal perforation may follow, leading to saddle nose deformity. Serous otitis media may occur as a result of eustachian tube blockage.

Pulmonary involvement may be manifested as asymptomatic infiltrates or may be clinically expressed as cough, hemoptysis, dyspnea, and chest discomfort. It is present in 85 to 90% of patients. Subglottic stenosis resulting from active disease or scarring occurs in approximately 16% of patients and may result in severe airway obstruction.

Eye involvement (52% of patients) may range from a mild conjunctivitis to dacryocystitis, episcleritis, scleritis, granulomatous sclerouveitis, ciliary vessel vasculitis, and retroorbital mass lesions leading to proptosis.

Skin lesions (46% of patients) appear as papules, vesicles, palpable purpura, ulcers, or subcutaneous nodules; biopsy reveals vasculitis, granuloma, or both. Cardiac involvement (8% of patients) manifests as pericarditis, coronary vasculitis, or, rarely, cardiomyopathy. Nervous system manifestations (23% of patients) include cranial neuritis, mononeuritis multiplex, or, rarely, cerebral vasculitis and/or granuloma.

Renal disease (77% of patients) generally dominates the clinical picture and, if left untreated, accounts directly or indirectly for most of the mortality in this disease. Although it may smolder in some cases as a mild glomerulitis with proteinuria, hematuria, and red blood cell casts, it is clear that once clinically detectable renal functional impairment occurs, rapidly progressive renal failure usually ensues unless appropriate treatment is instituted.

While the disease is active, most patients have nonspecific symptoms and signs such as malaise, weakness, arthralgias, anorexia, and weight loss. Fever may indicate activity of the underlying disease but more often reflects secondary infection, usually of the upper airway.

Characteristic laboratory findings include a markedly elevatedESR, mild anemia and leukocytosis, mild hypergammaglobulinemia (particularly of the IgA class), and mildly elevated rheumatoid factor. Thrombocytosis may be seen as an acute-phase reactant. In typical Wegener's granulomatosis with granulomatous vasculitis of the respiratory tract and glomerulonephritis, approximately 90% of patients have a positive c-ANCA. However, in the absence of renal disease, the sensitivity drops to approximately 70%.

DIAGNOSIS

The diagnosis of Wegener's granulomatosis is a clinicopathologic one made by the demonstration of necrotizing granulomatous vasculitis on biopsy of appropriate tissue in a patient with the clinical findings of upper and lower respiratory tract disease together with evidence of glomerulonephritis. Pulmonary tissue, preferably obtained by open thoracotomy, offers the highest diagnostic yield, almost invariably revealing granulomatous vasculitis. Biopsy of upper airway tissue usually reveals granulomatous inflammation with necrosis but may not show vasculitis. Renal biopsy confirms the presence of glomerulonephritis.

The specificity of a positive c-ANCAtiter for Wegener's granulomatosis is very high, especially if active glomerulonephritis is present. However, the presence of c-ANCA should be adjunctive and, with very rare exceptions, should not substitute for a tissue diagnosis. False-positive ANCA titers have been reported in certain infectious and neoplastic diseases.

In its typical presentation, the classic clinicopathologic complex of Wegener's granulomatosis usually provides ready differentiation from other disorders. However, if all the typical features are not present at once, it needs to be differentiated from the other vasculitides, particularly allergic angiitis and granulomatosis, Goodpasture's syndrome (Chap. 275), tumors of the upper airway or lung, and infectious diseases such as histoplasmosis (Chap. 201), mucocutaneous leishmaniasis (Chap. 215), and rhinoscleroma (Chap. 30) as well as noninfectious granulomatous diseases.

Of particular note is the differentiation from *midline granuloma* and *upper airway neoplasms*, which are part of the spectrum of *midline destructive diseases*. These diseases lead to extreme tissue destruction and mutilation localized to the midline upper airway structures including the sinuses; erosion through the skin of the face commonly occurs, a feature that is extremely rare in Wegener's granulomatosis. Although blood vessels may be involved in the intense inflammatory reaction and necrosis, primary vasculitis is seen rarely. When systemic involvement occurs, it usually declares itself as a neoplastic process. In this regard, it is likely that midline granuloma is part of the spectrum of *angiocentric immunoproliferative lesions* (AIL). The latter are considered to represent a spectrum of postthymic T cell proliferative lesions and should be treated as such (Chap. 112). The term *idiopathic* has been applied to midline granuloma when extensive diagnostic workup including multiple biopsies has failed to reveal anything other than inflammation and necrosis. Under these circumstances, it is possible that the tumor cells were masked by the intensive inflammatory response. Such cases have responded to local irradiation with 50 Gy (5000 rad). Upper airway lesions should never be irradiated in Wegener's granulomatosis.

Wegener's granulomatosis must also be differentiated from *lymphomatoid granulomatosis*, the latter also being a part of the spectrum ofAIL. Lymphomatoid granulomatosis is characterized by lung, skin, central nervous system, and kidney involvement in which atypical lymphocytoid and plasmacytoid cells infiltrate tissue in an angioinvasive manner. In this regard, it clearly differs from Wegener's granulomatosis in that it is not an inflammatory vasculitis in the classic sense but an infiltration of vessels with atypical mononuclear cells; granuloma may be present in involved tissues. Approximately 50% of patients develop a true malignant lymphoma. The presence of c-ANCAin Wegener's granulomatosis proves extremely helpful in the differentiation from all the preceding diseases.

TREATMENT

Wegener's granulomatosis was formerly universally fatal, usually within a few months after the onset of clinically apparent renal disease. Glucocorticoids alone led to some symptomatic improvement, with little effect on the ultimate course of the disease. It has been well established that the most effective therapy in this disease is cyclophosphamide given in doses of 2 mg/kg per day orally together with glucocorticoids. The leukocyte count should be monitored closely during therapy, and the dosage of cyclophosphamide should be adjusted in order to maintain the count above 3000/uL, which generally maintains the neutrophil count at approximately 1500/uL. With this approach, clinical remission can usually be induced and maintained without causing severe leukopenia with its associated risk of infection. Cyclophosphamide should be continued for 1 year following the induction of complete remission and gradually tapered and discontinued thereafter.

At the initiation of therapy, glucocorticoids should be administered together with cyclophosphamide. This can be given as prednisone, 1 mg/kg per day initially (for the first month of therapy) as a daily regimen, with gradual conversion to an alternate-day schedule followed by tapering and discontinuation after approximately 6 months.

Using the above regimen, the prognosis of this disease is excellent; marked improvement is seen in more than 90% of patients, and complete remissions are achieved in 75% of patients. A number of patients who developed irreversible renal failure but who achieved subsequent remission on appropriate therapy have undergone successful renal transplantation.

Despite the dramatic remissions induced by the therapeutic regimen described above, long-term follow-up of patients has revealed that approximately 50% of remissions are later associated with one or more relapses. Reinduction of remission is almost always achieved; however, a high percentage of patients ultimately have some degree of morbidity from irreversible features of their disease, such as varying degrees of renal insufficiency, hearing loss, tracheal stenosis, saddle nose deformity, and chronically impaired sinus function. In evaluating patients for relapse, theANCAtiter can be misleading. Many patients who achieve remission continue to have elevated titers for years. In addition, over 40% of patients who were in remission and had a fourfold increase in c-ANCA titer did not have a relapse in disease. In this regard, therapy should not be reinstituted or increased on the basis of a rise in the ANCA titer alone; however, such a finding should prompt the clinician to examine the patient carefully for any objective evidence of active disease and to monitor that patient more closely.

Certain types of morbidity are related to toxic side effects of treatment. Since the preceding therapeutic regimen calls for conversion to alternate-day glucocorticoid therapy within 3 months and ultimate discontinuation within 6 to 12 months, glucocorticoid-related side effects such as diabetes mellitus, cataracts, life-threatening infectious disease complications, serious osteoporosis, and severe cushingoid features are infrequently encountered except in those patients requiring prolonged courses of daily glucocorticoids. However, cyclophosphamide-related toxicities are more frequent and severe. Cystitis to varying degrees occurs in 50% of patients, bladder cancer in 6%, and myelodysplasia in 2%.

Some reports have indicated therapeutic success with less frequent and severe toxic side effects using intermittent boluses of intravenous cyclophosphamide (l g/m$_2$ per month) in place of daily drug administered orally. However, we and others have found an increased rate of relapse with bolus intravenous cyclophosphamide. We therefore strongly recommend that the drug be given as daily oral therapy.

Despite concerns regarding toxicity, a regimen of daily cyclophosphamide and glucocorticoids is clearly the treatment of choice in patients with immediately life-threatening disease such as rapidly progressive glomerulonephritis. However, methotrexate together with glucocorticoids may be considered as an alternative for initial therapy for certain patients whose disease is not immediately life-threatening or as a switch regimen in those patients who have experienced significant cyclophosphamide toxicity. In one study, patients in this category were given oral prednisone as described above, and methotrexate was administered orally starting at a dosage of 0.3 mg/kg, with a maximum of 15 mg/week. If the treatment was well tolerated after 1 to 2 weeks, the dosage was increased by 2.5 mg weekly up to a dosage of 20 to 25 mg/week and maintained at that level. Remissions were achieved in 33 of 42 patients (79%). Nineteen patients relapsed; 15 of these 19 relapses occurred when patients were receiving 15 mg or less of methotrexate per week; 13 of these 19 were treated with a second course of methotrexate and prednisone and 10 of 13 achieved a second remission. Toxicities of methotrexate included elevated transaminase levels (24%), leukopenia (7%), opportunistic infection (9.5%), methotrexate pneumonitis (7%), and stomatitis (2%).

Azathioprine, in doses of 1 to 2 mg/kg per day, has proven effective in some patients, particularly in maintaining remission in those in whom remission was induced by cyclophosphamide. The drug should be administered together with the glucocorticoid regimen described above. Although certain reports have indicated that trimethoprim-sulfamethoxazole may be of benefit in the treatment of Wegener's granulomatosis, there are no firm data to substantiate this, particularly in patients with serious renal and pulmonary disease. In a study examining the effect of trimethoprim-sulfamethoxazole on relapse, decreased relapses were shown only with regard to upper airway disease, and no differences in major organ relapses were observed. Trimethoprim-sulfamethoxazole alone should never be used to treat active Wegener's granulomatosis outside of the upper airway.

TEMPORAL ARTERITIS

DEFINITION

Temporal arteritis, also referred to as *cranial arteritis* or *giant cell arteritis*, is an inflammation of medium- and large-sized arteries. It characteristically involves one or more branches of the carotid artery, particularly the temporal artery; hence the name. However, it is a systemic disease that can involve arteries in multiple locations.

INCIDENCE AND PREVALENCE

The incidence of temporal arteritis varies widely in different studies and in different geographic regions. A high incidence has been found in Scandanavia and in regions of the United States with large Scandanavian populations, compared to a lower incidence in southern Europe. The annual incidence rates in individuals 50 years of age and older range from 0.49 to 23.3 per 100,000 population. It occurs almost exclusively in individuals older than 55 years; however, well-documented cases have occurred in patients 40 years old or younger. It is more common in women than in men and is rare in blacks. Familial aggregation has been reported, as has an association with HLA-DR4. In addition, genetic linkage studies have demonstrated an association of temporal arteritis with alleles at the HLA-DRB1 locus, particularly HLA-DRB1*04 variants. The disease is closely associated with *polymyalgia rheumatica*, which is more common than temporal arteritis. In Olmsted County, Minnesota, the annual incidence of polymyalgia rheumatica in individuals 50 years of age and older is 52.5 per 100,000 population.

PATHOPHYSIOLOGY AND PATHOGENESIS

Although the temporal artery is most frequently involved in this disease, patients often have a systemic vasculitis of multiple medium- and large-sized arteries, which may go undetected. Histopathologically, the disease is a panarteritis with inflammatory mononuclear cell infiltrates within the vessel wall with frequent giant cell formation. There is proliferation of the intima and fragmentation of the internal elastic lamina. Pathophysiologic findings in organs result from the ischemia related to the involved vessels. Distinct cytokine patterns as well as T lymphocytes expressing specific antigen receptors have been described suggesting the involvement of immunopathogenic mechanisms in temporal arteritis.IL-6 and IL-1b expression has been detected in a majority of circulating monocytes of patients with temporal arteritis and polymyalgia rheumatica. T cells recruited to vasculitic lesions in patients with temporal arteritis produce predominantly IL-2 andIFN-g, and the latter has been suggested to be involved in the progression to overt arteritis. Sequence analysis of the T cell receptor of tissue-infiltrating T cells in lesions of temporal arteritis indicates restricted clonal expansion, suggesting that an antigen residing in the arterial wall is recognized by a small fraction of T cells.

CLINICAL AND LABORATORY MANIFESTATIONS

The disease is characterized clinically by the classic complex of fever, anemia, highESR, and headaches in an elderly patient. Other manifestations include malaise, fatigue, anorexia, weight loss, sweats, and arthralgias. The polymyalgia rheumatica syndrome is characterized by stiffness, aching, and pain in the muscles of the neck, shoulders, lower back, hips, and thighs.

In patients with involvement of the temporal artery, headache is the predominant symptom and may be associated with a tender, thickened, or nodular artery, which may pulsate early in the disease but become occluded later (Figs. 28-CD1 and28-CD2). Scalp pain and claudication of the jaw and tongue may occur. A well-recognized and dreaded complication of temporal arteritis, particularly in untreated patients, is ocular involvement due primarily to ischemic optic neuropathy, which may lead to serious visual symptoms, even sudden blindness in some patients. However, most patients have complaints relating to the head or eyes for months before objective eye involvement. Attention to such symptoms with institution of appropriate therapy (see below) will usually avoid this complication. Claudication of the extremities, strokes, myocardial infarctions, and infarctions of visceral organs have been reported. Of note, temporal arteritis is associated with a markedly increased risk of aortic aneurysm, which is usually a late complication and may lead to dissection and death.

Characteristic laboratory findings in addition to the elevatedESRinclude a normochromic or slightly hypochromic anemia. Liver function abnormalities are common, particularly increased alkaline phosphatase levels. Increased levels of IgG and complement have been reported. Levels of enzymes indicative of muscle damage such as serum creatine kinase are not elevated.

DIAGNOSIS

The diagnosis of temporal arteritis and its associated clinicopathologic syndrome can often be made clinically by the demonstration of the classic picture of fever, anemia, and highESR with or without symptoms of polymyalgia rheumatica in an elderly patient. The diagnosis is confirmed by biopsy of the temporal artery. Since involvement of the vessel may be segmental, the diagnosis may be missed on routine biopsy; serial sectioning of biopsy specimens is recommended. When the temporal arteries appear clinically normal, but temporal arteritis is strongly suspected, a biopsy segment of a few centimeters may be required to establish the diagnosis. Ultrasonography of the temporal artery has been reported to be helpful in diagnosis. A temporal artery biopsy should be obtained as quickly as possible in the setting of ocular signs and symptoms, and under these circumstances therapy should not be delayed pending a biopsy. In this regard, it has been reported that temporal artery biopsies may show vasculitis even after more than 14 days of glucocorticoid therapy. A dramatic clinical response to a trial of glucocorticoid therapy can confirm the diagnosis.

TREATMENT

Temporal arteritis and its associated symptoms are exquisitely sensitive to glucocorticoid therapy. Treatment should begin with prednisone, 40 to 60 mg per day for approximately 1 month, followed by a gradual tapering to a maintenance dose of 7.5 to 10 mg per day. In order to lessen glucocorticoid side effects in elderly individuals, conversion to alternate-day therapy may be attempted, but only after the disease has been put into remission with daily therapy. When ocular signs and symptoms occur, it is important that therapy be initiated or adjusted to control them. Because of the possibility of relapse, therapy should be continued for at least 1 to 2 years. TheESR can serve as a useful indicator of inflammatory disease activity in monitoring and tapering therapy and can be used to judge the pace of the tapering schedule. However, minor increases in the ESR can occur as glucocorticoids are being tapered and do not necessarily reflect an exacerbation of arteritis, particularly if the patient remains symptom free. Under these circumstances, the tapering should continue with caution. If one attempts to maintain a normal ESR throughout the tapering period, glucocorticoid toxicity will almost surely occur. The prognosis is generally good, and most patients achieve complete remission that is often maintained after withdrawal of therapy.

TAKAYASU'S ARTERITIS

DEFINITION

Takayasu's arteritis is an inflammatory and stenotic disease of medium- and large-sized arteries characterized by a strong predilection for the aortic arch and its branches. For this reason, it is often referred to as the *aortic arch syndrome*.

INCIDENCE AND PREVALENCE

Takayasu's arteritis is an uncommon disease, much less common than temporal arteritis. It is most prevalent in adolescent girls and young women. Although it is more common in the Orient, it is neither racially nor geographically restricted.

PATHOPHYSIOLOGY AND PATHOGENESIS

The disease involves medium- and large-sized arteries, with a strong predilection for the aortic arch and its branches; the pulmonary artery may also be involved. The most commonly affected arteries seen by angiography are listed inTable 317-5. The involvement of the major branches of the aorta is much more marked at their origin than distally. The disease is a panarteritis with inflammatory mononuclear cell infiltrates and occasionally giant cells. There are marked intimal proliferation and fibrosis, scarring and vascularization of the media, and disruption and degeneration of the elastic lamina. Narrowing of the lumen occurs with or without thrombosis. The vasa vasorum are frequently involved. Pathologic changes in various organs reflect the compromise of blood flow through the involved vessels.

Immunopathogenic mechanisms, the precise nature of which is uncertain, are suspected in this disease. As with several of the vasculitis syndromes, circulating immune complexes have been demonstrated, but their pathogenic significance is unclear.

CLINICAL AND LABORATORY MANIFESTATIONS

Takayasu's arteritis is a systemic disease with generalized as well as local symptoms. The generalized symptoms include malaise, fever, night sweats, arthralgias, anorexia, and weight loss, which may occur months before vessel involvement is apparent. These symptoms may merge into those related to pain over the involved vessels, followed by symptoms of ischemia in organs supplied by the compromised vessels. Pulses are commonly absent in the involved vessels, particularly the subclavian artery. The frequency of arteriographic abnormalities and the potentially associated clinical manifestations are listed inTable 317-5.

The clinical course may be fulminant, may progress gradually, or may stabilize. Complications are related to the distribution of the involved vessels. Death usually occurs from congestive heart failure or cerebrovascular accidents.

Characteristic laboratory findings include an elevatedESR, mild anemia, and elevated immunoglobulin levels.

DIAGNOSIS

The diagnosis of Takayasu's arteritis should be suspected strongly in a young woman who develops a decrease or absence of peripheral pulses, discrepancies in blood pressure, and arterial bruits. The diagnosis is confirmed by the characteristic pattern on arteriography, which includes irregular vessel walls, stenosis, poststenotic dilatation, aneurysm formation, occlusion, and evidence of increased collateral circulation. Complete aortic arteriography should be obtained, unless this is renally contraindicated, in order to fully delineate the distribution and degree of arterial disease. Histopathologic demonstration of inflamed vessels adds confirmatory data; however, tissue is rarely readily available for examination.

TREATMENT

The course of the disease is variable, and spontaneous remissions may occur. Reported mortality statistics range from less than 10 to 75%. Although glucocorticoid therapy in doses of 40 to 60 mg prednisone per day alleviates symptoms, there are no convincing studies that indicate that they alone increase survival. The combination of glucocorticoid therapy for acute signs and symptoms and an aggressive surgical and/or angioplastic approach to stenosed vessels has markedly improved survival and decreased morbidity by lessening the risk of stroke, correcting hypertension due to renal artery stenosis, and improving blood flow to ischemic viscera and limbs. Unless it is urgently required, surgical correction of stenosed arteries should be undertaken only when the vascular inflammatory process is well controlled with medical therapy. Most recent mortality figures using this therapeutic approach are less than 10%. In individuals who are refractory to glucocorticoids, methotrexate in doses up to 25 mg per week has yielded encouraging results; however, long-term studies will be needed to confirm this.

HENOCH-SCHONLEIN PURPURA

DEFINITION

Henoch-Schonlein purpura, also referred to as *anaphylactoid purpura*, is a distinct systemic vasculitis syndrome that is characterized by palpable purpura (most commonly distributed over the buttocks and lower extremities), arthralgias, gastrointestinal signs and symptoms, and glomerulonephritis. It is a small vessel vasculitis.

INCIDENCE AND PREVALENCE

Henoch-Schonlein purpura is usually seen in children; most patients range in age from 4 to 7 years; however, the disease may also be seen in infants and adults. It is not a rare disease; in one series it accounted for between 5 and 24 admissions per year at a pediatric hospital. The male-to-female ratio is 1.5:1. A seasonal variation with a peak incidence in spring has been noted.

PATHOPHYSIOLOGY AND PATHOGENESIS

The presumptive pathogenic mechanism for Henoch-Schonlein purpura is immune-complex deposition. A number of inciting antigens have been suggested including upper respiratory tract infections, various drugs, foods, insect bites, and immunizations. IgA is the antibody class most often seen in the immune complexes and has been demonstrated in the renal biopsies of these patients.

CLINICAL AND LABORATORY MANIFESTATIONS

In pediatric patients, presenting symptoms related to the skin, gut, and joints are present in 50% of cases. In adults, presenting symptoms related to the skin are seen in over 70% of patients, while initial complaints related to the gut or the joints are noted in fewer than 20% of cases. The typical palpable purpura is seen in virtually all patients; most patients develop polyarthralgias in the absence of frank arthritis. Gastrointestinal involvement, which is seen in almost 70% of pediatric patients, is characterized by colicky abdominal pain usually associated with nausea, vomiting, diarrhea, or constipation and is frequently accompanied by the passage of blood and mucus per rectum; bowel intussusception may occur rarely. The renal involvement is usually characterized by mild glomerulonephritis leading to proteinuria and microscopic hematuria, with red blood cell casts in the majority of patients (Chap. 275); it usually resolves spontaneously without therapy. Rarely, a progressive glomerulonephritis will develop. Renal failure is the most common cause of death in the rare patient who dies of Henoch-Schonlein purpura. Although certain studies have found that renal disease is more severe in adults, this has not been a consistent finding. However, the course of renal disease in adults may be more insidious and thus requires close follow-up. Myocardial involvement can occur in adults but is rare in children.

Routine laboratory studies generally show a mild leukocytosis, a normal platelet count, and occasionally eosinophilia. Serum complement components are normal, and IgA levels are elevated in about one-half of patients.

TREATMENT

The prognosis of Henoch-Schonlein purpura is excellent. Most patients recover completely, and some do not require therapy. Treatment is similar for adults and children. When glucocorticoid therapy is required, prednisone in doses of 1 mg/kg per day and tapered according to clinical response has been shown to be useful in decreasing tissue edema, arthralgias, and abdominal discomfort; however, it has not proven beneficial in the treatment of skin or renal disease and does not appear to shorten the duration of active disease or lessen the chance of recurrence. Patients with rapidly progressive glomerulonephritis have been anecdotally reported to benefit from intensive plasma exchange combined with immunosuppressive drugs.

PREDOMINANTLY CUTANEOUS VASCULITIS

DEFINITION

The term *predominantly cutaneous vasculitis* has been used interchangeably with the terms *hypersensitivity vasculitis* and *cutaneous leukocytoclastic vasculitis*. Due to the heterogeneity of this group of disorders, none of these terms is totally adequate. The common denominator of this group of diseases is the involvement of small vessels of the skin. The syndrome is presumed to be associated with an aberrant hypersensitivity reaction to an antigen such as an infectious agent, a drug, or other foreign or endogenous substances. In most instances, however, an antigen is never identified and the disease remains idiopathic. The term *hypersensitivity vasculitis* is a misleading term since most of the other groups of vasculitis syndromes are probably also associated with some form of hypersensitivity or aberrant immunologic reaction to as yet unidentified antigens. The term *cutaneous leukocytoclastic vasculitis* is a better term; however, not all of these vasculitides are truly leukocytoclastic in nature. We have elected to use the term *"predominantly" cutaneous vasculitis* since skin involvement generally dominates the clinical picture, but the skin is not always the exclusive organ involved. Indeed, any organ system can be involved with this type of vasculitis; however, the extracutaneous involvement is usually much less severe than that of the systemic necrotizing vasculitides.

INCIDENCE AND PREVALENCE

Although the exact incidence of this group of vasculitis syndromes is uncertain, it is clearly more common than the systemic necrotizing vasculitis group. The disease can occur at any age and in both sexes; however, different subgroups have a higher incidence in certain age groups, and some are more common in males than females, or vice versa.

PATHOPHYSIOLOGY AND PATHOGENESIS

The typical histopathologic feature of the predominantly cutaneous vasculitides is the presence of vasculitis of small vessels. Postcapillary venules are the most commonly involved vessels; capillaries and arterioles may be involved less frequently. This vasculitis is characterized by a *leukocytoclasis*, a term that refers to the nuclear debris remaining from the neutrophils that have infiltrated in and around the vessels during the acute stages. In the subacute or chronic stages, mononuclear cells predominate; in certain subgroups, eosinophilic infiltration is seen. Erythrocytes often extravasate from the involved vessels, leading to palpable purpura.

Immune-complex deposition is generally considered to be the immunopathogenic mechanism of this type of vasculitis; however, formal proof that this is the case has not been established for all subgroups (see above). The predominantly cutaneous vasculitides can be broken down empirically into two major categories depending on the type of putative antigen involved in the hypersensitivity reaction. In the originally described group, the antigen was foreign to the host, i.e., a drug, microbe, or foreign protein. In this regard, essential mixed cryoglobulinemia has been associated with hepatitis C virus infection. In the second category, the antigen is felt to be endogenous to the host. Examples of these are the "self" proteins such as DNA or immunoglobulin, which form immune complexes with their respective antibodies and lead to vasculitic complications in systemic lupus erythematosus and rheumatoid arthritis, respectively; other examples are the recognized and putative tumor antigens that form immune complexes with antibody and lead to vasculitis associated with certain neoplasms. Certain lymphoid malignancies may also secrete cytokines that contribute to the pathogenic process.

CLINICAL AND LABORATORY MANIFESTATIONS

The hallmark of this broad group of vasculitides is the predominance of skin involvement. Skin lesions may appear typically as palpable purpura; however, other cutaneous manifestations of the vasculitis may occur, including macules, papules, vesicles, bullae, subcutaneous nodules, ulcers, and recurrent or chronic urticaria. Despite the fact that skin lesions predominate, other organ systems may be involved to varying degrees, and the extent to which this occurs may define a relatively distinct subgroup. Even in patients with isolated cutaneous involvement, the disease may be characterized by systemic signs and symptoms such as fever, malaise, myalgia, and anorexia. The skin lesions may be pruritic or even quite painful, with a burning or stinging sensation. Lesions most commonly occur in the lower extremities in ambulatory patients or in the sacral area in bedridden patients due to the effects of hydrostatic forces on the postcapillary venules. Edema may accompany certain lesions, and hyperpigmentation often occurs in areas of recurrent or chronic lesions.

There are no specific laboratory tests diagnostic of this category of vasculitis. A mild leukocytosis with or without eosinophilia is characteristic, as is an elevatedESR. Cryoglobulins and rheumatoid factor may be seen in certain cases, and serum complement levels follow no definite pattern. Laboratory abnormalities related to specific organ dysfunction reflect the involvement of these organs in the particular syndrome in question.

Drug-Induced Vasculitis Cutaneous drug reactions take a number of forms, and vasculitis is only one of these (Chaps. 57 and59). Vasculitis associated with drug reactions usually presents as palpable purpura that may be generalized or limited to the lower extremities or other dependent areas; however, urticarial lesions, ulcers, and hemorrhagic blisters may also occur (Chap. 59). Signs and symptoms may be limited to the skin, although systemic manifestations such as fever, malaise, and polyarthralgias may occur. Although the skin is the predominant organ involved, systemic vasculitis may result from drug reactions. Drugs that have been implicated in vasculitis include allopurinol, thiazides, gold, sulfonamides, phenytoin, and penicillin (Chap. 59).

Serum Sickness and Serum Sickness-Like Reactions These reactions are characterized by the occurrence of fever, urticaria, polyarthralgias, and lymphadenopathy 7 to 10 days after primary exposure and 2 to 4 days after secondary exposure to a heterologous protein (classic serum sickness) or a nonprotein drug such as penicillin or sulfa (serum sickness-like reaction). Most of the manifestations are not due to a vasculitis; however, occasional patients will have typical cutaneous venulitis that may progress rarely to a systemic vasculitis.

Vasculitis Associated with Other Underlying Primary Diseases A number of diseases have vasculitis as a secondary manifestation of the underlying primary process. Foremost among these are the connective tissue diseases, particularly *systemic lupus erythematosus* (Chap. 311), *rheumatoid arthritis* (Chap. 312), and *Sjogren's syndrome* (Chap. 314). The most common form of vasculitis in these conditions is the small vessel venulitis isolated to the skin and clinically indistinguishable from the predominantly cutaneous vasculitides noted in response to an exogenous antigen. However, certain patients may develop a fulminant systemic necrotizing vasculitis indistinguishable from thePANgroup.

Cryoglobulinemia may be seen in a number of the diverse vasculitic syndromes. *Essential mixed cryoglobulinemia* (Chap. 275) may present as a predominantly cutaneous vasculitis. However, typically, it is associated with glomerulonephritis, arthralgias, hepatosplenomegaly, and lymphadenopathy in addition to skin involvement.

Vasculitis can be associated with certain *malignancies*, particularly lymphoid or reticuloendothelial neoplasms. Leukocytoclastic venulitis confined to the skin is the most common finding; however, widespread systemic vasculitis may occur. Of particular note is the association of *hairy cell leukemia* (Chap. 112) with classicPAN.

A leukocytoclastic vasculitis predominantly involving the skin with occasional involvement of other organ systems may be a minor component of many other diseases. These include *subacute bacterial endocarditis*, *Epstein-Barr virus infection*, *HIV infection*, *chronic active hepatitis*, as well as a number of other infections; *ulcerative colitis*, *congenital deficiencies of various complement components*, *retroperitoneal fibrosis*, and *primary biliary cirrhosis*. Association of predominantly cutaneous vasculitis with*a_1-antitrypsin deficiency*, *intestinal bypass surgery*, and *relapsing polychondritis* has been reported.

DIAGNOSIS

The diagnosis of this category of vasculitis is made by the demonstration of vasculitis on biopsy. Given the predominance of cutaneous involvement, biopsy material is generally readily available. An important principle in the diagnostic approach to patients with presumed isolated cutaneous vasculitis is to search for an etiology of the vasculitis -- be it an exogenous agent such as a drug or an infection or an endogenous condition such as an underlying disease (Fig. 317-1). In addition, a careful physical and laboratory examination should be performed to rule out the possibility of systemic disease. This should start with the least invasive diagnostic approach and proceed to the more invasive only if clinically indicated.

TREATMENT

Most cases of predominantly cutaneous vasculitis resolve spontaneously, and others remit and relapse before finally remitting completely. In those patients in whom persistent cutaneous disease evolves or in whom extracutaneous organ system involvement occurs, a variety of therapeutic regimens have been tried with variable results. In general, the treatment of this type of vasculitis has not been satisfactory. Fortunately, since the disease is generally limited to the skin, this lack of consistent response to therapy usually does not lead to a life-threatening situation. When an antigenic stimulus is recognized as the precipitating factor in the vasculitis, it should be removed; if this is a microbe, appropriate antimicrobial therapy should be instituted. If the vasculitis is associated with another underlying disease, treatment of the latter often results in resolution of the former. In situations where disease is apparently self-limited, no therapy, except possibly symptomatic therapy, is indicated. When disease persists and when there is no evidence of an inciting agent, an associated disease, or an underlying systemic vasculitis, the decision to treat should be based on weighing the balance between the degree of symptoms and the risk of treatment. If the decision is made to treat, glucocorticoid therapy should be instituted, usually as prednisone, 1 mg/kg per day, in a regimen aimed at rapid tapering where possible, either directly to discontinuation or by conversion to an alternate-day regimen followed by ultimate discontinuation. In cases that prove refractory to glucocorticoids, a trial of an immunosuppressive agent may be indicated. Patients with chronic vasculitis isolated to cutaneous venules rarely respond dramatically to any therapeutic regimen, and immunosuppressive agents should be used only as a last resort in these patients. Methotrexate and azathioprine have been used in such situations in anecdotal reports (see above for specific regimens). Although cyclophosphamide is the most effective therapy for the systemic vasculitides, it should almost never be used for predominantly cutaneous vasculitis because of the potential toxicity. Plasmapheresis has been used with some success in fulminant cases. Dapsone has been tried in a number of patients with isolated cutaneous vasculitis with rare anecdotal reports of success. However, this drug has been consistently beneficial as therapy for cutaneous vasculitis only in patients with erythema elevatum diutinum (see below).

KAWASAKI DISEASE

Kawasaki disease (mucocutaneous lymph node syndrome) is an acute, febrile, multisystem disease of children. It is characterized by unresponsiveness to antibiotics, nonsuppurative cervical adenitis, and changes in the skin and mucous membranes such as edema; congested conjunctivae; erythema of the oral cavity, lips, and palms; and desquamation of the skin of the fingertips. Although the disease is generally benign and self-limited, it is associated with coronary artery aneurysms in approximately 25% of cases, with an overall case fatality rate of 0.5 to 2.8%. These complications usually occur between the third and fourth weeks of illness during the convalescent stage. Vasculitis of the coronary arteries is seen in almost all the fatal cases that have been autopsied. There is typical intimal proliferation and infiltration of the vessel wall with mononuclear cells. Beadlike aneurysms and thromboses may be seen along the artery. Most investigators agree that many of the cases ofPANformerly reported in children were actually arteritic complications of unrecognized Kawasaki disease. Other manifestations include pericarditis, myocarditis, myocardial ischemia and infarction, and cardiomegaly.

It is likely that immune-mediated injury to blood vessel endothelium is involved in the pathogenesis of this disease. Patients with Kawasaki disease have been demonstrated to have evidence of increased immune activation characterized by increased activated helper T cells and monocytes, elevated serum-solubleIL-2 receptor levels, elevated levels of spontaneous IL-1 production by peripheral blood mononuclear cells, anti-endothelial cell antibodies, and increased cytokine-inducible activation antigens on their vascular endothelium. A strong association has been reported between a novel form of *S. aureus* that releases toxic shock syndrome toxin 1 and Kawasaki disease, suggesting that this was the causative organism and was acting as a superantigen similar to the superantigen effect in toxic shock syndrome. However, analysis of the T cell receptor repertoire of patients with Kawasaki disease has yielded conflicting data as to whether the T cell response is driven by a superantigen or by a conventional antigen.

Apart from the up to 2.8% of patients who develop fatal complications, the prognosis of this disease for uneventful recovery is excellent. High-dose intravenousg globulin (2 g/kg as a single infusion over 10 h) together with aspirin (100 mg/kg per day for 14 days followed by 3 to 5 mg/kg per day for several weeks) have been shown to be effective in reducing the prevalence of coronary artery abnormalities when administered early in the course of the disease.

ISOLATED VASCULITIS OF THE CENTRAL NERVOUS SYSTEM

Isolated vasculitis of the central nervous system is an uncommon clinicopathologic entity characterized by vasculitis restricted to the vessels of the central nervous system without other apparent systemic vasculitis. Although the arteriole is most commonly affected, vessels of any size can be involved. The inflammatory process is usually composed of mononuclear cell infiltrates with or without granuloma formation. Cases have been associated with cytomegalovirus, syphilis, pyogenic bacterial, and varicella-zoster infections, as well as with Hodgkin's disease and amphetamine abuse; however, in most cases no underlying disease process has been identified.

Patients may present with severe headaches, altered mental function, and focal neurologic defects. Systemic symptoms are generally absent. Devastating neurologic abnormalities may occur depending on the extent of vessel involvement. The diagnosis is generally made by demonstration of characteristic vessel abnormalities on arteriography and confirmed by biopsy of the brain parenchyma and leptomeninges. In the absence of a brain biopsy, care should be taken not to misinterpret as true primary vasculitis angiographic abnormalities that might actually be vessel spasm related to another cause. The prognosis of this disease is poor; however, in certain patients the disease may remit spontaneously, and some reports indicate that glucocorticoid therapy alone or together with cyclophosphamide in steroid-resistant patients administered as described above for the systemic vasculitides has induced sustained clinical remissions in a small number of patients.

THROMBOANGIITIS OBLITERANS (BUERGER'S DISEASE)

Thromboangiitis obliterans is an inflammatory occlusive peripheral vascular disease of unknown etiology that affects arteries and veins. Thrombosis of the vessels is likely the primary event, and so this disease is not a classic vasculitis. However, it is considered among the vasculitides because of the intense inflammatory response within the thrombus and the fact that there is often a vasculitis of the vasa vasorum in the arterial wall.**The disease is discussed in detail in Chap. 248.*

BEHCET'S SYNDROME

Behcet's syndrome is a clinicopathologic entity characterized by recurrent episodes of oral and genital ulcers, iritis, and cutaneous lesions. The underlying pathologic process is a leukocytoclastic venulitis, although vessels of any size and in any organ can be involved.**This disorder is described in detail in Chap. 316.*

MISCELLANEOUS VASCULITIDES

A variety of disorders, many of which are uncommon, are characterized by varying degrees of inflammatory responses involving blood vessels. *Cogan's syndrome* is a disease characterized by nonsyphilitic interstitial keratitis together with vestibuloauditory symptoms. It may be associated with a systemic vasculitis involving vessels of different sizes as well as the aortic valve.

Erythema elevatum diutinum is a rare chronic skin disorder of unknown etiology characterized by persistent red, purple, and yellowish papules, plaques, and nodules usually distributed symmetrically over the extensor surface of the limbs; on biopsy, they demonstrate a leukocytoclastic venulitis together with a marked dermal inflammatory infiltrate. An association with streptococcal infections has been reported. The disease responds dramatically to dapsone therapy.

Certain *infections* may directly trigger an inflammatory vasculitic process. For example, rickettsias can invade and proliferate in the endothelial cells of small blood vessels causing a vasculitis (Chap. 177). In addition, the inflammatory response around blood vessels associated with certain systemic fungal diseases such as histoplasmosis (Chap. 201) may mimic a primary vasculitic process.

PRINCIPLES OF TREATMENT

Once a diagnosis of vasculitis has been established, a decision regarding therapeutic strategy must be made (Fig. 317-1). The vasculitis syndromes represent a wide spectrum of diseases with varying degrees of severity. Some require immediate and aggressive therapy with glucocorticoids and immunosuppressive agents, while others should be treated conservatively and symptomatically, usually with nonsteroidal anti-inflammatory drugs. Since the potential toxic side effects of certain therapeutic regimens may be substantial, the risk-versus-benefit ratio of any therapeutic approach should be weighed carefully. Specific therapeutic regimens are discussed above for the individual vasculitis syndromes; however, certain general principles regarding therapy should be considered. On the one hand, glucocorticoids and/or immunosuppressive therapy should be instituted immediately in diseases where irreversible organ system dysfunction and high morbidity and mortality have been clearly established. Wegener's granulomatosis is the prototype of a severe systemic vasculitis requiring such a therapeutic approach (see above). On the other hand, when feasible, aggressive therapy should be avoided for vasculitic manifestations that rarely result in irreversible organ system dysfunction and that usually do not respond to such therapy. For example, isolated cutaneous vasculitis usually resolves with symptomatic treatment, and prolonged courses of glucocorticoids uncommonly result in clinical benefit. Immunosuppressive agents have not proven to be beneficial in isolated cutaneous vasculitis, and their toxic side effects generally outweigh any potential beneficial effects. Glucocorticoids should be initiated in those systemic vasculitides that cannot be specifically categorized or for which there is no established standard therapy; immunosuppressive therapy should be added in these diseases only if an adequate response does not result or if remission can only be achieved and maintained with an unacceptably toxic regimen of glucocorticoids. When remission is achieved, one should continually attempt to taper glucocorticoids to an alternate-day regimen and discontinue when possible. When using immunosuppressive regimens, one should taper and discontinue the drug as soon as is feasible upon induction of remission (see below).

When glucocorticoids are used, prednisone is generally the formulation of choice and is administered as 1 mg/kg per day orally, first in divided doses and then converted to a single daily dose. After clinical improvement is noted (usually within a month), the regimen is gradually converted to an alternate-day schedule, followed by tapering and discontinuation after approximately 6 months or as the clinical response dictates. When an immunosuppressive agent is required, cyclophosphamide is the drug of choice and its efficacy has been clearly established in Wegener's granulomatosis and the severe systemic vasculitides (see above). It should be given in doses of 2 mg/kg per day orally. It is recommended that the drug be taken as a single dose in the morning together with large amounts of fluid. Dose adjustments should be based on the leukocyte count, which should be maintained above 3000/uL. Leukocyte counts at any given time will reflect the dosage of cyclophosphamide taken the previous week. Of note, neutropenia may become more pronounced as glucocorticoids are tapered. The regimen that has proven successful in Wegener's granulomatosis (see above) and that should be followed for the other severe systemic vasculitides has called for continuation of cyclophosphamide for approximately 1 year following the induction of complete remission, with gradual tapering (by 25-mg decrements of the daily dose) over several months until discontinuation. No other drug has proven to be as effective as cyclophosphamide for severe life-threatening vasculitis. However, immediate and long-range toxic side effects may be severe. Alternative immunosuppressive regimens may be instituted where indicated in those patients who cannot tolerate cyclophosphamide due to unacceptable side effects or who do not wish to take cyclophosphamide because of the potential side effects, particularly infertility or sterility in individuals of child-bearing age. Methotrexate has been shown to be an acceptable alternative to cyclophosphamide when the latter drug cannot be used. Methotrexate is administered as a single weekly dose initially at a dosage of 0.3 mg/kg, not to exceed 15 mg/week. If the treatment is well tolerated after 1 to 2 weeks, the dosage should be increased by 2.5 mg weekly up to a dosage of 20 to 25 mg/week and maintained at that level. Azathioprine at a dosage of 2 mg/kg per day orally has also been employed as an alternative to cyclophosphamide in severe systemic vasculitis with less favorable results. In unusual cases in which none of the above regimens have resulted in remission of the vasculitis, certain experimental approaches have been used, such as plasmapheresis together with immunosuppressive drugs, with anecdotal reports of limited success. In addition, other immunosuppressive agents such as cyclosporine have been employed with minimal success.

Physicians should be thoroughly aware of the toxic side effects of therapeutic agents employed (Table 317-6). Side effects of glucocorticoid therapy are markedly decreased in frequency and duration in patients on alternate-day regimens compared to daily regimens. When cyclophosphamide is administered chronically in doses of 2 mg/kg per day for substantial periods of time (one to several years), the incidence of cystitis as defined by nonglomerular hematuria is approximately 50% and the incidence of bladder cancer is 6%. Bladder cancer can occur several years after discontinuation of cyclophosphamide therapy; therefore, monitoring for bladder cancer should continue indefinitely in patients who have received prolonged courses of daily cyclophosphamide. Significant alopecia is unusual in the chronically administered, low-dose regimen. When patients are receiving low-dose cyclophosphamide, the white blood count (WBC) is maintained above 3000/uL, and the patient is not receiving daily glucocorticoids, the incidence of life-threatening opportunistic infections is very low. However, the WBC is not an accurate predictor of risk of opportunistic infections in patients receiving methotrexate; infections with *Pneumocystis carinii* and certain fungi can be seen in the face of WBC that are within normal limits. All vasculitis patients who are not allergic to sulfa and who are receiving daily glucocorticoids in combination with an immunosuppressive drug should receive trimethoprim-sulfamethoxazole as prophylaxis against *P. carinii* infection.

Finally, it should be emphasized that each patient is unique and requires individual decision-making. The above outline should serve as a framework to guide therapeutic approaches; however, flexibility should be practiced in order to provide maximal therapeutic efficacy with minimal toxic side effects in each patient.

(Bibliography omitted in Palm version)

Back to Table of Contents

318. SARCOIDOSIS - *Ronald G. Crystal*

DEFINITION

Sarcoidosis is a chronic, multisystem disorder of unknown cause characterized in affected organs by an accumulation of T lymphocytes and mononuclear phagocytes, noncaseating epithelioid granulomas, and derangements of the normal tissue architecture. Although there are usually skin anergy and depressed cellular immune processes in the blood, sarcoidosis is characterized at the sites of disease by exaggerated T helper 1 (T_H1) lymphocyte immune processes. All parts of the body can be affected, but the organ most frequently affected is the lung. Involvement of the skin, eye, liver, and lymph nodes is also common. The disease is often acute or subacute and self-limiting, but in many individuals it is chronic, waxing and waning over many years.

ETIOLOGY

The cause of sarcoidosis is unknown. Various infectious and noninfectious agents have been implicated, but there is no proof that any specific agent is responsible. However, all available evidence is consistent with the concept that the disease results from an exaggerated cellular immune response (acquired, inherited, or both) to a limited class of persistent antigens or self-antigens.

INCIDENCE AND PREVALENCE

Sarcoidosis is a relatively common disease affecting individuals of both sexes and almost all ages, races, and geographic locations. Females appear to be slightly more susceptible than males. Cases of sarcoidosis have been described in all of the major races, and the disease is found throughout the world. It has been suggested that sarcoidosis is more common in certain geographic areas such as the southeastern part of the United States, but when case-matched controls have been used, these geographic differences are less convincing. There is a remarkable diversity of the prevalence of sarcoidosis among certain ethnic and racial groups, with a range of <1 to 64 per 100,000 worldwide. The prevalence of sarcoidosis is from 10 to 40 per 100,000 in the United States and Europe. In the United States, most patients are black, with a ratio of blacks to whites ranging from 10:1 to 17:1. In Europe, however, the disease affects mostly whites. Furthermore, while the prevalence per 100,000 in Sweden is 64, in France it is 10, in Poland 3, yet for Irish females living in London it is 200. In contrast, the disease is very rare among Inuit, Canadian Indians, New Zealand Maoris, and Southeast Asians.

Most patients present with sarcoidosis between the ages of 20 and 40, but the disease can occur in children and in the elderly. Several hundred kindred groups with familial sarcoidosis have been described, and the disease has been observed in twins, more commonly in monozygotic than in dizygotic pairs. There also have been several instances of husband-wife pairs identified, and geographic foci of sarcoidosis among unrelated individuals living closely within a community, arguing for some environmental factors in the pathogenesis of the disease. Although the disease is believed to result from exaggerated cellular immune responses to a limited class of antigens, no clear patterns in any HLA locus have emerged. Unlike many diseases in which the lung is involved, sarcoidosis favors nonsmokers.

PATHOPHYSIOLOGY AND IMMUNOPATHOGENESIS

The first manifestation of the disease is an accumulation of mononuclear inflammatory cells, mostly CD4+ T_H1 lymphocytes and mononuclear phagocytes, in affected organs. This inflammatory process is followed by the formation of granulomas, aggregates of macrophages and their progeny, epithelioid cells, and multinucleated giant cells. The typical sarcoid granuloma is a compact structure composed of an aggregate of mononuclear phagocytes surrounded by a rim of CD4+ T lymphocytes and, to a far lesser extent, B lymphocytes. The overall structure is relatively discrete and is interspersed with fine collagen fibrils, presumably remnants of the underlying connective tissue matrix. The giant cells within the granuloma can be of the Langhans' or foreign-body variety and often contain inclusions such as Schaumann bodies (conchlike structures), asteroid bodies (stellate-like structures), and residual bodies (refractile calcium-containing inclusions).

Together the accumulated T cells, mononuclear phagocytes, and granulomas represent the active disease. Other than the fact that they take up space and thus their bulk modifies the local architecture, for all except late stage cases, there is no evidence that the mononuclear inflammatory cells dispersed in the tissue or in the granuloma injure the affected organ by releasing mediators that damage the normal parenchymal cells or the extracellular matrix. Rather, organ dysfunction in sarcoidosis results mostly from the accumulated inflammatory cells distorting the architecture of the affected tissue; if a sufficient number of structures vital to the function of the tissue are involved, the disease becomes clinically apparent in that organ. Thus, while autopsy series show that, to some extent, sarcoidosis involves most organs in the majority of patients, the disease manifests clinically only in organs where it affects function (such as the lung and eye) or in organs where it is readily observed (such as the skin or, by x-ray, the hilar nodes). For example, in the lung, the inflammatory cells and granulomas distort the walls of the alveoli, bronchi, and blood vessels (Fig. 318-1*A*), thus altering the intimate relationships between air and blood necessary for normal gas exchange. When a sufficient amount of pulmonary tissue is involved, it is sensed by the individual as dyspnea. In contrast, most individuals with sarcoidosis have granulomatous mononuclear cell inflammation in the liver but usually do not have symptoms or significant functional derangements referable to that organ, likely because the disease process does not modify the local structures sufficiently to affect function.

If the disease is suppressed, either spontaneously or with therapy, the mononuclear inflammation is reduced in intensity and the number of granulomas is reduced. The granulomas resolve either by dispersion of the cells or by centripetal proliferation of fibroblasts from the periphery of the granuloma inward, to form a small scar. In chronic cases, the mononuclear cell inflammation persists for years. If the intensity of the inflammation is sufficiently high for a sufficiently long period, the derangements to the affected tissues result in extensive damage, the development of fibrosis, and permanent loss of organ function.

All available evidence suggests that active sarcoidosis results from an exaggerated cellular immune response to a variety of antigens or self-antigens, in which the process of T lymphocyte triggering, proliferation, and activation is skewed in the direction of CD4+ TH1 lymphocyte processes (Fig. 318-1*B*). The result is an exaggerated TH1 T lymphocyte response and thus the accumulation of large numbers of activated TH1 cells in the affected organs. Since the activated TH1 lymphocyte releases mediators that attract and activate mononuclear phagocytes, it is likely that the process of granuloma formation is a secondary phenomenon that is a consequence of the exaggerated TH1 cell process. In this context, the current hypotheses of the cause of sarcoidosis, not mutually exclusive, include the following: (1) The disease is caused by a class of persistent antigens, nonself or self, that trigger only the TH1 cell arm of the immune response; (2) the disease results from an inadequate suppressor arm of the immune response, such that TH1 cell processes cannot be shut down in a normal fashion; or (3) the disease results from inherited (and/or acquired) differences in immune response genes, such that the response to a variety of antigens is an exaggerated, TH1 cell process.

Independent of the inciting agent(s) or the reason why there is an exaggerated TH1 cell response, there is a general understanding of the processes responsible for the maintenance of the inflammation and the development of the granuloma. The TH1 lymphocytes accumulate at the sites of disease, at least in part, because they proliferate in these sites at an exaggerated rate. This T cell proliferation is maintained by the spontaneous release of interleukin (IL) 2, the T cell growth factor, by activated TH1 cells in the local milieu. In this regard, sarcoidosis is a remarkable example of compartmentalization of the immune system and a dramatic illustration of why disease activity of sarcoidosis cannot be assessed by evaluating the immune system only in the blood. Whereas the TH1 cells in the involved organs are releasing IL-2 and proliferating at an enhanced rate, the T cells in other sites, such as blood, are quiescent. Furthermore, while there is a marked enhancement of the number of TH1 cells at the sites of disease, the numbers of TH1 cells in the blood are normal or slightly reduced. In the involved organs, the ratio of CD4+ to CD8+ T cells may be as high as 10:1 compared to the ratio of 2:1 found in normal tissues or in the blood of affected individuals.

In addition to driving other TH1 cells in the affected organs to proliferate, the TH1 cells at the sites of disease are activated and release mediators that both recruit and activate mononuclear phagocytes. The activated TH1 cells accomplish this by releasing a variety of mediators (lymphokines) including proteins capable of recruiting blood monocytes to the local milieu of the activated T cells and interferon g, a protein that, among its many actions, activates mononuclear phagocytes. Together with cytokines such as interleukin (IL)-12 and others released locally, these mediators recruit blood monocytes to the affected organs and activate them, providing the building blocks for the formation of the granuloma.

In addition to these exaggerated cellular immune processes, active sarcoidosis is also characterized by hyperglobulinemia. Included among the immunoglobulins are antibodies against a variety of infectious agents as well as IgM anti-T cell antibodies. However, there is no evidence that any of these antibodies plays a role in the pathogenesis of the disease, and they are thought to result from the nonspecific polyclonal stimulation of B cells by the activated T cells at the site of disease.

If the damage in the affected organs is sufficiently extensive that the remaining parenchymal cells cannot reestablish the normal tissue architecture, the usual result is fibrosis, the proliferation of mesenchymal cells, and deposition of their connective tissue products. There is convincing evidence that the fibroblast proliferation is directed by tissue macrophages spontaneously releasing growth signals for fibroblasts, including platelet-derived growth factor, fibronectin, and insulin-like growth factor 1. It is not known, however, why this fibrotic process occurs only in a relatively small proportion of individuals with sarcoidosis.

CLINICAL MANIFESTATIONS

Sarcoidosis is a systemic disease, and thus the clinical manifestations may be generalized or focused on one or more organs. However, because the lung is almost always involved, most patients have symptoms referable to the respiratory system. Independent of the site, the clinical manifestations of the disease relate directly to the exaggerated T_H1 lymphocyte-mononuclear phagocyte granulomatous inflammatory process itself or to the sequelae resulting from the permanent damage caused by this process.

Sarcoidosis is occasionally discovered in a completely asymptomatic individual, but more commonly it presents abruptly over 1 to 2 weeks or the affected individual develops symptoms insidiously over several months. Independent of the mode of presentation, ~75% of all cases present in individuals younger than 40 years.

The asymptomatic form is usually detected by a routine examination, such as a chest film. In the United States, this form represents about 10 to 20% of all cases, but in countries where chest films are mandatory in preemployment screening programs, the proportion of asymptomatic patients is higher.

So-called acute or subacute sarcoidosis develops abruptly over a period of a few weeks and represents 20 to 40% of all cases. These individuals usually have constitutional symptoms such as fever, fatigue, malaise, anorexia, or weight loss. These symptoms are usually mild, but in approximately 25% of the acute cases the constitutional complaints are extensive. Many patients have respiratory symptoms, including cough, dyspnea, a vague retrosternal chest discomfort and/or polyarthritis. Two syndromes have been identified in the acute group. Lofgren's syndrome, frequent in Scandinavian, Irish, and Puerto Rican females, includes the complex of erythema nodosum (Plate IIE-70) and x-ray findings of bilateral hilar adenopathy, often accompanied by joint symptoms, including arthritis at the ankles, knees, wrists, or elbows. The Heerfordt-Waldenstrom syndrome describes individuals with fever, parotid enlargement, anterior uveitis, and facial nerve palsy.

The insidious form of sarcoidosis develops over months and is associated usually with respiratory complaints without constitutional symptoms. In the United States, 40 to 70% of all patients with sarcoidosis patients are in this category. About 10% of these individuals have symptoms referable to organs other than the lung. It is the individuals who present with the insidious form of sarcoidosis who most commonly go on to develop chronic sarcoidosis, with permanent damage to the lung and other organs.

Despite the fact that sarcoidosis is a systemic disease and some evidence of inflammation can be detected in most organs in the majority of patients, sarcoidosis is important clinically because of the pulmonary abnormalities and, to a lesser extent, lymph node, skin, liver, and eye involvement. Far less commonly, other organs are involved significantly.

Lung Of individuals with sarcoidosis, 90% have abnormal findings on chest x-ray at some time during their course (Fig. 318-2*A*). Overall, ~50% develop permanent pulmonary abnormalities, and 5 to 15% have progressive fibrosis of the lung parenchyma. Sarcoidosis of the lung is primarily an interstitial lung disease (Chap. 259) in which the inflammatory process involves the alveoli, small bronchi, and small blood vessels. These individuals typically have symptoms of dyspnea, particularly with exercise, and a dry cough. In acute and subacute cases, physical examination usually reveals dry rales. Hemoptysis is rare, as is production of sputum. Occasionally, the large airways are involved to a degree sufficient to cause dysfunction. Distal atelectasis can result from endobronchial sarcoidosis or from external compression from enlarged intrathoracic nodes. Rarely, wheezing is heard, incorrectly suggesting asthma. Large-vessel pulmonary granulomatous arteritis is common, but it rarely causes major problems. If it dominates the pulmonary lesions, it is sometimes called *necrotizing sarcoidal granulomatosis*. The pleura is involved in 1 to 5% of cases, almost always manifesting as a unilateral pleural effusion with characteristics of an exudate containing lymphocytes. The effusions usually clear within a few weeks, but chronic pleural thickening can result. Pneumothorax is very rare.

Lymph Nodes Lymphadenopathy is very common in sarcoidosis. Intrathoracic nodes are enlarged in 75 to 90% of all patients; usually this involves the hilar nodes, but the paratracheal nodes are commonly involved (Fig. 318-2*A*). Less frequently, there is enlargement of subcarinal, anterior mediastinal, or posterior mediastinal nodes. Peripheral lymphadenopathy is very common, particularly involving the cervical, axillary, epitrochlear, and inguinal nodes. The nodes in the retroperitoneal area and in the mesenteric chain also can enlarge. All these nodes are nonadherent, with a firm, rubbery texture. Palpation causes no pain. Unlike nodes in tuberculosis, the nodes do not ulcerate. The lymphadenopathy rarely causes a problem for the affected individual; however, if it is massive, it can be disfiguring and can impinge on other organs and lead to functional impairment.

Skin Sarcoidosis involves the skin in ~25% of patients. The most common lesions are erythema nodosum (Plate IIE-70), plaques, maculopapular eruptions, subcutaneous nodules, and lupus pernio. Erythema nodosum, comprising bilateral, tender red nodules on the anterior surface of the legs, is not specific for sarcoidosis but is common, particularly in acute sarcoidosis, in combination with systemic symptoms and polyarthralgias. Treatment is not required, since the lesions resolve spontaneously in 2 to 4 weeks. Erythema nodosum is much more common among patients with sarcoidosis in Europe than in the United States. Skin plaques associated with sarcoidosis are purple, indolent lesions, often raised, and usually occur on the face, buttocks, and extremities. The maculopapular eruptions occur on the face around the eyes and nose, on the back, and on the extremities. These are elevated lesions <1 cm in diameter with a flat, waxy top. Subcutaneous nodules are most common on the trunk and extremities.

Lupus pernio is characterized by indurated blue-purple, swollen, shiny lesions on the nose, cheeks, lips, ears, fingers, and knees. The lesions on the tip of the nose cause a bulbous appearance, sometimes associated with varicosities. The nasal mucosa is usually involved, and underlying bone can be destroyed. Sarcoidosis also can involve old surgical scars and tattoos. Although it may be disfiguring, cutaneous sarcoidosis rarely causes major problems. Clubbing of the fingers is occasionally observed in sarcoidosis, usually in association with extensive pulmonary fibrosis.

Eye Eye involvement occurs in ~25% of patients with sarcoidosis, and it can cause blindness. The usual lesions involve the uveal tract, iris, ciliary body, and choroid. Of those patients with eye involvement, ~75% have anterior uveitis and 25 to 35% have posterior uveitis. There is blurred vision, tearing, and photophobia. The uveitis can develop rapidly and may clear spontaneously over a 6- to 12-month period. It also can develop insidiously and be chronic. Conjunctival involvement is also common, usually with small, yellow nodules. When the lacrimal gland is involved, a keratoconjunctivitis sicca syndrome, with dry, sore eyes, can result.

Upper Respiratory Tract The nasal mucosa is involved in up to 20% of patients, usually presenting with nasal stuffiness. Any of the structures of the mouth can be involved, particularly the tonsils. Sarcoidosis involves the larynx in ~5% of patients. The epiglottis and areas around the true vocal cords are usually involved, but the cords themselves are not. These individuals are usually hoarse, and they have dyspnea, wheezing, and stridor; complete obstruction can occur.

Bone Marrow and Spleen Sarcoidosis of the marrow is reported in 15 to 40% of patients, but it rarely causes hematologic abnormalities other than a mild anemia, neutropenia, eosinophilia, and occasionally, thrombocytopenia. Although splenomegaly occurs in only 5 to 10% of patients, celiac angiography or splenic biopsy reveals involvement in 50 to 60% of patients. The presentation and complications of splenomegaly in sarcoidosis are similar to those of splenomegaly in general.

Liver Although liver biopsy reveals liver involvement in 60 to 90% of patients, liver dysfunction is usually not important clinically. Sarcoidosis involves generally the periportal areas. Isolated granulomatous hepatitis can occur. Approximately 20 to 30% have hepatomegaly and/or biochemical evidence of liver involvement. Usually these changes reflect a cholestatic pattern and include an elevated alkaline phosphatase level; the bilirubin and aminotransferase levels are only mildly elevated, and jaundice is rare. Rarely, portal hypertension can develop, as can intrahepatic cholestasis with cirrhosis.

Kidney Clinically apparent primary renal involvement in sarcoidosis is rare, although tubular, glomerular, and renal artery diseases have been reported. More commonly, but still in only 1 to 2% of all patients, there is a disorder of calcium metabolism with hypercalciuria, with or without hypercalcemia. If chronic, nephrocalcinosis and nephrolithiasis can result. It is believed that the calcium abnormalities are associated with enhanced calcium absorption in the gut, which is related to an abnormally high level of circulating 1,25-dihydroxyvitamin D produced by mononuclear phagocytes in the granulomas.

Nervous System All components of the nervous system can be involved in sarcoidosis. Neurologic findings are observed in about 5% of patients. Seventh nerve involvement with unilateral facial paralysis is most common. It occurs suddenly and is usually transient. Other common manifestations of neurosarcoid include optic nerve dysfunction, papilledema, palate dysfunction, hearing abnormalities, hypothalamic and pituitary abnormalities, chronic meningitis, and, occasionally, space-occupying lesions. Psychiatric disturbances have been described, and seizures can occur. Rarely, multiple lesions occur that mimic multiple sclerosis, spinal cord abnormalities, and peripheral neuropathy.

Musculoskeletal System The bones, joints, and/or muscles can be involved in sarcoidosis. Bone lesions are observed in 5% of patients and include variable-sized cysts in areas of expanded bone; well-defined, round, punched-out lesions; or lattice-like changes. Hand and foot bones are the common sites, but most bones can be involved. Occasionally, the bone lesions are tender and painful. Joint involvement is more common, with an incidence of 25 to 50% in known cases of sarcoidosis. Arthralgias and frank arthritis occur mostly in large joints; they can be migratory and are usually transient, but they can be chronic and result in deformities. Although muscle biopsy frequently demonstrates granulomatous inflammation, muscle dysfunction is rare. However, nodules, polymyositis, and chronic myopathy have been described.

Heart Approximately 5% of patients have significant heart involvement, with clinical evidence of cardiac dysfunction. Left ventricular wall involvement is common. Arrhythmias are frequent, and serious conduction disturbances, including complete heart block, can occur. Papillary muscle dysfunction, pericarditis, and congestive heart failure are also observed. Cor pulmonale secondary to chronic pulmonary fibrosis may occur but is uncommon.

Endocrine and Reproductive System The hypothalamic-pituitary axis is the part of the endocrine system most commonly involved; this condition usually presents as diabetes insipidus. Anterior pituitary dysfunction is also seen, manifesting as a deficiency in one or more pituitary hormones. Complete hypopituitarism is rare. Much less frequently, sarcoidosis can cause primary dysfunction of other endocrine glands. Adrenal cortical involvement resulting in Addison's syndrome has been described. The reproductive organs may be involved, but infertility is rare. Pregnancy is not affected by sarcoidosis, and common with sarcoidosis who become pregnant usually improve during pregnancy. However, the disease may flare post partum; presumably this variation results from fluctuations in endogenous glucocorticoid production.

Exocrine Glands Parotid enlargement is a classic feature of sarcoidosis, but clinically apparent parotid involvement occurs in<10% of patients. Bilateral involvement is the rule. The gland is usually nontender, firm, and smooth. Xerostomia can occur; other exocrine glands are affected only rarely.

Gastrointestinal Tract Although sarcoidosis involvement of the gastrointestinal tract is found occasionally at autopsy, it rarely has clinical importance. Occasionally, patients have esophageal or gastric symptoms.

COMPLICATIONS

The respiratory tract abnormalities cause most of the morbidity and mortality associated with sarcoidosis. The major problems are those characteristic of interstitial lung disease (Chap. 259), particularly dyspnea and insufficient oxygen delivery to vital organs. Respiratory failure with carbon dioxide retention is rare. In some patients, lung destruction results in formation of bullae that may harbor mycetomas, which are usually aspergillomas; erosion into the parenchyma can result in massive bleeding. The most common complications apart from the lung are associated with the eye; however, with therapy, blindness is rare. Complications of other organs include a gamut of abnormalities. The most serious are central nervous system (CNS) lesions or cardiac involvement leading to congestive heart failure or sudden death.

LABORATORY ABNORMALITIES

Common abnormalities in the blood include lymphocytopenia, an occasional mild eosinophilia, an increased erythrocyte sedimentation rate, hyperglobulinemia, and an elevated level of angiotensin-converting enzyme (ACE). False-positive tests for rheumatoid factor or antinuclear antibodies can be observed. Hypercalcemia is rare. Other serum abnormalities relate to involvement of specific organs such as liver, kidney, or endocrine glands.

Because the lung is involved so commonly, the routine chest film is almost always abnormal (Fig. 318-2*A*). The three classic x-ray patterns of pulmonary sarcoidosis are type I -- bilateral hilar adenopathy with no parenchymal abnormalities; type II -- bilateral hilar adenopathy with diffuse parenchymal changes; and type III -- diffuse parenchymal changes without hilar adenopathy. The type III pattern is sometimes split into two categories, with films that show fibrosis and upper lobe retraction classified separately. Although patients with type I x-ray patterns tend to have the acute or subacute, reversible form of the disease while those with types II and III often have the chronic, progressive disease, these patterns do not represent consecutive "stages" of sarcoidosis. Thus, except for epidemiologic purposes, this x-ray categorization is mostly of historic interest. The hilar adenopathy is almost always bilateral, but unilateral node enlargement can be seen. Nodes are also common in the paratracheal region. The diffuse parenchymal changes are typically reticulonodular infiltrates, but an acinar pattern is observed occasionally. Large nodules, similar to those of metastatic disease, are unusual but can occur. When there is massive fibrosis, the hila are pulled upward and there are conglomerate masses in the midlung zones. Some of the unusual chest x-ray findings in sarcoidosis include "egg shell" calcification of hilar nodes, pleural effusions, cavitation, atelectasis, pulmonary hypertension, pneumothorax, and cardiomegaly. Computed tomography of the chest is rarely helpful for either diagnosis or prognosis but can identify early fibrosis, and a "ground-glass" appearance is thought to be consistent with an active alveolitis.

The lung function abnormalities of sarcoidosis are typical for interstitial lung disease (Chap. 259) and include decreased lung volumes and diffusing capacity with a normal or supernormal ratio of the forced expiratory volume in 1 s to the forced vital capacity. Occasionally there is evidence of airflow limitation. There is usually mild hypoxemia and a mild, compensated hypocarbia.

The gallium-67 lung scan is usually abnormal, showing a pattern of diffuse uptake. If present, enlarged nodes are detected in these scans, as is inflammation in a variety of extrathoracic sites that usually have no clinical importance (Fig. 318-2*B*). Bronchoalveolar lavage typically demonstrates an increased proportion of lymphocytes, most of which are members of the activated T_H1 subset of CD4+ T lymphocytes (Fig. 318-2*C*). The remaining cells are mostly alveolar macrophages. In patients with significant fibrosis, a few neutrophils are also found. Eosinophils are rare.

The other laboratory features of sarcoidosis depend on the specific organ involved.

DIAGNOSIS

For a typical case, the diagnosis of sarcoidosis is made by a combination of clinical, radiographic, and histologic findings (Fig. 318-3*A*). In a young adult with constitutional complaints, respiratory symptoms, erythema nodosum, blurred vision, and bilateral hilar adenopathy, the diagnosis is almost always sarcoidosis. Commonly, however, the findings are more subtle. Furthermore, because sarcoidosis can occur in almost any place in the body, like tuberculosis or syphilis, it can be confused with many other disorders. In this context, the differential diagnosis of sarcoidosis must cover a wide range. However, it is confused most commonly with neoplastic diseases such as lymphoma or with disorders characterized also by a mononuclear cell granulomatous inflammatory process, such as the mycobacterial and fungal disorders.

The presence of skin anergy is typical but not diagnostic of sarcoidosis. Individuals with sarcoidosis who develop active tuberculosis react strongly to skin tests with purified protein derivative. The Kveim-Siltzbach skin test, the intradermal injection of a heat-treated suspension of a sarcoidosis spleen extract which is biopsied 4 to 6 weeks later, yields sarcoidosis-like lesions in 70 to 80% of individuals with sarcoidosis, with <5% false-positive results. However, the material is not widely available, and with the use of the transbronchial biopsy to obtain lung parenchyma for diagnostic purposes, the Kveim-Siltzbach test is not in general use.

No blood findings are diagnostic of the disease. Serum levels ofACE are elevated in approximately two-thirds of patients with sarcoidosis. Approximately 5% of all positive tests are not sarcoidosis and are seen in a variety of disorders, including asbestosis, silicosis, berylliosis, fungal infection, granulomatous hepatitis, hypersensitivity pneumonitis, leprosy, lymphoma, and tuberculosis. Hypercalcemia or an elevated 24-h urine calcium level is consistent with the diagnosis but is not specific.

The chest x-ray cannot be used as the sole criterion for the diagnosis of sarcoidosis. While the finding of bilateral hilar adenopathy is the hallmark of this disease, a similar pattern is occasionally observed in lymphoma, tuberculosis, coccidioidomycosis, brucellosis, and bronchogenic carcinoma.

The pattern of the gallium-67 scan is not diagnostic for sarcoidosis, nor is the finding of an increased proportion of lymphocytes among the cells recovered by bronchoalveolar lavage. However, the typical patterns of these tests (Fig. 318-2*B* and *C*) put the diagnosis in the general category of granulomatous lung disorders.

Whether or not the presentation is "classic," biopsy evidence of a mononuclear cell granulomatous inflammatory process is mandatory to make a definitive diagnosis of sarcoidosis. Because the lung is involved so frequently, it is the most common site to be biopsied, usually through a fiberoptic bronchoscope. Less common, but acceptable, sites for biopsy are the hilar nodes (by mediastinoscopy), the skin, conjunctiva, or lip. Rarely, the spleen, intraabdominal nodes, muscle, parotid or other salivary glands, upper respiratory tract, or the heart is biopsied for diagnostic purposes. At any of these sites, the findings must include the typical noncaseating granulomas. However, although histologic evidence is mandatory for a definitive diagnosis of sarcoidosis, the histologic findings are not sufficiently specific to make the diagnosis by themselves, since noncaseating granulomas are found in a number of other diseases, including infections and malignancy. Furthermore, although the liver or scalene nodes often reveal "positive" biopsies in cases of sarcoidosis, noncaseating granulomas from other causes are so frequent in these sites that they are not considered acceptable sites for establishing the diagnosis. Thus the definitive diagnosis of sarcoidosis is based on the biopsy in the context of the history, physical examination, blood tests, x-ray, lung function, and, if available, gallium-67 scan and bronchoalveolar lavage. Patients with HIV infection commonly have lymphocytopenia, chest x-ray abnormalities, positive gallium-67 chest scans, and increased proportions of lavage lymphocytes (early in the course of the disease), and they can have lung granulomas; thus, serologic testing for HIV infection should always be done in individuals suspected of having sarcoidosis.

PROGNOSIS

Overall, the prognosis in sarcoidosis is good. Most individuals who present with the acute disease are left with no significant sequelae. Approximately half of all patients have some permanent organ dysfunction, but for most this is mild, stable, and progresses rarely. In ~15 to 20% of patients, the disease remains active or recurs intermittently. Death is attributable directly to the disease in ~10% of all those affected.

TREATMENT

The therapy of choice for sarcoidosis is glucocorticoids (Fig. 318-3*B*), Various other drugs have been tried, including indomethacin, oxyphenbutazone, chloroquine, hydroxychloroquine, methotrexate, *p*-aminobenzoate, allopurinol, levamisole, azothioprine, and cyclophosphamide; but there is no evidence, apart from anecdotal, uncontrolled reports, to support their efficacy. Cyclosporine is ineffective for the pulmonary manifestations of the disease; anecdotal reports suggest that it may be useful in extrathoracic sarcoid not responding to glucocorticoids.

The major problem in treating sarcoidosis is in deciding when to treat. Because the disease clears spontaneously in ~50% of patients, and because the permanent organ derangements often do not improve with glucocorticoid treatment, there is controversy among clinicians as to the criteria for treatment. However, there is no question that glucocorticoids suppress effectively the activated TH1 lymphocyte processes occurring at the sites of disease. Thus, the major problem in making decisions concerning therapy in sarcoidosis is to determine the extent and activity of the inflammatory process in the organs at greatest risk, such as the lung, eye, heart, and CNS.

For the lung, this is based on a combination of history, physical findings, chest x-ray, and pulmonary function tests. Centers that see large numbers of these individuals sometimes use criteria based on gallium-67 lung scans and bronchoalveolar lavage findings. The serum level ofACE has been suggested as a criterion for disease activity, but it is not specific for the lung. Unless the respiratory impairment is devastating, active pulmonary sarcoidosis is observed usually without therapy for 2 to 3 months; if the inflammation does not subside spontaneously, therapy is instituted. For the eye, decisions concerning therapy are based on slit-lamp examination and tests for visual acuity. For the heart andCNS, decisions are based on an estimate of the severity of the involvement; patients with minor dysfunction are usually observed, while patients with significant cardiac or neurologic abnormalities are treated. Usually, it is not necessary to treat the systemic symptoms, but occasionally the extent of the fevers, fatigue, and/or weight loss necessitate therapy.

The usual therapy for sarcoidosis is prednisone, 1 mg/kg, for 4 to 6 weeks, followed by a slow taper over 2 to 3 months. This regimen is repeated if the disease again becomes active. Alternate-day therapy is used by some clinicians, but there is no evidence that it is as effective. High-dose bolus intravenous glucocorticoids are used occasionally but are probably not as effective as oral therapy. There is no evidence that inhaled glucocorticoids are efficacious. Mild ocular disease responds usually to local therapy, but suppression of the uveitis often requires systemic glucocorticoids.

(Bibliography omitted in Palm version)

Back to Table of Contents

319. AMYLOIDOSIS - *Jean D. Sipe, Alan S. Cohen*

DEFINITION AND CLASSIFICATION

Amyloidosis results from the deposition of insoluble, fibrous amyloid proteins, mainly in the extracellular spaces of organs and tissues. Named by Virchow in 1854 on the basis of color after staining with iodine and sulfuric acid, all amyloid fibrils share an identical secondary structure, the b-pleated sheet conformation, and a unique ultrastructure. All amyloid deposits contain an identical nonfibrillar component, the pentraxin serum amyloid P (SAP), and are associated with glycosaminoglycans. Abnormal protein folding and assembly can also result in protein deposition (e.g., in brain or kidney) that lacks the classic fibrillar morphology of amyloid and the presence of SAP. Depending upon the biochemical nature of the amyloid precursor protein, amyloid fibrils can be deposited locally or may involve virtually every organ system of the body. Amyloid fibril deposition may have no apparent clinical consequences or may lead to severe pathophysiologic changes. Often the disease falls between these two extremes. Regardless of etiology, the clinical diagnosis of amyloidosis is usually not made until the disease is far advanced.

Although the fibril precursors differ in their amino acid sequences, the polypeptide backbones of these protein precursors assume similar fibrillar morphologies that render them resistant to proteolysis.

The amyloidoses are classified according to the biochemical nature of the fibril-forming protein (Table 319-1). *Systemic amyloidoses* include biochemically distinct forms that are neoplastic, inflammatory, genetic, or iatrogenic in origin, while *localized* or *organ-limited amyloidoses* are associated with aging and diabetes and occur in isolated organs, often endocrine, without evidence of systemic involvement.

Despite their biochemical and clinical differences, the various amyloidoses share common pathophysiologic features: (1) an amyloidogenic precursor in appropriate concentration; (2) appropriate host genetic background; (3) abnormalities in proteolysis of fibril precursors and nascent amyloid fibrils; and (4) alterations in extracellular matrix constituents such as glycosaminoglycans, including the presence of amyloid-enhancing factor and Apo E. The guidelines for nomenclature and classification of amyloid and amyloidosis were updated in 1998 by the Nomenclature Committee of the International Society for Amyloidosis (Table 319-1). Amyloid deposits should be classified using the capital letter A as the first letter of designation followed by the protein designation without any open space; for example, AL for amyloidosis involving immunoglobulin light chains.

ETIOLOGY AND PATHOGENESIS

Light Chain Amyloidosis (AL) The most common form of systemic amyloidosis seen in current clinical practice is AL (primary idiopathic amyloidosis, or that associated with multiple myeloma) resulting from fibril formation by monoclonal antibody light chains in primary amyloidosis and in some cases of multiple myeloma (Chap. 113). Fewer than 20% of patients with AL have myeloma. The rest have other monoclonal gammopathies, light chain disease, or even agammaglobulinemia (producing light chains, but not intact immunoglobulin). About 15 to 20% of patients with myeloma have amyloidosis. A monoclonal population of bone marrow plasma cells is present and consistently produces either small lambda or kappa fragments or immunoglobulins that are processed (cleaved) in an abnormal fashion by macrophage enzymes to produce the partially degraded light chains responsible for AL amyloidosis. Lambda chain class predominates over kappa in AL by a 2:1 ratio, whereas in multiple myeloma and normal immunoglobulin synthesis, the reverse is true. Indeed, almost all lambda VI family chains have been associated with amyloid. The primary structure of each amyloid-forming light chain is unique, reflecting the features of the B cell clone that produced it. In patients with multiple myeloma, light chains can be deposited as casts in kidney tubules or as punctate deposits on basement membranes. Also, nonfibrillar deposition diseases have been described; thus there are three forms of human light chain-associated renal and systemic diseases: AL amyloidosis, cast nephropathy, and light chain deposition disease. Rarely, heavy chain amyloid deposition has been reported.

Amyloid A Amyloidosis (AA) AA amyloidosis (secondary, reactive, or acquired amyloidosis) occurs most frequently as a complication of chronic inflammatory disease. Effective treatment of the underlying inflammatory condition has reduced incidence in developed countries. In the past in the United States, tuberculosis (Chap. 169), osteomyelitis (Chap. 129), and leprosy (Chap. 170) were the most common precipitating diseases, and they remain so in developing countries. During inflammation, proinflammatory cytokines such as interleukin (IL) 1, IL-6, and tumor necrosis factor (TNF) stimulate the synthesis in liver of serum amyloid A, an injury-specific component of high-density lipoprotein. Thus, effective treatment of the underlying inflammatory disorder blocks the stimulus for precursor synthesis. Familial deposition of the AA protein occurs in some groups of patients with familial Mediterranean fever (FMF) and Familial Hibernian Fever (FHF) (Chap. 289). Colchicine treatment has been very effective both in blocking attacks of FMF and in reducing the incidence of AA amyloidosis in association with FMF. FMF is an autosomal recessive disorder subdivided into phenotype I, with irregularly occurring fever and abdominal, chest, or joint pain, preceding or accompanying renal amyloid; and phenotype II, in which renal amyloidosis is the first or only manifestation of the disease (Chap. 289). FMF is caused by mutations (16 identified to date) in the gene designated *MEFV* that encodes a 781-amino-acid protein named *pyrin* that appears to be a transcription factor. There is a strong correlation between the M694V mutation in MEFV and development of amyloidosis. FHF is an autosomal dominant disorder characterized by missense mutations in the TNF receptor.

Heredofamilial Amyloidoses Heredofamilial amyloidoses other than the AA form associated withFMF andFHFprimarily involve the nervous system, and their mode of inheritance is autosomal dominant. Familial amyloid polyneuropathies (FAP) are dominant hereditary diseases affecting kinships originating in Portugal, Japan, Sweden, Finland, Greece, Italy, and elsewhere. FAP can be subclassified based on clinical symptoms and the biochemical nature of the fibrils; in nearly all cases the fibrils are variants of transthyretin (TTR), apolipoprotein AI, gelsolin, cystatin C, and rarely the a chain of fibrinogen A or lysozyme. The mutant proteins, although present from birth, are associated with a delayed onset of disease symptoms, usually after three to seven decades of life. The FAP transthyretin prototype is the lower limb neuropathy first described in Portugal. It has a poor prognosis and is characterized by progressively severe neuropathy, including marked autonomic nervous system involvement. In some of these individuals, bilateral "scalloped" pupils are pathognomonic of the disease.

ATTR The most frequently occurring form ofFAPinvolvesTTR, a 14-kDa protein originally described as prealbumin, that transports thyroxine and retinol-binding protein in the blood. The first mutation to be identified in Portuguese families and in families of Swedish origin was a single amino acid substitution, methionine for valine at position 30. To date, more than 60 TTR variants have been defined, several of which are nonamyloidogenic. Variant TTR gene carriers exhibit clinically heterogeneous amyloidoses according to the position and nature of the amino acid substitution. Substitution of proline for leucine at position 55 results in an early onset and rapidly progressing disease, whereas substitution of methionine for threonine at position 119 appears to protect against amyloid fibril formation. In Denmark, patients with a methionine substitution for leucine at position 111 have a severe cardiopathy. Nonpathogenic TTR mutants such as the substitution of serine for glycine at position 6 also exist, and several are associated with changes in association with retinol-binding protein.

AApoAI Deposition of one of five apolipoprotein AI variants (G26R, W50R, L60R, L90P, and deletion of residues 61-7 with VT inserts) can be associated with peripheral neuropathy that is clinically similar to the type of familial amyloidosis that is caused by variants ofTTR. In some kindreds, the clinical presentation is renal failure without neurologic symptoms.

AGel A unique form of hereditary systemic amyloidosis has been reported primarily in Finland but also in patients of Japanese and Dutch backgrounds. Fibrils of gelsolin fragments, a calcium-binding protein that binds to and fragments actin filaments, are deposited in blood vessels and basement membranes, leading to clinical manifestations of lattice corneal dystrophy and cranial neuropathy, followed by peripheral neuropathy, dystrophic skin changes, and involvement of other organs. Two mutations at position 187, within the actin-binding domain of gelsolin, are associated with the disease.

ALys Hereditary nonneuropathic systemic amyloidosis has been described in English families in which lysozyme is the major fibril protein. Two mutations have been described -- I56T and D67H.

AFib Hereditary nonneuropathic renal amyloidosis has been described in families with one of three mutations in the fibrinogen A a chains, R524L, E526V, or R554L.

Ab$_2$M In long-term hemodialysis, amyloidosis is now well recognized as a serious bone and joint complication.b$_2$-microglobulin is the major constituent of the amyloid fibrils, and formation of advanced glycation end products ofb$_2$-microglobulin has been implicated in the pathogenesis of Ab$_2$M.

Localized or Organ-Limited Amyloidoses Depending upon the biochemical nature of the amyloid fibril protein, instead of systemic deposition involving the cardiovascular and gastrointestinal systems along with lymph nodes, spleen, liver, kidneys, and adrenals, amyloid deposition may be limited to a single organ such as the pancreas, brain, or heart. Recently, lactoferrin has been found to occur as amyloid fibrils in a rare form of corneal amyloidosis, and amyloid fibrils of prolactin have been identified in the pituitary gland and in a prolactin-producing tumor.

Polypeptide Hormone-Derived Amyloidosis Amyloid deposits are common in polypeptide hormone-producing tissues and tumors. Calcitonin is deposited in the hereditary amyloid syndrome, medullary carcinoma of the thyroid (ACal) (Chap. 330). Also AANF (atrial natriuretic factor-derived) amyloid deposits are found in the sarcolemma of ~80% of persons over 80 years of age. AIAPP (islet amyloid polypeptide-derived, or amylin) is deposited as amyloid fibrils in 90% of individuals with type 2 diabetes (Chap. 333), in endocrine tumors (Chap. 93), and in insulinoma (Chap. 93). It is produced in bcells of the pancreas and stored and released together with insulin. Human insulin does not naturally form amyloid fibrils, although fibrils of porcine insulin, AIns, are sometimes found as subcutaneous nodules at sites of insulin injection in diabetic individuals.

Amyloidosis Associated with Alzheimer's Disease A novel protein, b-amyloid protein (Ab), is the major fibril protein in the amyloid deposits of the cerebrovascular walls and the cores of the neuritic plaques of Alzheimer's disease (AD) patients and also in individuals with Down's syndrome (Chap. 66). The intracellular neurofibrillary tangles are composed of paired helical filaments arranged in a twisted conformation and have as their major component an abnormally phosphorylatedt-protein, a microtubule-associated protein whose semantic relation to the Ab of AD is arguable. Ab varies in length from 39 to 43 amino acids and is derived from a large transmembrane glycoprotein called amyloidb-precursor protein (AbPP). Mutations in AbPP are associated with familial AD and also with a different type of amyloidosis, hereditary cerebral hemorrhage with amyloidosis (Dutch type). Other forms of familial AD are associated with mutations in genes that encode presenilin proteins.

Prion Diseases Prions are a unique class of infectious proteins associated with a group of neurodegenerative diseases, the transmissible spongiform encephalopathies. In humans, these diseases include kuru, Creutzfeldt-Jakob disease, Gerstmann-Straussler-Scheinker syndrome, and fatal familial insomnia (Chap. 373); in animals, scrapie and bovine spongiform encephalopathy (mad cow disease). PrPSc is a pathogenic, transmissible spongiform encephalopathy-specific form of the host-encoded prion protein (PrP); PrPScdiffers from PrP in that it contains a high amount ofb-pleated sheet structure and is insoluble and resistant to proteolytic enzymes. PrPScdeposits either consist of or can be readily converted to amyloid fibrils. APrP is similar to Ab and ATTR in that both familial and sporadic forms occur. In addition, infectious prior diseases have resulted from the transmission of PrPSc by ritualistic cannibalism, corneal transplantation, treatment with cadaveric human growth hormone, and a variety of neurosurgical procedures. It has been suggested that the earlier onset familial forms of amyloidosis are due to accelerated fibril formation from mutant precursors, whereas in sporadic cases, amyloid fibrils are formed more slowly from normal precursor molecules. The mutant PrP molecules are nearer the threshold for transition to the amyloidogenic PrPScthan are the normal. The transition from normal to amyloidogenic PrPSc is irreversible but very slow. The disease progresses because, once formed, amyloidogenic PrPSccan seed the conversion of normal molecules into an amyloidogenic form.

CLINICAL MANIFESTATIONS

The clinical manifestations of amyloidosis are varied and depend entirely on the biochemical nature of the fibril protein and thus the area of the body that is involved (Table 319-2). The diagnosis of amyloidosis is usually not made until after the point of irreversible organ damage. Proteinuria is often the first symptom associated with systemic amyloidosis, particularly of the AA and AL type; peripheral neuropathies are associated withFAP, and dementia and cognitive dysfunction with amyloid deposits in brain. Organ enlargement, especially of the liver, kidney, spleen, and heart, may be prominent; however, this does not occur in FAP,AD, or PrP diseases.

Kidney Renal involvement may consist of mild proteinuria or frank nephrosis. In some cases, the urinary sediment may show a few red blood cells. The renal lesion is usually not reversible and in time leads to progressive azotemia and death. The prognosis does not appear to be related to the degree of the proteinuria; when azotemia finally develops, the prognosis is grave. Treatment by peritoneal or hemodialysis or kidney transplantation improves the prognosis considerably. Hypertension is rare, except in long-standing amyloidosis. Renal tubular acidosis or renal vein thrombosis may occur. Localized accumulation of amyloid may be noted in the ureter, bladder, or other parts of the genitourinary tract.

Heart Cardiac amyloidosis can present as intractable heart failure. Electrocardiographic abnormalities include a low-voltage QRS complex and abnormalities in atrioventricular and intraventricular conduction, often resulting in varying degrees of heart block. Owing to their propensity to develop conduction defects and arrhythmias, patients with cardiac amyloidosis appear to be especially sensitive to digitalis, and this drug should be used with caution.

With respect to systemic amyloidoses, cardiac amyloidosis is common in primary (AL) and heredofamilial amyloidosis and very rare in the secondary (AA) form. With respect to localized amyloidosis, cardiac amyloidosis of the wild type or nonvariantTTR type is common after 80 years of age; also atrial natriuretic factor may be present in the atria. In systemic amyloidosis, cardiac manifestations consist primarily of congestive failure and cardiomegaly (with or without murmurs) and a variety of arrhythmias and are comparable in AL andFAP, the predominant forms with cardiomyopathy (Chap. 238). Although these manifestations predominantly reflect diffuse myocardial amyloid, the endocardium, valves, and pericardium may also be involved. Pericarditis with effusion is rare, although the differential diagnosis of constrictive pericarditis versus restrictive cardiomyopathy frequently arises. Echocardiography has demonstrated symmetric thickening of the left ventricular wall, hypokinesia and decreased systolic contraction and thickening of the interventricular septum and left ventricular posterior wall, and left ventricular cavities of small to normal size. Two-dimensional echocardiography produces the characteristic findings of thickened right and left ventricles, a normal left ventricular cavity, and, especially, a diffuse hyperrefractile "granular sparkling" appearance. Hearts that are heavily infiltrated with amyloid may or may not show an enlarged silhouette. Fluoroscopy usually shows decreased mobility of the ventricular wall; angiographic studies usually demonstrate thickened ventricular wall, decreased ventricular mobility, and absence of rapid ventricular filling in early diastole.

Liver While hepatic involvement is common except in heredofamilial amyloidosis of theTTRtype, liver function abnormalities are minimal and occur late in the disease. Portal hypertension occurs but is uncommon. Intrahepatic cholestasis has been noted in about 5% of patients with AL (primary) amyloidosis. Hepatomegaly is common, and AL hepatic amyloid is usually accompanied by the nephrotic syndrome and congestive heart failure with poor prognosis. Amyloidosis of the spleen characteristically is not associated with leukopenia and anemia.

Skin Involvement of the skin is one of the most characteristic manifestations of primary (AL) amyloidosis (Chap. 57). Other forms of amyloidosis such as lichen amyloidosis are thought to involve forms of keratin. In AL amyloidosis, the usually nonpruritic lesions may consist of slightly raised, waxy papules or plaques that are usually clustered in the folds of the axillae, anal, or inguinal regions; the face and neck; or mucosal areas such as ear or tongue. Periorbital ecchymoses ("black eye" or "raccoon syndrome") have been reported.

Gastrointestinal Tract Gastrointestinal symptoms are common in all systemic types of amyloidosis either from direct involvement of the gastrointestinal tract at any level or from infiltration of the autonomic nervous system with amyloid. Symptoms include obstruction, ulceration, malabsorption, hemorrhage, protein loss, and diarrhea (Chap. 286). Infiltration of the tongue is characteristic of primary amyloidosis (AL) or amyloidosis accompanying multiple myeloma and occasionally leads to macroglossia (Fig. 319-CD1). When not enlarged, the tongue may become stiffened and firm to palpation. Gastrointestinal bleeding may occur from any of a number of sites, notably the esophagus, stomach, or large intestine, and may be severe. Amyloid infiltration of the esophagus may lead to an incompetent or nonrelaxing lower esophageal sphincter, nonspecific motility disorders of the esophageal body, or rarely achalasia. Small-bowel lesions may lead to clinical and x-ray changes of obstruction. A malabsorption syndrome is common. Amyloidosis (AA or secondary) may also develop in association with other entities involving the gastrointestinal tract, especially tuberculosis (Chap. 169), granulomatous enteritis (Chap. 287), lymphoma (Chap. 112), and Whipple's disease (Chap. 286); differentiation of these conditions, which give rise to secondary amyloidosis, from diffuse primary amyloidosis of the small bowel may be difficult. Similarly, amyloidosis of the stomach may closely mimic gastric carcinoma, with obstruction, achlorhydria, and the radiologic appearance of tumor masses.

Nervous System Neurologic manifestations, especially prominent in the heredofamilial amyloidoses may include peripheral neuropathy, postural hypotension, inability to sweat, Adies's pupil, hoarseness, and sphincter incompetence. The cranial nerves are generally spared, except in the Finnish hereditary amyloidosis. Carpal tunnel syndrome may be caused by several amyloidoses, especially primary (AL) and chronic hemodialysis (Ab$_2$M) amyloid. Peripheral neuropathy is frequent in the former type. Ab amyloid occurs in the central nervous system as a component of senile plaques and in blood vessels ("congophilic angiopathy"). The protein concentration in the cerebrospinal fluid may be increased. Infiltrates of the cornea or vitreous body may be present in hereditary amyloid syndromes. Certain of these syndromes (advancedFAP) are characterized by a bilateral scalloping appearance of the pupil.

Endocrine Amyloid may infiltrate the thyroid or other endocrine glands but rarely causes endocrine dysfunction. Local amyloid deposits almost invariably accompany medullary carcinoma of the thyroid. Amyloid is often found in the adrenal gland, pituitary gland, and pancreas. Pancreatic islet amyloid as a complication of type 2 diabetes is especially prominent and is caused by the b cell peptide islet amyloid polypeptide. Little if any clinical dysfunction is present unless there is massive replacement of the gland by amyloid.

Joints and Muscles Amyloid can directly, although rarely, involve articular structures by its presence in the synovial membrane and synovial fluid or in the articular cartilage. In these cases it is almost always of the AL type and associated with multiple myeloma. Amyloid arthritis can mimic a number of the rheumatic diseases because it can present as a symmetric arthritis of small joints with nodules, morning stiffness, and fatigue (Chap. 320). The synovial fluid usually has a low white blood cell count, a good to fair mucin clot, a predominance of mononuclear cells, and no crystals. Studies of surgical specimens suggest a significant incidence of amyloid in cartilage, capsule, and synovium in osteoarthritis (Chap. 321). Amyloid infiltration of muscle may lead to a pseudomyopathy. Shoulder muscle infiltration can produce the "shoulder pad" sign. Amyloid is found in muscle inclusion body disease, where Ab and/or PrP have been identified.

Deposition ofb_2-microglobulin as amyloid fibrils in the musculoskeletal systems is a serious complication of long-term hemodialysis. Ab_2M presents as the carpal tunnel syndrome, cystic bone lesions, and even destractive spondyloarthropathy.

Respiratory System The nasal sinuses, larynx, and trachea may be involved by accumulation of AL amyloid, which blocks the ducts, in the case of the sinuses, or the air passages. Amyloidosis of the lung involves the bronchi and alveolar septa diffusely. The lower respiratory tract is affected most frequently in primary (AL) amyloidosis and in the disease associated with dysproteinemia. Pulmonary symptoms attributable to amyloid are present in about 30% of cases. Amyloid may be localized in the bronchi or pulmonary parenchyma and may resemble a neoplasm. In these cases, local excision should be attempted and, when successful, may be followed by prolonged remissions.

Hematopoietic System Hematologic changes may include fibrinogenopenia, increased fibrinolysis, and selective deficiency of clotting factors. Deficient factor X seems to be due to nonspecific calcium-dependent binding to the polyanionic amyloid fibrils. Splenectomy in the patient with such a factor X deficiency can relieve the deficiency and the associated bleeding disorder, since factor X has been shown to bind to the large masses of splenic amyloid. Endothelial damage together with the clotting abnormalities lead to a propensity toward abnormal bleeding.

DIAGNOSIS

Amyloid fibrils are identified in biopsy or necropsy tissue sections (Table 319-3). The systemic amyloidoses offer a choice of biopsy sites; abdominal fat aspirates or renal or rectal biopsies are often performed. Microscopically, amyloid deposits stain pink with the hematoxylin-eosin stain and show metachromasia with crystal violet. The widely used and useful Congo red stain imparts a unique green birefringence when stained tissue sections are viewed using the polarizing microscope (Fig. 319-1). Fluorescent dyes such as thioflavin are sensitive screening stains for amyloid deposits in brain and other tissues; however, specificity should be confirmed. After amyloid has been identified by staining, it can be chemically classified by genomic DNA and protein studies and by immunohistochemistry. In the case ofFAP, the presence of mutantTTR (or gelsolin, Apo AI, etc.) establishes the specific diagnosis of the disease. Isoelectric focusing is used as a simple screening test for variant transthyretins associated with familial TTR amyloidosis. In order to establish the relationship of immunoglobulin-related amyloid to multiple myeloma, electrophoretic and immunoelectrophoretic studies on serum and urine should be performed when the biopsy reveals amyloid deposition. Most of these patients will have only relatively small paraprotein components, and only a few will have frank multiple myeloma.

PROGNOSIS

Generalized amyloidosis is usually a slowly progressive disease that leads to death in several years, but in some instances, prognosis is improving. The average survival in most large series of AL amyloid is ~12 months and inFAP is ~7 to 15 years. A number of individuals with amyloid have been followed 5 to 10 years and longer. The course of amyloidosis is difficult to document, because dating the time of origin of the disease is rarely possible. When amyloidosis develops in patients with rheumatoid arthritis, it seldom becomes evident when the arthritis is of less than 2 years' duration. When amyloidosis develops in patients with multiple myeloma, manifestations leading to initial hospitalization are more apt to be related to amyloid disease than to myeloma. In these cases, prognosis is very poor, and life expectancy is usually less than 6 months.

TREATMENT

Rational therapy should be directed at (1) reducing precursor production, (2) inhibiting the synthesis and extracellular deposition of amyloid fibrils, and (3) promoting lysis or mobilization of existing amyloid deposits. There are new specific therapies for the various amyloidoses. In certain of the heredofamilial amyloidoses, genetic counseling is an important aspect of treatment, and the removal of the site of synthesis of the mutant protein by liver transplantation has proven remarkably successful. Liver transplantation has been carried out since 1990 forFAPpatients in Sweden, the United States, Portugal, Spain, and other countries. It appears that disease progression is halted and that there is some improvement in autonomic nervous system function. The utilization of chronic hemodialysis and of kidney transplantation has clearly improved the prognosis of renal amyloid.

In the case of AL amyloid, the fact that immunoglobulin light chain is made by plasma cells has led to the use of alkylating agents. However, these agents are toxic and not very effective. The most effective form of treatment currently is stem cell transplantation and immunosuppressive drugs (melphalan). Several long-term remissions have been reported, but serious complications, even death, can occur. A novel anthracycline, iododoxorubicin (IDOX), has been shown to bind to AL amyloid (similar to Congo red) in vivo and promote amyloid resorption. A subset of AL patients responds transiently to this experimental agent; and it is thought that IDOX may prove useful in combination with other forms of treatment. Cardiac tranplantation in selected cases of AL orFAPamyloidosis has its advocates and has been successful.

Colchicine has been shown to be effective in preventing acute attacks and amyloidosis in patients withFMF(Chap. 289).

The major causes of death are heart disease and renal failure. Sudden death, presumably due to arrhythmias, is common. Occasionally, gastrointestinal hemorrhage, respiratory failure, intractable heart failure, and superimposed infections are the terminal events.

(Bibliography omitted in Palm version)

Back to Table of Contents

SECTION 3 -DISORDERS OF THE JOINTS

320. APPROACH TO ARTICULAR AND MUSCULOSKELETAL DISORDERS - *John J. Cush*, *Peter E. Lipsky*

Musculoskeletal complaints account for more than 315,000,000 outpatient visits per year. Many of the musculoskeletal complaints that cause patients to seek medical attention are related to self-limited conditions requiring minimal evaluation and only symptomatic therapy and reassurance. However, some patients with similar symptoms have a more serious condition that requires further evaluation or additional laboratory testing to confirm the suspected diagnosis or determine the extent and nature of the pathologic process. A primary objective is to determine if a "red flag" or urgent rheumatologic condition is present and, if not, to formulate a differential diagnosis that leads to accurate diagnosis and timely therapy while avoiding excessive diagnostic testing and unnecessary treatment (Table 320-1) There are several urgent conditions that must be diagnosed promptly to avoid significant morbid or mortal sequelae. These red flag diagnoses include septic arthritis, acute crystal-induced arthritis (e.g., gout), and fracture. Each of these may be suspected by an acute onset with a monoarticular or focal presenting complaint (see below).

Individuals with musculoskeletal complaints should be evaluated in a uniform, logical manner by means of a thorough history, a comprehensive physical examination, and, if appropriate, laboratory testing. The goals of the initial encounter are to determine whether the musculoskeletal complaint is (1) *articular* or *nonarticular* in origin, (2) *inflammatory* or *noninflammatory* in nature, (3) *acute* or *chronic* in duration, and (4) *localized* or *widespread* (*systemic*) in distribution.

With such an approach and an understanding of the pathophysiologic processes that underlie musculoskeletal complaints, an adequate diagnosis can be made in the vast majority of individuals. However, some patients will not fit immediately into an established diagnostic category. Many musculoskeletal disorders resemble each other at the outset, and some take weeks or months to evolve into a readily recognizable diagnostic entity. This consideration should temper the desire always to establish a definitive diagnosis at the first encounter.

ARTICULAR VERSUS NONARTICULAR

The musculoskeletal evaluation must discriminate the anatomic site(s) of origin of the patient's complaint. For example, ankle pain can result from a variety of pathologic conditions involving disparate anatomic structures, including gonococcal arthritis, calcaneal fracture, Achilles tendinitis, cellulitis, and peripheral neuropathy. Articular structures include the synovium, synovial fluid, articular cartilage, intraarticular ligaments, joint capsule, and juxtaarticular bone. Nonarticular (or periarticular) structures, such as supportive extraarticular ligaments, tendons, bursae, muscle, fascia, bone, nerve, and overlying skin, may be involved in the pathologic process. Pain from nonarticular structures may mimic true articular pain because of their proximity to the joint. Distinguishing between articular and nonarticular disease requires a careful and detailed examination. Articular disorders may be characterized by deep or diffuse joint pain, limited range of motion on active and passive movement, swelling caused by synovial proliferation or effusion or bony enlargement, crepitation, instability, locking, or deformity. By contrast, nonarticular disorders tend to be painful on active but not passive range of motion, demonstrate point or focal tenderness in regions distinct from articular structures, and have physical findings remote from the joint capsule. Moreover, nonarticular disorders seldom demonstrate crepitus, instability, deformity, or swelling.

INFLAMMATORY VERSUS NONINFLAMMATORY

In the course of a musculoskeletal evaluation, the examiner should elicit symptoms and signs that will narrow or establish the diagnosis. A primary objective is to identify the nature of the underlying pathologic process. Musculoskeletal disorders are generally classified as inflammatory or noninflammatory. Inflammatory disorders may be infectious (infection with *Neisseria gonorrhoea* or *Mycobacterium tuberculosis*), crystal-induced (gout, pseudogout), immune-related [rheumatoid arthritis (RA), systemic lupus erythematosus (SLE)], reactive (rheumatic fever, Reiter's syndrome), or idiopathic. Inflammatory disorders may be identified by the presence of some or all of the four cardinal signs of inflammation (erythema, warmth, pain, and swelling), by systemic symptoms (prolonged morning stiffness, fatigue, fever, weight loss), or by laboratory evidence of inflammation (elevated erythrocyte sedimentation rate or C-reactive protein level, thrombocytosis, anemia of chronic disease, or hypoalbuminemia). Articular stiffness is common in chronic musculoskeletal disorders. However, the chronology and magnitude of stiffness may be diagnostically important. Morning stiffness related to inflammatory disorders (such as RA) is precipitated by prolonged rest, often lasts several hours, and may improve with activity and anti-inflammatory medications. By contrast, intermittent stiffness associated with noninflammatory conditions, such as osteoarthritis, is precipitated by brief periods of rest, usually lasts less than 60 min, and is exacerbated by activity. Noninflammatory disorders may be related to trauma (rotator cuff tear), ineffective repair (osteoarthritis), cellular overgrowth (pigmented villonodular synovitis), or pain amplification (fibromyalgia). They are often characterized by pain without swelling or warmth, the absence of inflammatory or systemic features, little or no morning stiffness, and normal laboratory findings.

Identification of the nature of the underlying process and the site of the complaint will enable the examiner to narrow the diagnostic considerations and to assess the need for immediate diagnostic or therapeutic intervention or for continued observation.Figure 320-1 presents a logical approach to the evaluation of patients with musculoskeletal complaints.

CLINICAL HISTORY

Additional historic features may be helpful in establishing the nature and extent of the pathologic process and may provide important clues to the diagnosis. When evaluating patients with musculoskeletal complaints, the clinician should always consider the most common conditions (e.g., low back pain, osteoarthritis) seen in the general population (Fig. 320-2). Aspects of the patient profile, including age, sex, race, and family history, can provide important information. Certain diagnoses are more frequent in specific age groups.SLE, rheumatic fever, and Reiter's syndrome are more common in the young, whereas fibromyalgia andRA are most common in middle age, and osteoarthritis and polymyalgia rheumatica in the elderly. Some diseases are more common in a particular gender or race. Gout and the spondyloarthropathies (e.g., ankylosing spondylitis, Reiter's syndrome) are more common in men, whereas SLE, RA, and fibromyalgia are more common in women. Polymyalgia rheumatica, giant cell arteritis, and Wegener's granulomatosis preferentially affect whites, whereas sarcoidosis and SLE are more common in blacks. *Familial aggregation* occurs in some disorders, such as ankylosing spondylitis, gout, RA, and Heberden's nodes of osteoarthritis.

The chronology of the complaint (*onset*, *evolution*, and *duration*) is an important diagnostic feature. The onset of disorders such as septic arthritis and gout tends to be abrupt, whereas osteoarthritis,RA, and fibromyalgia may develop more indolently. In terms of evolution, disorders are classified as acute (e.g., septic arthritis), chronic (e.g., osteoarthritis), intermittent (e.g., gout), migratory (e.g., rheumatic fever, gonococcal or viral arthritis), or additive (e.g., RA, Reiter's syndrome). Musculoskeletal disorders typically are called *acute* if they last less than 6 weeks and *chronic* if they last longer. Acute and intermittent arthropathies tend to be infectious, crystal-induced, or reactive. Noninflammatory and immune-related arthritides, such as osteoarthritis and RA, respectively, are often chronic. The duration of the patient's complaints may alter the diagnostic considerations. For example, the musculoskeletal signs and symptoms of hepatitis B virus infection may be identical with those of early RA at the onset but rarely persist beyond 3 weeks.

The *number and distribution* of involved articulations should be noted. Articular disorders are classified as *monarticular* (one joint involved), *oligoarticular* or *pauciarticular* (two to three joints involved), or *polyarticular* (more than three joints involved). Nonarticular disorders can be classified as either *focal* or *widespread*. Complaints secondary to trauma and gout are typically focal or monarticular, whereas polymyositis,RA, and fibromyalgia are more diffuse or polyarticular. Joint involvement tends to be symmetric in RA but is often asymmetric in the spondyloarthropathies and in gout. The upper extremities are frequently involved in RA, whereas lower extremity arthritis is characteristic of Reiter's syndrome and gout at their onset. Involvement of the axial skeleton is common in osteoarthritis and ankylosing spondylitis but infrequent in RA, with the notable exception of the cervical spine.

The clinical history should also identify *precipitating events*, such as trauma, drug administration (Table 320-2), or antecedent or intercurrent illnesses, that may have contributed to the patient's complaint. Last, a thorough *rheumatic review of systems* may disclose associated features outside the musculoskeletal system and provide useful diagnostic information. A variety of musculoskeletal disorders may be associated with systemic features such as fever (SLE, infection), rash (SLE, Reiter's syndrome, dermatomyositis), myalgias, weakness (polymyositis, polymyalgia rheumatica), and morning stiffness (inflammatory arthritis). In addition, some conditions are associated with involvement of other organ systems, including the eyes (Behcet's disease, sarcoidosis, Reiter's syndrome), gastrointestinal tract (scleroderma, inflammatory bowel disease), genitourinary tract (Reiter's syndrome, gonococcemia, Behcet's disease), and nervous system (Lyme disease, SLE, vasculitis).

PHYSICAL EXAMINATION

The goal of the physical examination is to ascertain the structures involved, the nature of the underlying pathology, the extent and functional consequences of the process, and the presence of systemic or extraarticular manifestations. A knowledge of topographic anatomy is necessary to identify the primary site(s) of involvement and differentiate articular from nonarticular disorders. The musculoskeletal examination depends largely on careful inspection, palpation, and a variety of specific physical maneuvers to elicit diagnostic signs (Table 320-3). Although most articulations of the appendicular skeleton can be examined in this manner, adequate inspection and palpation are not possible for many axial (e.g., zygapophyseal) and inaccessible (e.g., sacroiliac or hip) joints. For such joints, there is a greater reliance on specific maneuvers and imaging for assessment.

Examination of involved and uninvolved joints will determine whether *warmth*, *erythema*, or *swelling* is present. The examination should distinguish true articular swelling caused by synovial effusion or synovial proliferation from nonarticular or periarticular involvement, which usually extends beyond the normal joint margins or the full extent of the synovial space. Synovial effusion can be distinguished from synovial hypertrophy or bony hypertrophy by palpation or specific maneuvers. For example, small to moderate knee effusions may be identified by the "bulge sign" or "ballottement of the patella." Bursal effusions (e.g., effusions of the olecranon or prepatellar bursa) overlie bony prominences and are fluctuant with sharply defined borders. Joint *stability* can be assessed by palpation and by the application of manual stress to assess displacement in different planes. Subluxation or dislocation, which may be secondary to traumatic, mechanical, or inflammatory causes, can be assessed by inspection and palpation. Joint *volume* can be assessed by palpation. Distention of the articular capsule usually causes pain. The patient will attempt to minimize the pain by keeping the joint in the position of least intraarticular pressure and greatest volume, usually partial flexion. Clinically, joint distention may be detected as obvious swelling, voluntary or fixed flexion deformities, or diminished range of motion -- especially on extension, which decreases joint volume. Active and passive *range of motion* should be assessed in all planes, with contralateral comparison. Serial evaluations of joint motion may be made using a goniometer to quantify the arc of movement. Each joint should be passively manipulated through its full range of motion (including, as appropriate, flexion, extension, rotation, abduction, adduction, inversion, eversion, supination, pronation, and medial or lateral deviation or bending). Limitation of motion is frequently caused by effusion, pain, deformity, or contracture. *Contractures* may reflect antecedent synovial inflammation or trauma. Joint *crepitus* may be felt during palpation or maneuvers and may be prominent or coarse in osteoarthritis. Joint *deformity* usually indicates a long-standing or aggressive pathologic process. Deformities may result from ligamentous destruction, soft tissue contracture, bony enlargement, ankylosis, erosive disease, or subluxation. Examination of the musculature will permit assessment of strength and reveal atrophy, pain, or spasm. The examiner should look carefully for nonarticular or periarticular involvement, especially when articular complaints are not supported by objective findings referable to the joint capsule. The identification of musculoskeletal pain of soft tissue origin (nonarticular pain) will prevent unwarranted and often expensive additional evaluations. Specific maneuvers may reveal nonarticular abnormalities, such as a carpal tunnel syndrome (which can be identified by Tinel's or Phalen's sign). Other examples of soft tissue abnormalities include olecranon bursitis, epicondylitis (tennis elbow), enthesitis (e.g., Achilles tendinitis), and trigger points associated with fibromyalgia.

LABORATORY INVESTIGATIONS

The vast majority of musculoskeletal disorders can be diagnosed easily by a complete history and physical examination. An additional objective of the initial encounter is to determine whether additional investigations or immediate therapy are required. A number of features indicate the need for additional evaluation. *Monarticular* conditions require additional evaluation, as do *traumatic* or *inflammatory* conditions and conditions accompanied by *neurologic changes* or *systemic manifestations* of serious disease. Finally, individuals with *chronic* symptoms (lasting more than 6 weeks), especially when there has been a lack of response to symptomatic measures, are candidates for additional evaluation. The extent and nature of the additional investigation should be dictated by the clinical features and suspected pathologic process. Laboratory tests should be used to confirm a specific clinical diagnosis and not be used as a tool to screen or evaluate patients with vague rheumatic complaints. Indiscriminate use of broad batteries of diagnostic tests and radiographic procedures are rarely useful or cost-effective.

Besides a complete blood count, including a white blood cell (WBC) and differential count, the routine evaluation should include determination of an acute-phase indicator, such as the erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP), which can be useful in discriminating inflammatory from noninflammatory musculoskeletal disorders. Both tests are inexpensive and easily performed; the resulting values may be elevated with infections, inflammatory arthritis, autoimmune disorders, neoplasia, pregnancy, and advanced age. Serum uric acid determinations are only useful when gout has been diagnosed and therapy contemplated.

Serologic tests for rheumatoid factor, antinuclear antibodies (ANA), complement levels, Lyme disease antibodies, or antistreptolysin O (ASO) titer should be carried out only when there is substantive clinical evidence suggesting a relevant associated diagnosis, as these tests have poor predictive value when used in a screening fashion, especially when the pretest probability is low. They should not be performed arbitrarily in patients with minimal or nonspecific musculoskeletal complaints. For example, 4 to 5% of the general population will have positive tests for rheumatoid factor and ANAs, yet only 1% or 0.04% will haveRA orSLE, respectively. IgM rheumatoid factor (autoantibodies against the Fc portion of IgG) is found in 80% of patients with RA and may also be seen in low titers in patients with chronic infections (tuberculosis, leprosy); other autoimmune diseases (SLE, Sjogren's syndrome); or chronic pulmonary, hepatic, or renal diseases. ANAs are found in nearly all patients with SLE and may also be seen in patients with other autoimmune diseases (polymyositis, scleroderma, antiphospholipid syndrome), drug-induced lupus (resulting from hydralazine, procainamide, or quinidine administration), or chronic hepatitic or renal disorders. The interpretation of a positive ANA determination may depend on the titer and on the pattern observed by immunofluorescence microscopy. Diffuse and speckled patterns are most common but least specific, whereas a peripheral, or rim, pattern is highly specific and is suggestive of autoantibodies against double-stranded (native) DNA. This pattern is seen only in patients with SLE.

Aspiration and analysis of synovial fluid are always indicated in acute monarthritis or when an infectious or crystal-induced arthropathy is suspected. Synovial fluid analysis may be crucial in distinguishing between noninflammatory and inflammatory processes. This distinction can be made on the basis of the appearance, viscosity, and cell count of the synovial fluid. Tests for synovial fluid glucose, protein, lactate dehydrogenase, lactic acid, or autoantibodies are not recommended, as they are insensitive or have little discriminatory value. Normal synovial fluid is clear or a pale straw color and is viscous, primarily because of the high levels of hyaluronate. Noninflammatory synovial fluid is clear, viscous, and amber-colored, with aWBCcount of <2000/uL and a predominance of mononuclear cells. The viscosity of synovial fluid is assessed by expressing fluid from the syringe one drop at a time. Normally there is a stringing effect, with a long tail behind each drop. Effusions due to osteoarthritis or trauma usually have normal viscosity. Inflammatory fluid is turbid and yellow, with an increased WBC (2000 to 50,000/uL) and a predominance of polymorphonuclear leukocytes. Inflammatory fluid has a reduced viscosity, diminished hyaluronate, and little or no tail following each drop of synovial fluid. Such effusions are found inRA, gout, other inflammatory arthritides, and septic arthritis. Infectious fluid is turbid and opaque, with a WBC count usually>50,000/uL, a predominance of polymorphonuclear leukocytes (>75%), and low viscosity. Such effusions are typical of septic arthritis, but they occur rarely with sterile inflammatory arthritides such as RA or gout. In addition, hemorrhagic synovial fluid may be seen with trauma, hemarthrosis, or neuropathic arthritis. An algorithm for synovial fluid aspiration and analysis is shown inFig. 320-3. Synovial fluid should be analyzed immediately for appearance, viscosity, and cell count. Cellularity and the presence of crystals may be assessed by light or polarizing microscopy, respectively. Monosodium urate crystals, seen in gouty effusions, are long, needle-shaped, negatively birefringent, and usually intracellular, whereas calcium pyrophosphate dihydrate crystals, found in chondrocalcinosis and pseudogout, are usually short, rhomboid-shaped, and positively birefringent. Whenever infection is suspected, synovial fluid should be Gram-stained and cultured appropriately. If gonococcal arthritis is suspected, immediate plating of the fluid on appropriate culture medium is indicated. Synovial fluid from chronic monarthritis patients should also be cultured for *M. tuberculosis* and fungi. Last, it should be noted that crystal-induced arthritis and infection occasionally occur together in the same joint.

DIAGNOSTIC IMAGING IN JOINT DISEASES

Conventional radiography has been a valuable tool in the diagnosis and staging of articular disorders. Plain x-rays are most appropriate when there is a history of trauma, suspected chronic infection, progressive disability, or monarticular involvement; when therapeutic alterations are considered; or when a baseline assessment is desired for what appears to be a chronic process. However, in most inflammatory disorders, early radiography is rarely helpful in establishing a diagnosis and may only reveal soft tissue swelling or juxtaarticular demineralization. As the disease progresses, calcification (of soft tissues, cartilage, or bone), joint space narrowing, erosions, bony ankylosis, new bone formation (sclerosis, osteophyte formation, or periostitis), or subchondral cysts may develop and suggest specific clinical entities. Consultation with a radiologist will help define proper technique and positioning and prevent the need for further studies.

Additional imaging techniques may possess greater diagnostic sensitivity and facilitate early diagnosis in a limited number of articular disorders and are indicated in selected circumstances when conventional radiography is not adequate (Table 320-4). *Ultrasonography* is useful in the detection of soft tissue abnormalities that cannot be appreciated fully by clinical examination. Although ultrasonography is inexpensive and easily performed, only in a limited number of circumstances is it the preferred method of evaluation. The foremost application of ultrasound is in the diagnosis of synovial (Baker's) cysts, although rotator cuff tears and various tendon injuries may be evaluated with ultrasound by an experienced operator. *Radionuclide scintigraphy* provides useful information regarding the metabolic status of bone and, along with radiography, is well suited for total-body assessment of the extent and distribution of musculoskeletal involvement. It is a very sensitive but poorly specific means of detecting inflammatory or metabolic alterations in bone or periarticular soft tissue structures. The limited tissue resolution of scintigraphy may obscure the distinction between bony and periarticular processes and may necessitate the use of additional imaging modalities. Scintigraphy, using^{99m}Tc,^{67}Ga, orWBCslabeled with^{111}In, has been applied to a variety of articular disorders with variable success (Table 320-4). [^{99m}Tc]pertechnetate or [^{99m}Tc]diphosphonate scintigraphy may be useful in identifying infection, neoplasia, inflammation, increased blood flow, bone remodeling, heterotopic bone formation, or avascular necrosis (Fig. 320-4). However, the poor specificity of^{99m}Tc scanning has limited its use to investigational and serial assessments of joint or bone involvement, assessment of inflammatory or infectious processes, and surveys for bone metastases.^{67}Ga binds to serum and cellular transferrin and lactoferrin and is preferentially taken up by neutrophils, macrophages, bacteria, and tumor tissue (e.g., lymphoma) and is useful in the identification of infection and malignancies. Scanning with^{111}In-labeled WBCs has been used to detect both infectious and inflammatory arthritis. Although both have been used with success,^{111}In-labeled WBC scanning is superior to^{67}Ga in the early diagnosis of osteomyelitis and infected prosthetic joints. Prior treatment with antibiotics may reduce the diagnostic sensitivity of both^{67}Ga and^{111}In-labeled WBC scintigraphy.

Computed tomography (CT) provides rapid reconstruction of sagittal, coronal, and axial images and thus of the spatial relationships among anatomic structures. It has proved most useful in the assessment of the axial skeleton because of its ability to visualize in the axial plane. Articulations that are difficult to visualize by conventional radiography, such as the zygapophyseal, sacroiliac, sternoclavicular, and hip joints, can be evaluated effectively using CT. CT has been demonstrated to be useful in the diagnosis of low back pain syndromes, sacroiliitis, osteoid osteoma, tarsal coalition, osteomyelitis, intraarticular osteochondral fragments, and advanced osteonecrosis.

Magnetic resonance imaging (MRI) has significantly advanced the ability to image musculoskeletal structures. MRI can provide multiplanar images with fine anatomic detail and contrast resolution (Fig. 320-5). Other advantages are the absence of ionizing radiation and adverse effects and the superior ability to visualize bone marrow and soft tissue periarticular structures. However, the high cost and long procedural time of MRI limit its use in the evaluation of musculoskeletal disorders. MRI should be used only when it will provide necessary information that cannot be obtained by less expensive and noninvasive means.

MRIcan image fascia, vessels, nerve, muscle, cartilage, ligaments, tendons, pannus, synovial effusions, cortical bone, and bone marrow. Visualization of particular structures can be enhanced by altering the pulse sequence to produce either T1-weighted or T2-weighted spin echo, gradient echo, or inversion recovery [including short tau inversion recovery (STIR) images. Because of its sensitivity to changes in marrow fat, MRI is a sensitive although nonspecific means of detecting osteonecrosis and osteomyelitis (Fig. 320-5). Because of its enhanced soft tissue resolution, MRI is more sensitive than arthrography orCT for the diagnosis of soft tissue injuries (e.g., meniscal and rotator cuff tears), intraarticular derangements, and spinal cord damage following injury, subluxation, or synovitis of the vertebral facet joints.

RHEUMATOLOGIC EVALUATION OF THE ELDERLY

Musculoskeletal disorders in elderly patients are often not diagnosed because the signs and symptoms may be insidious or chronic in these patients. In addition, the nature of the problem is often obscured by the presence of multiple interacting factors, including other medical conditions and therapies. These difficulties are compounded by the diminished reliability of laboratory testing in the elderly, who often manifest nonpathologic abnormal results. For example, erythrocyte sedimentation rates may be misleadingly elevated and low titer positive tests for rheumatoid factor and ANAs may be seen in up to 15% of elderly patients. Although nearly all rheumatic disorders can afflict the elderly, certain diseases and drug-induced disorders (Table 320-2) are more common in this age group. The elderly should be approached in the same manner as other patients with musculoskeletal complaints but with additional inquiries to exclude common geriatric musculoskeletal disorders. An emphasis on identifying the rheumatic consequences of intercurrent medical conditions and therapies is extremely important. Osteoarthritis, gout, polymyalgia rheumatica, drug-induced lupus erythematosus, and chronic salicylate toxicity are all more common in the elderly than in other individuals. The physical examination should identify the nature of the musculoskeletal complaint, as well as coexisting diseases that may influence the diagnosis and choice of treatment.

Approach to the Patient

Regional Rheumatic Complaints Although all patients should be evaluated in a logical and thorough manner, many cases of focal musculoskeletal complaints are caused by commonly encountered disorders that exhibit a predictable pattern of onset, evolution, and localization and that can often be diagnosed immediately on the basis of limited historic information and selected maneuvers or tests. Although nearly every joint can be approached in this manner, the evaluation of four commonly involved anatomic regions -- the hand, shoulder, hip, and knee -- are reviewed here.

HAND PAIN Focal or unilateral hand pain may result from trauma, overuse, infection, or a reactive or crystal-induced arthritis. By contrast, bilateral hand complaints suggest a degenerative (e.g., osteoarthritis), systemic, or inflammatory/immune etiology. Patterns of joint involvement are highly suggestive of certain disorders. The distribution of affected joints in the hand may provide important diagnostic information (Fig. 320-6). Thus, osteoarthritis (or degenerative arthritis) may manifest as distal interphalangeal (DIP) and proximal interphalangeal (PIP) joint pain with bony hypertrophy sufficient to produce Heberden's and Bouchard's nodes, respectively. Pain, with or without bony swelling, involving the base of the thumb (first carpometacarpal joint) is also highly suggestive of osteoarthritis. By contrast,RA tends to involve the PIP, metacarpophalangeal, intercarpal, and carpometacarpal joints (wrist) with pain, prolonged stiffness, and palpable synovial tissue hypertrophy. Psoriatic arthritis may also involve the DIP and PIP joints and the carpus with inflammatory pain, stiffness, and synovitis. Moreover, the diagnosis of psoriatic arthritis can be suggested by nail pitting or onycholysis. Soft tissue swelling may also be noted over the dorsum of the hand and wrist and may suggest an inflammatory extensor tendon tenosynovitis, possibly caused by gonococcal infection, gout, or inflammatory arthritis. The diagnosis of tenosynovitis may be suggested by local warmth and edema and is confirmed when pain is induced by maintaining the wrist in a fixed, neutral position and flexing the digits distal to the metacarpophalangeal joints to stretch the extensor tendon sheaths.

Focal wrist pain localized to the radial aspect may be caused by DeQuervain's tenosynovitis resulting from inflammation of the tendon sheath(s) involving the abductor pollicis longus or extensor pollicis brevis (Fig. 320-6). This condition commonly results from overuse or develops after pregnancy and may be diagnosed with Finkelstein's test. A positive result in Finkelstein's test is present when local wrist pain is induced after the thumb is flexed across the palm and placed inside a clenched fist and the patient actively moves the hand downward with ulnar deviation at the wrist. Carpal tunnel syndrome is another common disorder of the upper extremity and results from compression of the median nerve within the carpal tunnel. Manifestations include paresthesias in the thumb and the second, third, and radial half of the fourth fingers, and sometimes, atrophy of thenar musculature. Carpal tunnel syndrome is commonly associated with pregnancy, edema, trauma, osteoarthritis, inflammatory arthritis, and infiltrative disorders (e.g., amyloidosis). The diagnosis is suggested by a positive Tinel's or Phalen's sign. With each test, paresthesia in a median nerve distribution is induced or increased by either "thumping" the volar aspect of the wrist (Tinel's sign) or pressing the extensor surfaces of the two flexed wrists against each another (Phalen's sign).

SHOULDER PAIN During the evaluation of shoulder disorders, the examiner should carefully note any history of trauma, infection, inflammatory disease, occupational hazards, or previous cervical disease. In addition, the patient should be questioned as to the activities or movement(s) that elicit shoulder pain. Shoulder pain is frequently referred from the cervical spine, but it may also be referred from intrathoracic lesions (e.g., a Pancoast tumor) or from gallbladder, hepatic, or diaphragmatic disease. The shoulder should be put through its full range of motion both actively and passively (with examiner assistance): forward flexion, extension, abduction, adduction, and rotation. Manual inspection of the periarticular structures will often provide important diagnostic information. The examiner should apply direct manual pressure over the subacromial bursa, which lies lateral to and immediately beneath the acromion. Subacromial bursitis is a frequent cause of shoulder pain. Anterior to the subacromial bursa, the bicipital tendon traverses the bicipital groove. This tendon is best identified by palpating it in its groove as the patient rotates the humerus internally and externally. Direct pressure over the tendon may reveal pain indicative of bicipital tendinitis. Palpation of the acromioclavicular joint may disclose local pain, bony hypertrophy, or synovial swelling. Whereas osteoarthritis andRAcommonly affect the acromioclavicular joint, osteoarthritis seldom involves the glenohumeral joint, unless there is a traumatic or occupational cause. The glenohumeral joint is best palpated anteriorly by placing the thumb over the humeral head (just medial and inferior to the coracoid process) and having the patient rotate the humerus internally and externally. Pain localized to this region is indicative of glenohumeral pathology. Synovial effusion or tissue is seldom palpable but, if present, may suggest infection, RA, or an acute tear of the rotator cuff.

Rotator cuff tendinitis or tear is a very common cause of shoulder pain. The rotator cuff is formed by the tendons of the supraspinatus, infraspinatus, teres minor, and subscapularis muscles. Rotator cuff tendinitis is suggested by pain on active abduction (but not passive abduction), pain over the lateral deltoid muscle, night pain, and evidence of the impingement sign. This maneuver is performed by the examiner raising the patient's arm into forced flexion while stabilizing the scapula and preventing it from rotating. A positive sign is present if pain develops before 180° of forward flexion. A complete tear of the rotator cuff, which often results from trauma, may manifest in the same manner but is less common than tendinitis. The diagnosis is suggested by the drop arm test, in which the patient is asked to maintain the arm outstretched after it has been passively abducted. If the patient is unable to hold the arm up once 90° of abduction is reached, the test is positive. Tendinitis or tear of the rotator cuff can be confirmed byMRI or ultrasonography.

KNEE PAIN A careful history should delineate the chronology of the knee complaint and whether there are predisposing conditions, trauma, or medications that might underlie the complaint. Observation of the patient's gait is also important. The knee should be carefully inspected in the upright (weight-bearing) and prone positions for swelling, erythema, contusion, laceration, and malalignment. The most common form of malalignment in the knee is genu varum (bow-legs) and genu valgum (knock-knees). Bony swelling of the knee joint commonly results from hypertrophic osseous changes seen with disorders such as osteoarthritis and neuropathic arthropathy. Swelling caused by hypertrophy of intrasynovial structures (synovial enlargement or effusion) may manifest as a fluctuant, ballotable, or soft tissue enlargement in the suprapatellar pouch (superior reflection of the synovial cavity) or lateral and medial to the patella. Synovial effusions may also be detected by balloting the patella downward toward the femoral groove or by eliciting a bulge sign. To elicit this sign, the examiner positions the knee in extension and manually compresses or milks synovial fluid down from the suprapatellar pouch and lateral to the patellae. Manual pressure lateral to the patella may cause an observable shift in synovial fluid (bulge) to the medial aspect. This maneuver is only effective for detecting small to moderate effusions (<100 mL). Inflammatory disorders such asRA, gout, and Reiter's syndrome may involve the knee joint and produce significant pain, stiffness, swelling, or warmth. A popliteal or *Baker's cyst* is best palpated with the knee partially flexed and is best seen with the patient standing with knees fully extended to visualize popliteal swelling or fullness from a posterior view.

Anserine bursitis is an often missed cause of knee pain in adults. The pes anserine bursa underlies the semimembranosus tendon and may become inflamed or painful owing to trauma, overuse, or inflammation. Anserine bursitis manifests primarily as point tenderness inferior and medial to the patella and overlying the medial tibial plateau. Swelling and erythema may not be present. Other forms of bursitis may also present as knee pain. The prepatellar bursa is superficial and is located over the inferior portion of the patella. The infrapatellar bursa is deeper and lies beneath the patellar ligament before its insertion on the tibial tubercle.

Internal derangement of the knee may result from trauma or degenerative processes.

Damage to the meniscal cartilage (medial or lateral) frequently presents as chronic or intermittent knee pain. Such an injury should be suspected when there is a history of trauma or athletic activity and when the patient relates symptoms of locking, clicking, or "giving way" of the joint. Pain may be detected during direct palpation over the medial or lateral joint line. The diagnosis may also be suggested by ipsilateral joint-line pain when the knee is stressed laterally or medially. A positive McMurray test may indicate a meniscal tear. To perform this test, the knee is first flexed at 90°, and the leg is then extended while simultaneously the lower extremity is torqued medially or laterally. A painful click during inward rotation may indicate a lateral meniscus tear, and pain during outward rotation may indicate a tear in the medial meniscus. Finally, damage to the cruciate ligaments should be suspected if there is pain of acute onset, possibly with swelling, a history of trauma, or a synovial fluid aspirate that is grossly bloody. Examination of the cruciate ligaments is best accomplished by eliciting a drawer sign. With the patient recumbent, the knee should be partially flexed and the foot stabilized on the examining surface. The examiner should manually attempt to displace the tibia anteriorly or posteriorly with respect to the femur. If anterior movement is detected, then anterior cruciate ligament damage is likely. Conversely, significant posterior movement may indicate posterior cruciate damage. Contralateral comparison will assist the examiner in detecting significant anterior or posterior movement.

HIP PAIN The hip is best evaluated by observing the patient's gait and assessing range of motion. The vast majority of patients reporting "hip pain" localize their pain unilaterally to the posterior or gluteal musculature (Fig. 320-7). Such pain may or may not be associated with low back pain and tends to radiate down the posterolateral aspect of the thigh. This presentation frequently results from degenerative arthritis of the lumbosacral spine and commonly follows a dermatomal distribution with involvement of nerve roots between L5 and S1. Some individuals instead localize their "hip pain" laterally to the area overlying the trochanteric bursa. Because of the depth of this bursa, swelling and warmth are usually absent. Diagnosis of trochanteric bursitis can be confirmed by inducing point tenderness over the trochanteric bursa. Range of movement may be limited by pain. Pain in the hip joint is less common and tends to be located anteriorly, over the inguinal ligament; it may radiate medially to the groin or along the anteromedial thigh. Uncommonly, iliopsoas bursitis may mimic true hip joint pain. Diagnosis of iliopsoas bursitis may be suggested by a history of trauma or inflammatory arthritis. Pain associated with an iliopsoas bursitis is localized to the groin or anterior thigh and tends to worsen with hyperextension of the hip; many patients prefer to flex and externally rotate the hip to reduce the pain from a distended bursa.

(Bibliography omitted in Palm version)

Back to Table of Contents

321. OSTEOARTHRITIS - *Kenneth D. Brandt*

Osteoarthritis (OA), also erroneously called degenerative joint disease, represents failure of the diarthrodial (movable, synovial-lined) joint. In idiopathic (primary) OA, the most common form of the disease, no predisposing factor is apparent. Secondary OA is pathologically indistinguishable from idiopathic OA but is attributable to an underlying cause (Table 321-1).

EPIDEMIOLOGY AND RISK FACTORS

OAis the most common joint disease of humans. Among the elderly, knee OA is the leading cause of chronic disability in developed countries; some 100,000 people in the United States are unable to walk independently from bed to bathroom because of OA of the knee or hip.

Under the age of 55 years the joint distribution ofOA in men and women is similar; in older individuals, hip OA is more common in men, while OA of interphalangeal joints and the thumb base is more common in women. Similarly, radiographic evidence of knee OA and, especially *symptomatic* knee OA, is more common in women than in men (Table 321-2).

Racial differences exist in both the prevalence ofOA and the pattern of joint involvement. The Chinese in Hong Kong have a lower incidence of hip OA than whites; OA is more frequent in native Americans than in whites. Interphalangeal joint OA and, especially, hip OA are much less common in South African blacks than in whites in the same population. Whether these differences are genetic or are due to differences in joint usage related to life-style or occupation is unknown.

In some cases, the relation of heredity toOA is less ambiguous. Thus, the mother and sister of a woman with distal interphalangeal joint OA (Heberden's nodes) are, respectively, twice and thrice as likely to exhibit OA in these joints as the mother and sister of an unaffected woman. Point mutations in the cDNA coding for articular cartilage collagen have been identified in families with chondrodysplasia and polyarticular secondary OA.

Age is the most powerful risk factor forOA. In a radiographic survey of women less than 45 years old, only 2% had OA; between the ages of 45 to 64 years, however, the prevalence was 30%, and for those older than 65 years it was 68%. In males, the figures were similar but somewhat lower in the older age groups.

Major trauma and repetitive joint use are also important risk factors forOA. Anterior cruciate ligament insufficiency or meniscus damage (and meniscectomy) may lead to knee OA. Although damage to the articular cartilage may occur at the time of injury or subsequently, with use of the affected joint, even normal cartilage will degenerate if the joint is unstable. A person with a trimalleolar fracture will almost certainly develop ankle OA.

The pattern of joint involvement inOA is influenced by prior vocational or avocational overload. Thus, while ankle OA is common in ballet dancers, elbow OA in baseball pitchers, and metacarpophalangeal joint OA in prize fighters, OA is not very common at any of these sites in the general population.

Given the growing participation of the population of this country in cardiovascular fitness programs, it is important to note that there are no convincing data to support an association between specific athletic activities and arthritis if major trauma is excluded. Neither long-distance running nor jogging has been shown to causeOA. This apparent lack of association may, however, be due to the lack of good long-term studies, the difficulty of retrospective assessment of activities, and selection bias, i.e., early discontinuation of the activity by those incurring joint damage. In contrast, vocational activities, such as those performed by jackhammer operators, cotton mill and shipyard workers, and coal miners, may lead to OA in the joints exposed to repetitive occupational use. Men whose jobs required knee bending and at least medium physical demands had a higher rate of radiographic evidence of knee OA, and more severe radiographic changes, than men whose jobs required neither.

Obesity is a risk factor for kneeOA and hand OA. For those in the highest quintile for body mass index at baseline examination, the relative risk for developing knee OA in the ensuing 36 years was 1.5 for men and 2.1 for women. For *severe* knee OA, the relative risk rose to 1.9 for men and 3.2 for women, suggesting that obesity plays an even larger role in the etiology of the most serious cases of knee OA. Furthermore, obese individuals who have not yet developed OA can reduce their risk: A weight loss of only 5 kg was found to be associated with a 50% reduction in the odds of developing symptomatic knee OA.

The correlation between the pathologic severity ofOA and symptoms is poor. Many individuals with radiographic changes of advanced OA have no symptoms. The risk factors for *pain* and *disability* in affected individuals are poorly understood. Disability in those with knee OA is more strongly associated with quadriceps muscle weakness than with either joint pain or radiographic severity of the disease. For the same degree of pathologic severity, women are more likely to be symptomatic than men, those on welfare more likely than those who are working, and those who are divorced more likely than those who are married. For individuals with OA who had poor social support, periodic telephone calls from a trained lay interviewer were as effective as a nonsteroidal anti-inflammatory drug (NSAID) in reducing joint pain, emphasizing the importance of psychosocial factors as determinants of pain.

PATHOLOGY

Although the cardinal pathologic feature ofOA is a progressive loss of articular cartilage, OA is not a disease of any single tissue but a disease of an *organ*, the synovial joint, in which all of the tissues are affected: the subchondral bone, synovium, meniscus, ligaments, and supporting neuromuscular apparatus as well as the cartilage.

The most striking morphologic changes inOA are usually seen in load-bearing areas of the articular cartilage. In the early stages the cartilage is thicker than normal, but with progression of OA the joint surface thins, the cartilage softens, the integrity of the surface is breached, and vertical clefts develop (fibrillation) (Fig. 321-1). Deep cartilage ulcers, extending to bone, may appear. Areas of fibrocartilaginous repair may develop, but the repair tissue is inferior to pristine hyaline articular cartilage in its ability to withstand mechanical stress. All of the cartilage is metabolically active, and the chondrocytes replicate, forming clusters (clones). Later, however, the cartilage becomes hypocellular.

Remodeling and hypertrophy of bone are also major features ofOA. Appositional bone growth occurs in the subchondral region, leading to the bony "sclerosis" seen radiographically. The abraded bone under a cartilage ulcer may take on the appearance of ivory (eburnation). Growth of cartilage and bone at the joint margins leads to osteophytes (spurs), which alter the contour of the joint and may restrict movement. A patchy chronic synovitis and thickening of the joint capsule may further restrict movement. Periarticular muscle wasting is common and may play a major role in symptoms and, as indicated above, in disability.

PATHOGENESIS

The main load on articular cartilage -- the major target tissue inOA -- is produced by contraction of the muscles that stabilize or move the joint. Although cartilage is an excellent shock absorber in terms of its bulk properties, at most sites it is only 1 to 2 mm thick -- too thin to serve as the sole shock-absorbing structure in the joint. Additional protective mechanisms are provided by subchondral bone and periarticular muscles.

Articular cartilage serves two essential functions within the joint, both of which are mechanical. First, it provides a remarkably smooth bearing surface, so that, with joint movement, the bones glide effortlessly over each other. With synovial fluid as lubricant, the coefficient of friction for cartilage rubbed against cartilage, even under physiologic loading, is 15 times lower than that of two ice cubes passed across each other! Second, articular cartilage prevents the concentration of stresses, so the bones do not shatter when the joint is loaded.

OAdevelops in either of two settings: (1) the biomaterial properties of the articular cartilage and subchondral bone are normal, but excessive loading of the joint causes the tissues to fail, or (2) the applied load is reasonable, but the material properties of the cartilage or bone are inferior.

Although articular cartilage is highly resistant to wear under conditions of repeated oscillation, repetitive impact loading soon leads to joint failure. This fact accounts for the high prevalence ofOA at specific sites related to vocational or avocational overloading. In general, the earliest changes occur at the sites in the joint that are subject to the greatest compressive loads. Some cases of "idiopathic" OA of the hip may be due to subtle congenital or developmental defects, such as congenital subluxation/dislocation, acetabular dysplasia, Legg-Calve-Perthes disease, or slipped capital femoral epiphysis, which increase joint congruity and concentrate the dynamic load.

Clinical conditions that reduce the ability of the cartilage or subchondral bone to deform are associated with development ofOA. In ochronosis, for example, accumulation of homogentisic acid polymers leads to stiffening of the cartilage; in osteopetrosis, stiffness of the subchondral trabeculae occurs. In both conditions, severe generalized OA is usually apparent by the age of 40. If the subchondral bone is stiffened experimentally, repetitive impact loading soon leads to breakdown of the overlying cartilage. Conversely, osteoporosis, in which the bone is abnormally soft, may protect against OA.

The Extracellular Matrix of Normal Articular Cartilage Articular cartilage is composed of two major macromolecular species: proteoglycans (PGs), which are responsible for the compressive stiffness of the tissue and its ability to withstand load, and collagen, which provides tensile strength and resistance to shear. Although lysomal proteases (cathepsins) have been demonstrated within the cells and matrix of normal articular cartilage, their low pH optimum makes it likely that the proteoglycanase activity of these enzymes will be confined to intracellular sites or the immediate pericellular area. However, cartilage also contains a family of matrix metalloproteinases (MMPs), including stromelysin, collagenase, and gelatinase, which can degrade all the components of the extracellular matrix at neutral pH. Each is secreted by the chondrocyte as a latent proenzyme that must be activated by proteolytic cleavage of its N-terminal sequence. The level of MMP activity in the cartilage at any given time represents the balance between activation of the proenzyme and inhibition of the active enzyme by tissue inhibitors. It has recently become apparent that much of the total tissue pool of aggrecan, the majorPG in articular cartilage, is degraded by a proteinase that cleaves the protein core of the molecule at a site distinct from that at which the MMPs are active. The enzyme responsible for this cleavage is referred to as "aggrecanase" but has not been clearly identified.

The turnover of normal cartilage is effected through a degradative cascade, for which many investigators consider the driving force to be interleukin (IL) 1, a cytokine produced by mononuclear cells (including synovial lining cells) and synthesized by chondrocytes. IL-1 stimulates the synthesis and secretion of the latentMMPs and of tissue plasminogen activator. Plasminogen, the substrate for the latter enzyme, may be synthesized by the chondrocyte or may enter the cartilage from the synovial fluid. Both plasminogen and stromelysin may play a role in activation of the latent MMPs. In addition to its catabolic effect on cartilage, IL-1, at concentrations even lower than those needed to stimulate cartilage degradation, suppressesPGsynthesis by the chondrocyte, inhibiting matrix repair (see below).

The balance of the system lies with at least two inhibitors, tissue inhibitor of metalloproteinase (TIMP) and plasminogen activator inhibitor-1 (PAI-1), which are synthesized by the chondrocyte and limit the degradative activity ofMMPs and plasminogen activator, respectively. If TIMP or PAI-1 is destroyed or is present in concentrations that are insufficient relative to those of active enzymes, stromelysin and plasmin are free to act on matrix substrates. Stromelysin can degrade the protein core of thePG and activate latent collagenase. Conversion of latent stromelysin to an active, highly destructive protease by plasmin provides a second mechanism for matrix degradation.

Polypeptide mediators, e.g., insulin-like growth factor-1 (IGF-1) and transforming growth factorb (TGF-b), stimulate biosynthesis ofPGs. They regulate matrix metabolism in normal cartilage and may play a role in matrix repair inOA. Notably, these growth factors modulate catabolic as well as anabolic pathways of chondrocyte metabolism; by down-regulating chondrocyte receptors forIL-1, they may decrease PG degradation.

In addition to its responsiveness to cytokines and a variety of other biologic mediators, chondrocyte metabolism in normal cartilage can be modulated directly by mechanical loading. Whereas static loading and prolonged cyclic loading inhibit synthesis ofPGs and protein, loads of relatively brief duration may stimulate matrix biosynthesis.

Pathophysiology of Cartilage Changes in OA Most investigators feel that the primary changes inOA begin in the cartilage. A change in the arrangement and size of the collagen fibers is apparent. Biochemical data are consistent with the presence of a defect in the collagen network of the matrix, perhaps due to disruption of the "glue" that binds adjacent fibers. This is among the earliest matrix changes observed and appears to be irreversible.

Although "wear" may be a factor in the loss of cartilage, strong evidence supports the concept that lysosomal enzymes andMMPsaccount for much of the loss of cartilage matrix inOA. Whether their synthesis and secretion are stimulated byIL-1 or by other factors (e.g., mechanical stimuli), MMPs, plasmin, and cathepsins all appear to be involved in the breakdown of articular cartilage in OA.TIMP andPAI-1may work to stabilize the system, at least temporarily, while growth factors, such asIGF-1,TGF-b, and basic fibroblast growth factor, are implicated in repair processes that may heal the lesion or, at least, stabilize the process. A stoichiometric imbalance exists between the levels of active enzyme and the level of TIMP, which may be only modestly increased.

Of current interest is the possible role of nitric oxide (NO) in articular cartilage damage inOA, since NO has been shown to stimulate synthesis ofMMPs by chrondrocytes. Chondrocytes are a major source of NO, the synthesis of which is stimulated byIL-1 and tumor necrosis factor and by shear stresses on the tissue. In an experimental model of OA, treatment with a selective inhibitor of inducible NO synthase reduced the severity of cartilage damage.

The chondrocytes inOAcartilage undergo active cell division and are very active metabolically, producing increased quantities of DNA, RNA collagen,PG, and noncollagenous proteins. (For this reason, it is inaccurate to call OA a *degenerative* joint disease). Prior to cartilage loss and PG depletion, this marked biosynthetic activity may lead to an increase in PG concentration, which may be associated with thickening of the cartilage and a stage of homeostasis referred to as "compensated" OA. These mechanisms may maintain the joint in a reasonably functional state for years. The repair tissue, however, often does not hold up as well under mechanical stresses as normal hyaline cartilage and eventually, at least in some cases, the rate of PG synthesis falls off and "end-stage" OA develops, with full-thickness loss of cartilage.

CLINICAL FEATURES

The joint pain ofOA is often described as a deep ache and is localized to the involved joint. Typically, the pain of OA is aggravated by joint use and relieved by rest, but, as the disease progresses, it may become persistent. Nocturnal pain, interfering with sleep, is seen particularly in advanced OA of the hip and may be enervating. Stiffness of the involved joint upon arising in the morning or after a period of inactivity (e.g., an automobile ride) may be prominent but usually lasts less than 20 min. Systemic manifestations are not a feature of primary OA.

Because articular cartilage is aneural, the joint pain inOA must arise from other structures (Table 321-3). In some cases it may be due to stretching of nerve endings in the periosteum covering osteophytes; in others, to microfractures in subchondral bone or from medullary hypertension caused by distortion of blood flow by thickened subchondral trabeculae. Joint instability, leading to stretching of the joint capsule, and muscle spasm may also be sources of pain.

In some patients withOA, joint pain may be due to synovitis. In advanced OA, histologic evidence of synovial inflammation may be as marked as that in the synovium of a patient with rheumatoid arthritis. Synovitis in OA may be due to phagocytosis of shards of cartilage and bone from the abraded joint surface (wear particles), to release from the cartilage of soluble matrix macromolecules, or to crystals of calcium pyrophosphate or hydroxyapatite. In other cases, immune complexes, containing antigens derived from cartilage matrix, may be sequestered in collagenous tissue of the joint, leading to low-grade chronic synovitis. In contrast, in the earlier stages of OA, even in the patient with chronic joint pain, synovial inflammation may be absent, suggesting that the joint pain is due to one of the other factors mentioned above.

Physical examination of theOA joint may reveal localized tenderness and bony or soft tissue swelling. Bony crepitus (the sensation of bone rubbing against bone, evoked by joint movement) is characteristic. Synovial effusions, if present, are usually not large. Palpation may reveal some warmth over the joint. Periarticular muscle atrophy may be due to disuse or to reflex inhibition of muscle contraction. In the advanced stages of OA, there may be gross deformity, bony hypertrophy, subluxation, and marked loss of joint motion. The notion that OA is inexorably progressive, however, is incorrect. In many patients the disease stabilizes; in some, regression of joint pain and even of radiographic changes occurs.

Although the diagnosis ofOA is often straightforward because of the high prevalence of radiographic changes of OA in asymptomatic individuals, it is important to ensure that joint pain in a patient with radiographic evidence of OA is not due to some other cause, such as soft tissue rheumatism (e.g., anserine bursitis at the knee, trochanteric bursitis at the hip), radiculopathy, referral of pain from another joint (e.g., 25% of patients with hip disease have pain referred to the knee), entrapment neuropathy, vascular disease (claudication), or some other type of arthritis (e.g., crystal-induced synovitis, septic arthritis). These are all common pitfalls in the diagnosis of OA. It is usually not difficult to differentiate OA from a systemic rheumatic disease, such as rheumatoid arthritis, because, in the latter diseases, joint involvement is usually symmetric and polyarticular, with arthritis in wrists and metacarpophalangeal joints (which are generally not involved in OA), and there are also constitutional features such as prolonged morning stiffness, fatigue, weight loss, or fever.

LABORATORY AND RADIOGRAPHIC FINDINGS

The diagnosis ofOA is usually based on clinical and radiographic features. In the early stages, the radiograph may be normal, but joint space narrowing becomes evident as articular cartilage is lost. Other characteristic radiographic findings include subchondral bone sclerosis, subchondral cysts, and osteophytosis. A change in the contour of the joint, due to bony remodeling, and subluxation may be seen. Although tibiofemoral joint space narrowing has been considered to be a radiographic surrogate for articular cartilage thinning, in patients with early OA who do not have radiographic evidence of bony changes (e.g., subchondral sclerosis or cysts, osteophytes), joint space narrowing alone does not accurately indicate the status of the articular cartilage. Similarly, osteophytosis alone, in the absence of other radiographic features of OA, may be due to aging rather than to OA.

As indicated above, there is often great disparity between the severity of radiographic findings, the severity of symptoms, and functional ability inOA. Thus, while more than 90% of persons over the age of 40 have some radiographic changes of OA in weight-bearing joints, only 30% of these persons are symptomatic.

No laboratory studies are diagnostic forOA, but specific laboratory testing may help in identifying one of the underlying causes of secondary OA (Table 321-1). Because primary OA is not systemic, the erythrocyte sedimentation rate, serum chemistry determinations, blood counts, and urinalysis are normal. Analysis of synovial fluid reveals mild leukocytosis (<2000 white blood cells per microliter), with a predominance of mononuclear cells. Synovial fluid analysis is of particular value in excluding other conditions, such as calcium pyrophosphate dihydrate deposition disease (Chap. 322), gout (Chap. 322), or septic arthritis (Chap. 323).

Prior to the appearance of radiographic changes, the ability to diagnoseOAclinically without an invasive procedure (e.g., arthroscopy) is limited. Approaches such as magnetic resonance imaging (MRI) and ultrasonography have not been sufficiently validated to justify their routine clinical use for diagnosis of OA or monitoring of disease progression.

OA AT SPECIFIC JOINT SITES

Interphalangeal Joints Heberden's nodes, bony enlargements of the distal interphalangeal joints, are the most common form of idiopathicOA(Fig. 321-2). A similar process at the proximal interphalangeal joints leads to Bouchard's nodes. Often, these nodes develop gradually, with little or no discomfort. However, they may present acutely with pain, redness, and swelling, sometimes triggered by minor trauma. Gelatinous dorsal cysts filled with hyaluronic acid may develop at the insertion of the digital extensor tendon into the base of the distal phalanx.

Erosive OA In erosiveOAdistal and/or proximal interphalangeal joints of the hands are most prominently affected. Erosive OA is more destructive than typical nodal OA; x-ray evidence of collapse of the subchondral plate is characteristic, and bony ankylosis may occur. Joint deformity and functional impairment may be severe. Pain and tenderness are commonly episodic. The synovium is much more extensively infiltrated with mononuclear cells than in other forms of OA.

Generalized OA GeneralizedOA is characterized by involvement of three or more joints or groups of joints (distal interphalangeal and proximal interphalangeal joints are counted as one group each). Heberden's and Bouchard's nodes are prominent. Symptoms may be episodic, with "flare-ups" of inflammation marked by soft tissue swelling, redness, and warmth. The erythrocyte sedimentation rate may be elevated, but serum rheumatoid factor tests are negative.

Thumb Base The second most frequent area of involvement inOA is the thumb base. Swelling, tenderness, and crepitus on movement of the joint are typical. Osteophytes may lead to a "squared" appearance of the thumb base (Fig. 321-3). In contrast to Heberden's nodes, which usually do not interfere significantly with function, thumb base OA frequently causes loss of motion and strength. Pain with pinch leads to adduction of the thumb and contracture of the first web space, often resulting in compensatory hyperextension of the first metacarpophalangeal joint and swan-neck deformity of the thumb.

The Hip Congenital or developmental defects (e.g., acetabular dysplasia, Legg-Calve-Perthes disease, slipped capital epiphysis) can lead to cases of hipOA. Some 20% of patients will develop bilateral involvement. Pain from hip OA is generally referred to the inguinal area but may be referred to the buttock or proximal thigh. Less commonly, hip OA presents as knee pain. Pain can be evoked by putting the involved hip through its range of motion. Flexion may be painless initially, but internal rotation will exacerbate pain. Loss of internal rotation occurs early, followed by loss of extension, adduction, and flexion due to capsular fibrosis and/or buttressing osteophytes.

The Knee OA of the knee may involve the medial or lateral femorotibial compartment and/or the patellofemoral compartment. Palpation may reveal bony hypertrophy (osteophytes) and tenderness. Effusions, if present, are generally small. Joint movement commonly elicits bony crepitus. OA in the medial compartment may result in a varus (bow-leg) deformity; in the lateral compartment it may produce a valgus (knock-knee) deformity. A positive "shrug" sign (pain when the patella is compressed manually against the femur during quadriceps contraction) may be a sign of patellofemoral OA.

Chondromalacia patellae, which also is characterized by anterior knee pain and a positive shrug sign, is a syndrome of patellofemoral pain, often bilateral, in teenagers and young adults. It is more common in females than in males. It may be caused by a variety of factors (e.g., abnormal quadriceps angle, patella alta, trauma). Although exploration of the knee may reveal softening and fibrillation of cartilage on the posterior aspect of the patella, this change is usually not progressive; chondromalacia patellae is usually not a precursor ofOA. In most cases, analgesics orNSAIDsand physical therapy are effective; in some, pain may be relieved by surgical correction of patellar malalignment.

The Spine Degenerative disease of the spine can involve the apophyseal joint, intervertebral disks, and paraspinous ligaments. *Spondylosis* refers to degenerative *disk* disease. The diagnosis of spinalOAshould be reserved for patients with involvement of the apophyseal joints and not only disk degeneration. Symptoms of spinal OA include localized pain and stiffness. Nerve root compression by an osteophyte blocking a neural foramen, prolapse of a degenerated disk, or subluxation of an apophyseal joint may cause radicular pain and motor weakness.

Marked calcification and ossification of paraspinous ligaments occur in *diffuse idiopathic* *skeletal hyperostosis* (DISH). Although DISH is often categorized as a variant ofOA, diarthrodial joints are not involved. Ligamentous calcification and ossification in the anterior spinal ligaments give the appearance of "flowing wax" on the anterior vertebral bodies. However, a radiolucency may be seen between the newly deposited bone and the vertebral body, differentiating DISH from the marginal osteophytes in spondylosis. Intervertebral disk spaces are preserved, and sacroiliac and apophyseal joints appear normal, helping to differentiate DISH from spondylosis and from ankylosing spondylitis, respectively. DISH occurs in the middle-aged and elderly and is more common in men than in women. Patients are frequently asymptomatic but may have musculoskeletal stiffness. The radiographic changes are generally much more severe than might be predicted from the mild symptoms.

TREATMENT

Treatment of OA is aimed at reducing pain, maintaining mobility, and minimizing disability. The vigor of the therapeutic intervention should be dictated by the severity of the condition in the individual patient. For those with only mild disease, reassurance, instruction in joint protection, and an occasional analgesic may be all that is required; for those with more severe OA, especially of the knee or hip, a comprehensive program comprising a spectrum of nonpharmacologic measures supplemented by an analgesic and/orNSAID is appropriate.

Nonpharmacologic Measures

Reduction of Joint Loading OA may be caused or aggravated by poor body mechanics. Correction of poor posture and a support for excessive lumbar lordosis can be helpful. Excessive loading of the involved joint should be avoided. Patients with OA of the knee or hip should avoid prolonged standing, kneeling, and squatting. Obese patients should be counseled to lose weight. In patients with medial-compartment knee OA, a wedged insole may decrease joint pain.

Rest periods during the day may be of benefit, but complete immobilization of the painful joint is rarely indicated. In patients with unilateralOA of the hip or knee, a cane, held in the contralateral hand, may reduce joint pain by reducing the joint contact force. Bilateral disease may necessitate use of crutches or a walker.

Physical Therapy Application of heat to theOA joint may reduce pain and stiffness. A variety of modalities are available; often, the least expensive and most convenient is a hot shower or bath. Occasionally, better analgesia may be obtained with ice than with heat.

It is important to note that patients with OAof weight-bearing joints are less active and tend to be less fit with regard to musculoskeletal and cardiovascular status than normal controls. An exercise program should be designed to maintain range of motion, strengthen periarticular muscles, and improve physical fitness. The benefits of aerobic exercise include increases in aerobic capacity, muscle strength, and endurance; less exertion with a given workload; and weight loss. Those who exercise regularly live longer and are healthier than those who are sedentary. Patients with hip or knee OA can participate safely in conditioning exercises to improve fitness and health without increasing their joint pain or need for analgesics orNSAIDs.

Disuse of theOA joint because of pain will lead to muscle atrophy. Because periarticular muscles play a major role in protecting the articular cartilage from stress, strengthening exercises are important. In individuals with knee OA, strengthening of the periarticular muscles may result, within weeks, in a decrease in joint pain as great as that seen withNSAIDs.

Drug Therapy of OA Therapy forOA today is palliative; no pharmacologic agent has been shown to prevent, delay the progression of, or reverse the pathologic changes of OA in humans. Although claims have been made that someNSAIDshave a "chondroprotective effect," adequately controlled clinical trials in humans with OA to support this view are lacking. In management of OA pain, pharmacologic agents should be used as adjuncts to nonpharmacologic measures, such as those described above, which are the keystone of OA treatment.

AlthoughNSAIDsoften decrease joint pain and improve mobility inOA, the magnitude of this improvement is generally modest -- on average, about 30% reduction in pain and 15% improvement in function. In a double-blinded, controlled trial in patients with symptomatic knee OA, an anti-inflammatory dose of ibuprofen (2400 mg/d) was no more effective than a low (i.e., essentially analgesic) dose of ibuprofen (1200 mg/d) or than acetaminophen (4000 mg/d), a drug with essentially no anti-inflammatory effect. Other studies confirm that an analgesic dose of ibuprofen may be as effective as anti-inflammatory doses of other NSAIDs, including the potent agent, phenylbutazone (400 mg/d), in symptomatic treatment of OA. Even in the presence of clinical signs of inflammation (e.g., synovial effusion, tenderness), relief of joint pain by acetaminophen may be as effective as that achieved with an NSAID. Nonetheless, if simple analgesics are inadequate, it is reasonable to cautiously prescribe an NSAID for a patient with OA.

It should be recognized that concern over the use ofNSAIDsin OAhas grown in recent years because of side effects of these agents, especially those related to the gastrointestinal (GI) tract. Those at greatest risk for OA, i.e., the elderly, appear also to be at greater risk than younger individuals for GI symptoms, ulceration, hemorrhage, and death as a result of NSAID use. The annual rate of hospitalization for peptic ulcer disease among elderly current NSAID users was 16 per 1000 -- four times greater than that for persons not taking an NSAID. Among those age 65 and older, as many as 30% of all hospitalizations and deaths related to peptic ulcer disease have been attributed to NSAID use. In addition to age, risk factors for hemorrhage and other ulcer complications associated with NSAID use include a history of peptic ulcer disease or of upper GI bleeding, concomitant use of glucocorticoids or anticoagulants, and, possibly, smoking and alcohol consumption (Table 321-4).

In patients who carry risk factors for anNSAID-associatedGIcatastrophe, a cyclooxygenase (Cox)-2-specific NSAID may be preferable to even a low dose of a nonselective Cox inhibitor. In contrast to the NSAIDs available to date -- all of which inhibit Cox-1 as well as Cox-2 -- two Cox-2-specific inhibitors (CSIs), celecoxib and rofecoxib, are now available. Both appear to be comparable in efficacy to the nonselective NSAIDs. Endoscopic studies have shown that both agents are associated with an incidence of gastroduodenal ulcer lower than that of comparator NSAIDs and comparable to that of placebo. Of additional advantage with respect to the issue of upper GI bleeding, CSIs do not have a clinically significant effect on platelet aggregation or bleeding time, suggesting that CSIs may be especially advantageous in patients at high risk for incurring an NSAID-associated GI catastrophe. Long-term studies are now in progress that are designed to ascertain whether clinically important differences exist between CSIs and nonselective NSAIDs with respect to major GI clinical outcomes.

Systemic glucocorticoids have no place in the treatment ofOA. However, intra- or periarticular injection of a depot glucocorticoid preparation may provide marked symptomatic relief for weeks to months. Because studies in animal models have suggested that glucocorticoids may produce cartilage damage, and frequent injections of large amounts of steroids have been associated with joint breakdown in humans, the injection should generally not be repeated in a given joint more often than every 4 to 6 months.

Intraarticular injection of hyaluronic acid has been approved recently for treatment of patients with kneeOA who have failed a program of nonpharmacologic therapy and simple analgesics. Because the duration of benefit following treatment may exceed by months the synovial half-life of exogenous hyaluronic acid, the mechanism of action is unclear. The placebo response to intraarticular injection of hyaluronic acid is often large and sustained. Although relief of knee pain is achieved more slowly after hyaluronic acid injection than after intraarticular glucocorticoid injection, the effect may last much longer after hyaluronic acid injection than after glucocorticoid injection.

Capsaicin cream, which depletes local sensory nerve endings of substance P, a neuropeptide mediator of pain, may reduce joint pain and tenderness when applied topically by patients with hand or kneeOA, even when used as monotherapy, i.e., withoutNSAIDsor systemic analgesics.

A Rational Approach to the Nonsurgical Management of OA Nonpharmacologic management is the foundation of treatment ofOA pain and is as important as -- and often more important than -- drug treatment, which should play an adjunctive or complementary role in the management of this disease. Nonpharmacologic measures may comprise instruction of the patient in principles of joint protection; thermal modalities; exercises to strengthen periarticular muscles; weight reduction, if the patient is obese; avoidance of excessive loading of the arthritic hip or knee joint by use of shoes with well-cushioned soles and a cane or walker, when appropriate; and prescription of orthotics for the patient with varus or valgus knee deformity. Medial taping of the patella may reduce knee pain in patients with patellofemoral OA. In patients with painful knee OA, if the above measures are ineffective, tidal irrigation of the joint with a large quantity of saline or Ringer's lactate warrants consideration (see below). A health education program designed to assist the patient with self-management can reduce pain and decrease health care costs; the benefits may persist for years. At any point in the course of OA, if acute joint pain and effusion develop, intraarticular injection of glucocorticoids may be indicated once joint infection is excluded by synovial fluid analysis.

Figure 321-4 provides an algorithm that might be applied to treatment of a newly diagnosed patient with kneeOA. The progressive levels of treatment are associated with increasing cost, decreasing convenience for the patient, and increasing risk of side effects. The scheme should not be interpreted dogmatically as a fixed progression of steps; rather, treatment of OA must be individualized. The treatment program should be flexible. For example, in some patients it may be reasonable to institute patellar taping or prescribe a wedged insole on the initial visit, or an intraarticular glucocorticoid injection on a later visit. As indicated above, maintaining regular contact with the patient, e.g., via periodic telephone calls, may reduce joint pain to a level beyond what can be achieved with anNSAIDalone, and this, or some surrogate measure, warrants incorporation into the treatment program (Fig. 321-4).

Because of its low cost, excellent safety profile, and an efficacy in many patients comparable to that ofNSAIDs, when an analgesic is required for treatment ofOA pain it is reasonable to prescribe acetaminophen initially, in a dose up to 4000 mg/d. If this does not control joint symptoms within a reasonable period of time, a *low dose* of NSAID (e.g., ibuprofen, 1200 mg/d; naproxen, 500 mg/d) may be substituted for, or added to, the acetaminophen. If a nonselective NSAID is used, even in a low dose, it is reasonable to recommend coadministration of a gastroprotective agent, such as misoprostol, or a proton pump inhibitor, such as famotidine or omeprazole, which have been shown by endoscopy to be effective in treating and preventing NSAID gastropathy. Because the risk of an NSAID-associatedGIcatastrophe is dose-dependent, the lowest effective dose of NSAID should be employed. Salsalate and other nonacetylated salicylates, which have only a minimal effect on prostaglandin synthase, are as effective as other NSAIDs and have a lower rate of serious GI side effects. However, phototoxicity and central nervous system toxicity may limit their use.

If the above approach does not provide adequate symptomatic relief, tramadol, a weak opioid, for which the risks of tolerance and addiction appear to be minimal, may be prescribed. Mean daily doses have typically been in the range of 200 to 300 mg. Side effects (e.g., nausea and vomiting, constipation, and drowsiness) are common, but their frequency may be reduced by initiating treatment with a dose of only 25 mg/d, which is then gradually increased over the next several days. If this is not effective or opioids are contraindicated, an anti-inflammatory dose of aCSI or of a nonselective NSAID may be prescribed, with coadministration of a gastroprotective agent in the latter instance.

WhenNSAIDsare required, they may be prescribed on an "as needed" basis, rather than in a fixed daily dose; pain control has been shown to be comparable and the risk of toxicity will be reduced. Once treatment with an NSAID or simple analgesic is initiated, the need for continuation of that treatment requires ongoing assessment. For many patients withOA, it will be possible eventually to reduce the dose of drug or to use the agent only intermittently, during exacerbations of joint pain.

Tidal Irrigation Copious irrigation of theOA knee to flush out fibrin, cartilage shards, and other debris may provide months of comfort for the patient whose joint pain has been refractory to analgesics,NSAIDs, and intraarticular glucocorticoid injections. It should be recognized, however, that invasive procedures such as this are accompanied by a large placebo effect, and studies that include a sham lavage control group have not yet been reported.

Orthopedic Surgery Joint replacement surgery should be reserved for patients with advancedOA in whom aggressive medical management has failed. In such cases total joint arthroplasty may be remarkably effective in relieving pain and increasing mobility. Osteotomy, which is surgically more conservative, can eliminate concentrations of peak dynamic loading and may provide effective pain relief in patients with hip or knee OA. It is of greatest benefit when the disease is only moderately advanced. Arthroscopic removal of loose cartilage fragments can prevent locking and relieve pain. Chondroplasty (abrasion arthroplasty) has also had some popularity as treatment for OA, but well-controlled studies of its efficacy are lacking, and the fibrocartilage that resurfaces the abraded bone is inferior to normal hyaline cartilage in its ability to withstand mechanical loads. In patients who had undergone tibial osteotomy for medial compartment knee OA, knee pain and function were not related to the extent of cartilage regeneration 2 years later.

Autologous chondrocyte transplantation and attempts at cartilage repair using mesenchymal stem cells and autologous osteochondral plugs are currently being used experimentally for repair of focal chondral defects, but have not proved to be effective in treatment ofOA.

ACKNOWLEDGEMENT

Kathie Lane provided exemplary secretarial assistance during the preparation of this manuscript.

(Bibliography omitted in Palm version)

Back to Table of Contents

322. GOUT AND OTHER CRYSTAL ARTHROPATHIES - *Antonio J. Reginato*

"GOUT" CRYSTALLOGRAPHY AND ARTHRITIS

The use of polarizing microscopy during synovial fluid analysis and the application of other crystallographic techniques, such as electron microscopy, energy-dispersive elemental analysis, and x-ray diffraction, have established the role of different microcrystals, including monosodium urate (MSU), calcium pyrophosphate dihydrate (CPPD), calcium hydroxyapatite (HA), and calcium oxalate (CaOx), in inducing acute or chronic arthritis or periarthritis. In spite of differences in crystal morphology, chemistry, and physical properties, the clinical events that result from deposition of MSU, CPPD, HA, and CaOx may be indistinguishable (Table 322-1). Prior to the use of crystallographic techniques in rheumatology, much of what was considered to be MSU gouty arthritis in fact was not. Simkin has suggested that the generic term *gout* be used to describe the whole group of crystal-induced arthritides (MSU gout, CPPD gout, HA gout, and CaOx gout). This concept further emphasizes the identical clinical presentations of these entities (Table 322-1) and the need to perform synovial fluid analysis to distinguish the type of crystal involved. In the setting of acute articular or periarticular inflammation, aspiration and analysis of effusions are most important to assess the possibility of infection and to identify the type of crystals present. Polarization microscopy alone can identify most typical crystals and allow diagnosis. HA, however, is an exception. Apart from the identification of specific microcrystalline materials or organisms, synovial fluid characteristics are nonspecific, and synovial fluid can be inflammatory or noninflammatory.

MONOSODIUMURATE GOUT

MSUgout is a metabolic disease most often affecting middle-aged to elderly men. It is typically associated with an increased uric acid pool, hyperuricemia, episodic acute and chronic arthritis, and deposition of MSU crystals in connective tissue tophi and kidneys (Chap. 347).

Acute and Chronic Arthritis Acute arthritis is the most frequent early clinical manifestation ofMSUgout. Usually, only one joint is affected initially, but polyarticular acute gout is also seen in male hypertensive patients with ethanol abuse as well as in postmenopausal women. The metatarsophalangeal joint of the first toe is often involved, but tarsal joints, ankles, and knees are also commonly affected. In elderly patients, finger joints may be inflamed. Inflamed Heberden's or Bouchard's nodes may be a first manifestation of gouty arthritis. The first episode of acute gouty arthritis frequently begins at night with dramatic joint pain and swelling. Joints rapidly become warm, red, and tender, and the clinical appearance often mimics a cellulitis. Early attacks tend to subside spontaneously within 3 to 10 days, and most of the patients do not have residual symptoms until the next episode. Several events may precipitate acute gouty arthritis: dietary excess, trauma, surgery, excessive ethanol ingestion, adrenocorticotropic hormone (ACTH) and glucocorticoid withdrawal, hypouricemic therapy, and serious medical illnesses such as myocardial infarction and stroke.

After many acute mono- or oligoarticular attacks, a proportion of gouty patients may present with a chronic nonsymmetric synovitis, causing potential confusion with rheumatoid arthritis (Chap. 312). Less commonly, chronic gouty arthritis will be the only manifestation and, more rarely, the disease will manifest as inflamed or noninflamed periarticular tophaceous deposits in the absence of chronic synovitis (Table 322-1). Women represent only 5 to 17% of all patients with gout. Premenopausal gout is a rare occurrence and accounts for only about 17% of all women with gout; it is seen mostly in individuals with a strong family history of gout. A few kindreds of precocious gout in young females caused by decreased renal urate clearance and renal insufficiency have been described. Most women with gouty arthritis are postmenopausal and elderly, have arterial hypertension causing mild renal insufficiency, and are usually receiving diuretics. Also, most of these patients have underlying degenerative joint disease, and inflamed tophaceous deposits may be seen on Heberden's and Bouchard's nodes.

Laboratory Diagnosis Even if the clinical appearance strongly suggests gout, the diagnosis should be confirmed by needle aspiration of acutely or chronically inflamed joints or tophaceous deposits. Acute septic arthritis, several of the other crystalline-associated arthropathies, palindromic rheumatism, and psoriatic arthritis may present with similar clinical features. During acute gouty attacks, strongly birefringent needle-shapedMSUcrystals with negative elongation are largely intracellular (Fig. 322-1). Synovial fluid cell counts are elevated from 2000 to 60,000/uL. Effusions appear cloudy due to leukocytes, and large amounts crystals occasionally produce a thick pasty or chalky joint fluid. Bacterial infection can coexist with urate crystals in synovial fluid; if there is any suspicion of septic arthritis, joint fluid must also be cultured. MSU crystals can often be demonstrated in the first metatarsophalangeal (MTP) joint and in knees not acutely involved with gout. Arthrocentesis of these joints is a useful technique to establish the diagnosis of gout between attacks. Serum uric acid levels can be normal or low at the time of the acute attack, since lowering of uric acid with hypouricemic therapy or other medications limits the value of serum uric acid determinations for the diagnosis of gout. Despite these limitations, serum uric acid is almost always elevated at some time and can be used to follow the course of hypouricemic therapy. A 24-h urine collection for uric acid is valuable in assessing the risk of stones, in elucidating overproduction or underexcretion of uric acid, and in deciding which hypouricemic regimen to use (Chap. 347). Excretion of more than 800 mg of uric acid per 24 h on a regular diet suggests that causes of overproduction of purine should be considered. Urinalysis, blood urea nitrogen, serum creatinine, white blood cell (WBC) count, and serum lipids should be monitored because of possible pathologic sequelae of gout and other associated diseases requiring treatment.

Radiographic Features Cystic changes, well-defined erosions described as punched-out lytic lesions with overhanging bony edges (Martel's sign), associated with soft tissue calcified masses are characteristic radiographic features of chronic tophaceous gout. However, similar radiographic signs can also be observed in erosive osteoarthritis, destructive apatite arthropathies, and rheumatoid arthritis.

TREATMENT

Acute Gouty Arthritis The mainstay of treatment during an acute attack is the administration of an anti-inflammatory drug such as colchicine, nonsteroidal anti-inflammatory drugs (NSAIDs), or glucocorticoids depending on the age of the patient and comorbid conditions. Both colchicine and NSAIDs may be quite toxic in the elderly, particularly in the presence of renal insufficiency and gastrointestinal disorders. In elderly patients, one may favor the use of intraarticular glucocorticoid injections for attacks involving one or two larger joints or cool applications along with lower oral doses of colchicine for gouty synovitis affecting small joints. Colchicine given orally is a traditional and effective treatment, if used early in the attack, in at least 85% of patients. One tablet (0.6 mg) is given every hour until relief of symptoms or gastrointestinal toxicity occurs, or a total of four to eight tablets have been taken in accordance with the age of the patient. The drug must be stopped promptly at the first sign of loose stools, and symptomatic treatment must be given for the diarrhea. Intravenous colchicine is sometimes used and can reduce, though not eliminate, the gastrointestinal side effects. Intravenous colchicine is most reliable for pre- or postoperative prophylaxis in 1- to 2-mg doses when patients cannot take medications orally. Life-threatening colchicine toxicity and sudden death have been described with the administration of more than 4 mg/d intravenously. The intravenous dose for acute gouty arthritis is 1 to 2 mg given slowly through an established venous line over 10 min in a soluset, and two additional doses of 1 mg each may be given at 6-h intervals, but the total dose should never exceed 4 mg. NSAIDs are affective in about 90% of patients, and the resolution of signs and symptoms usually occurs in 5 to 7 days. The most effective drugs are those with a short half-life and include indomethacin, 25 to 50 mg tid, ibuprofen, 800 mg tid, or diclofenac, 50 mg tid. Cyclooxigenase-2-specific inhibitors are probably equally effective but with less short-term gastrointestinal toxicity. Oral glucocorticoids such as prednisone, 30 to 50 mg/d as the initial dose and tapered over 5 to 7 days, a single intravenous dose of methylprednisolone, 7 mg of betametasone, or 60 mg of triamcinolone acetonide have been equally effective.ACTH as an intramuscular injection of 40 to 80 IU in a single dose or every 12 h for 1 to 2 days is effective in patients with acute polyarticular refractory gout or with a contraindication for using colchicine or NSAIDs.

Hypouricemic Therapy Attempts to normalize serum uric acid to <300 umol/L (5.0 mg/dL) to prevent recurrent gouty attacks and eliminate tophaceous deposits entail a commitment to long-term hypouricemic regimens and medications that generally are required for life. Hypouricemic therapy should be considered when the hyperuricemia cannot be corrected by simple means (control of body weight, low-purine diet, increase in liquid ingestion, limitation of ethanol intake, and avoidance of diuretic use). The decision to initiate hypouricemic therapy is usually made taking into consideration the number of acute attacks, family history of gout, presence ofMSUtophaceous deposits, uric acid excretion >800 mg per 24 hours, presence of uric acid stones, and risk for acute uric acid nephropathy during chemotherapy for myeloproliferative disorders. Uricosuric agents, such as probenecid, can be used in patients with good renal function who underexcrete uric acid, with <600 mg in a 24-hour urine sample. Urine volume must be maintained by ingestion of 1500 mL of water every day. Probenecid can be started at a dosage of 200 mg twice daily and increased gradually as needed up to 2 g in order to maintain a serum uric acid level <300 umol/L (5 mg/dL). Probenecid is the drug of choice to treat elderly patients with hypertension and thiazide dependence; however, probenecid is not effective with a renal creatinine clearance<1 mL/s. These patients may require allopurinol or benzbromarone (not available in the United States), which is another uricosuric drug that is effective in patients with renal failure and who are receiving diuretics. Allopurinol is the best drug to lower serum urate in overproducers, stone formers, and patients with advanced renal failure. It can be given in a single morning dose, 300 mg initially and increasing up to 800 mg if needed. In most patients, it is not necessary to start at a lower dose; however, in patients with renal failure, the dosage should be adjusted depending on the serum creatinine concentration in order to minimize side effects. Patients with frequent acute attacks may require lower initial doses to prevent exacerbations. Toxicity of allopurinol has been recognized increasingly in patients with renal failure who use thiazide diuretics and in those patients allergic to penicillin and ampicillin. The most serious side effects include skin rash with progression to life-threatening toxic epidermal necrolysis, systemic vasculitis, bone marrow suppression, granulomatous hepatitis, and renal failure. Urate-lowering drugs should not be initiated during acute attacks. This is especially important in patients who have refractory acute arthritis or who had a flare-up previously with hypouricemic drugs. Colchicine prophylaxis in doses of 0.6 mg one to two times daily is usually continued, along with hypouricemic therapy, until the patient is normouricemic and without gouty attacks for 3 months. However, prophylactic colchicine treatment may be necessary as long as tophi are present.

CPPD DEPOSITION DISEASE

Pathogenesis The deposition of CPPDcrystals in articular tissues is most common in the elderly, affecting 10 to 15% of persons 65 to 75 years old and 30 to 60% of those more than 85 years old. In most cases this process is asymptomatic, and the cause of CPPD deposition is uncertain. Because over 80% of patients are more than 60 years old and 70% have preexisting joint damage from other conditions, it is likely that biochemical changes in aging cartilage favor crystal nucleation. Examples of such chemical alterations include the following. There is an increased production of inorganic pyrophosphate and decreased levels of pyrophosphatases in cartilage extracts from patients with CPPD arthritis. The increase in pyrophosphate production appears to be related to enhanced activity of ATP pyrophosphohydrolase and 5¢-nucleotidase, which catalyze the reaction of ATP to adenosine and pyrophosphate. This pyrophosphate could combine with calcium to form CPPD crystals in matrix vesicles or on collagen fibers. There is a diminution in the levels of cartilage glycosaminoglycans that normally inhibit and regulate crystal nucleation. These deficiencies may lead to increased crystal deposition. In vitro studies have demonstrated that transforming growth factor b1 and epidermal growth factor both stimulate the production of pyrophosphate by articular cartilage and thus may contribute to the deposition of CPPD crystals. The release of CPPD crystals into the joint space is followed by the phagocytosis of these crystals by neutrophils, which respond by releasing inflammatory substances. In addition, neutrophils release a glycopeptide that is chemotactic for other neutrophils, thus augmenting the inflammatory events. The same substance is present inMSUgout.

A minority of patients withCPPDarthropathy have metabolic abnormalities or hereditary CPPD disease (Table 322-2). These associations suggest that a variety of different metabolic products may enhance CPPD deposition. Included among these conditions are the "four H's" of hyperparathyroidism, hemochromatosis, hypophosphatasia, and hypomagnesemia. Hemochromatosis and hyperparathyroidism are good examples. Ferrous ions and hypercalcemia may either directly alter cartilage or inhibit inorganic pyrophosphatases, leading to enhanced susceptibility to CPPD deposition. The presence of CPPD arthritis in individuals less than 50 years old should lead to consideration of these metabolic disorders and inherited forms of disease, including those identified in a variety of ethnic groups (Table 322-2). Genomic DNA studies performed on four different kindreds have shown a possible location of the genetic defects on chromosome 8q in one, and on chromosome 5p in the other three. Identification of these genes will help elucidate the pathogenesis of both the familial and the more common sporadic form of the disease. Investigation should include inquiry for evidence of familial aggregation and evaluation of serum calcium, phosphorus, alkaline phosphatase, magnesium, serum ferritin, and transferritin saturation.

Clinical Manifestations CPPDarthropathy may be asymptomatic, acute, subacute, or chronic or cause acute synovitis superimposed on chronically involved joints. Acute CPPD arthritis was originally termed *pseudogout* by McCarty and coworkers because of its striking similarity toMSUgout. Other clinical manifestations of CPPD deposition include (1) induction or enhancement of peculiar forms of osteoarthritis; (2) induction of severe destructive disease that may radiographically mimic neuropathic arthritis; (3) production of symmetric proliferative synovitis, clinically similar to rheumatoid arthritis and frequently seen in familial forms with early onset; (4) intervertebral disk and ligament calcification with restriction of spine mobility, mimicking ankylosing spondylitis (also seen in hereditary forms); and (5) rarely spinal stenosis (most commonly seen in the elderly (Table 322-1).

The knee is the joint most frequently affected inCPPDarthropathy. Other sites include the wrist, shoulder, ankle, elbow, and hands. Rarely, the temporomandibular joint and ligamentum flavum of the spinal canal are involved. Clinical and radiographic evidence indicates that CPPD deposition is polyarticular in at least two-thirds of patients. When the clinical picture resembles that of slowly progressive osteoarthritis, diagnosis may be more difficult. Joint distribution may provide important clues suggesting CPPD disease. For example, primary osteoarthritis rarely involves a metacarpophalangeal, wrist, elbow, shoulder, or ankle joint. If radiographs reveal punctate and/or linear radiodense deposits in fibrocartilaginous joint menisci or articular hyaline cartilage (chondrocalcinosis), the diagnostic certainty of CPPD is further enhanced. *Definitive diagnosis* requires demonstration of typical crystals in synovial fluid or articular tissue (Fig. 322-2). In the absence of joint effusion or indications to obtain a synovial biopsy, chondrocalcinosis is presumptive of CPPD deposition. One exception is chondrocalcinosis due toCaOx in some patients with chronic renal failure.

Acute attacks ofCPPDarthritis may be precipitated by trauma, arthroscopy, or hyaluronate injections. Rapid diminution of serum calcium concentration, as may occur in severe medical illness or after surgery (especially parathyroidectomy), can also lead to pseudogout attacks.

In as many as 50% of cases,CPPDgout is associated with low-grade fever and, on occasion, temperatures as high as 40°C. Whether or not radiographic proof of chondrocalcinosis is evident in the involved joint(s), synovial analysis with microbial cultures is essential to rule out the possibility of infection. In fact, infection in a joint with any microcrystalline deposition process can lead to crystal shedding and subsequent synovitis from both crystals and microorganisms. Synovial fluid in acute CPPD gout has inflammatory qualities. The WBCcount can range from several thousand cells to 100,000 cells/uL, the mean being about 24,000 cells/uL and the predominant cell being the neutrophil. Polarization microscopy usually reveals rhomboid crystals with weak positive birefringence inside fibrim and in neutrophils (Fig. 322-2).

TREATMENT

Untreated acute attacks may last a few days to as long as a month. Treatment by joint aspiration andNSAIDs, or colchicine, or intraarticular glucocorticoid injection may result in return to prior status in 10 days or less. For patients with frequent recurrent attacks of CPPDgout, daily prophylactic treatment with low doses of colchicine may be helpful in decreasing the frequency of the attacks. Severe polyarticular attacks usually require short courses of glucocorticoids. Unfortunately, there is no effective way to remove CPPD deposits from cartilage and synovium. Uncontrolled studies suggest that radioactive synovectomy (with yttrium 90) or the administration of antimalarial agents may be helpful in controlling persistent synovitis. Patients with progressive destructive large-joint arthropathy usually require joint replacement.

CALCIUM HYDROXYAPATITE DEPOSITION DISEASE

Pathogenesis HA is the primary mineral of bone and teeth. Abnormal accumulation can occur in areas of tissue damage (dystrophic calcification), in hypercalcemic or hyperparathyroid states (metastatic calcification), and in certain conditions of unknown cause (Table 322-3). In chronic renal failure, hyperphosphatemia enhances HA deposition both in and around joints.

HAmay be released from exposed bone and cause the acute synovitis occasionally seen in chronic stable osteoarthritis (e.g., "hot" Heberden's nodes). HA deposition is also an important factor in an extremely destructive chronic arthropathy of the elderly that occurs most often in knees and shoulders (Milwaukee shoulder). Joint destruction is associated with attenuation or rupture of supporting structures, leading to instability and deformity. Progression tends to be indolent, and synovial fluidWBCcounts are usually less than 1000/uL. Symptoms range from minimal to severe pain and disability that may lead to joint replacement surgery. Whether severely affected patients merely represent an extreme synovial tissue response to the HA crystals that are so common in osteoarthritis is uncertain. Synovial membrane tissue cultures exposed to HA (orCPPD) crystals markedly increased the release of collagenases and neutral proteases, underscoring the destructive potential of abnormally stimulated synovial lining cells.

Clinical Manifestations Periarticular and articular deposits may coexist and be associated with acute and/or chronic damage to the joint capsule, tendons, bursa, or articular surfaces. The most common sites ofHAdeposition include bursae and tendons in and/or around the knees, shoulders, hips, and fingers. Clinical manifestations include asymptomatic radiographic abnormalities, acute synovitis, bursitis, tendinitis, and chronic destructive arthropathy. Most patients with HA arthropathy are elderly. Although the true incidence of HA arthritis is not known, 30 to 50% of patients with osteoarthritis have HA microcrystals in their synovial fluid. Such crystals can frequently be identified in clinically stable osteoarthritic joints, but they are more likely to come to attention in persons experiencing acute or subacute worsening of joint pain and swelling. The synovial fluidWBCcount in HA arthritis is usually low (<2000/uL) but may at times have as many as 50,000/uL. Most synovial fluid analyses reveal a predominance of mononuclear cells. Occasionally, neutrophils may dominate.

Diagnosis Radiographic findings inHAarthropathy are not diagnostic. Intra- and/or periarticular calcifications with or without erosive, destructive, or hypertrophic changes may be present.

Definitive diagnosis ofHAarthropathy depends on identification of crystals from synovial fluid or tissue (Fig. 322-3). Individual crystals are very small, nonbirefringent, and can only be seen by electron microscopy. Clumps of crystals may appear as 1- to 20-um shiny intra- or extracellular globules that stain purplish with Wright's stain and bright red with alizarin red S. Absolute identification depends on electron microscopy with energy-dispersive elemental analysis, x-ray diffraction, or infrared spectroscopy.

TREATMENT

Treatment of HAarthritis is nonspecific. Acute attacks of bursitis or synovitis may be self-limiting, resolving in from days to several weeks. Aspiration of effusions and the use of eitherNSAIDsor oral colchicine for 2 weeks or intra- or periarticular injection of glucocorticoid salts appear to shorten the duration and intensity of symptoms. In patients with underlying severe destructive articular changes, response to medical therapy is usually less rewarding.

CAOXDEPOSITION DISEASE

Pathogenesis *Primary oxalosis* is a rare hereditary metabolic disorder (Chap. 352). Enhanced production of oxalic acid may result from at least two different enzyme defects, leading to hyperoxalemia and deposition of calcium oxalate crystals in tissues. Nephrocalcinosis, renal failure, and death usually occur before age 20. Acute and/or chronicCaOxarthritis and periarthritis may complicate primary oxalosis during later years of illness.

Secondary oxalosis is more common than the primary disorder. It is one of the many metabolic abnormalities that complicate end-stage renal disease (ESRD). In ESRD, calcium oxalate deposits have long been recognized in visceral organs, blood vessels, bones, and even cartilage. However, it was not until 1982 that such deposits were demonstrated to be one of the causes of arthritis in chronic renal failure. Thus far, reported patients have been dependent on long-term hemodialysis or peritoneal dialysis (Chap. 272), and many had received ascorbic acid supplements. Ascorbic acid is metabolized to oxalate, which is inadequately cleared in uremia and by dialysis. Such supplements are now usually avoided in dialysis programs because of the risk of enhancing hyperoxalosis and its sequelae.

Clinical Manifestations and Diagnosis As was noted for the other calcium salts,CaOxaggregates can be found in bone, articular cartilage, synovium, and periarticular tissues. From these sites, crystals may be shed, causing acute synovitis. Persistent aggregates of CaOx may, likeHA andCPPD, stimulate synovial proliferation and enzyme release, resulting in progressive articular destruction. Deposits have been documented in fingers, wrists, elbows, knees, ankles, and feet.

Each of the known microcrystalline arthropathies may be a complication ofESRD, and rare patients have more than one type of crystal present in a joint effusion. The advent of crystallographic techniques has made it clear that most arthritic problems in ESRD are not, as was once believed, due toMSUgout. Clinical features of acuteCaOxarthritis may not be distinguishable from those due to sodium urate,CPPD, orHA. Radiographs may reveal chondrocalcinosis, a feature of either CPPD or CaOx deposition. CaOx-induced synovial effusions are usually noninflammatory, with fewer than 2000 leukocytes/uL. Neutrophils or mononuclear cells have predominated. CaOx crystals have a variable shape and variable birefringence to polarized light. The most easily recognized forms are bipyramidal and have strong positive birefringence (Fig. 322-4).

TREATMENT

Treatment of CaOxarthropathy withNSAIDs, colchicine, intraarticular glucocorticoids, and/or an increased frequency of dialysis has produced only slight improvement. In primary oxalosis, liver transplantation has induced a significant reduction in crystal deposits (Chap. 352).

(Bibliography omitted in Palm version)

Back to Table of Contents

323. INFECTIOUS ARTHRITIS - *Scott J. Thaler*, *James H. Maguire*

INTRODUCTION AND APPROACH TO THE PATIENT

While *Staphylococcus aureus*, *Neisseria gonorrhoeae*, and other bacteria are the most common causes of infectious arthritis, various mycobacteria, spirochetes, fungi, and viruses also infect joints. Since acute bacterial infection can rapidly destroy articular cartilage, all inflamed joints must be evaluated without delay to exclude noninfectious processes and to determine appropriate antimicrobial therapy and drainage procedures. For more detailed information on infectious arthritis due to specific organisms, the reader is referred to the chapters on those organisms.

Acute bacterial infection typically involves a single joint or a few joints. Subacute or chronic monarthritis or oligoarthritis suggests mycobacterial or fungal infection; episodic inflammation is seen in syphilis, Lyme disease, and the reactive arthritis that follows enteric infections and chlamydial urethritis (Table 323-1). Acute polyarticular inflammation occurs as an immunologic reaction during the course of endocarditis, rheumatic fever, disseminated neisserial infection, and acute hepatitis B. Bacteria and viruses occasionally infect multiple joints, the former most commonly in persons with rheumatoid arthritis.

Aspiration of synovial fluid, an essential element in the evaluation of potentially infected joints, can be performed without difficulty in most cases by the insertion of a large-bore needle into the site of maximal fluctuation or tenderness or by the route of easiest access. Ultrasonography or fluoroscopy may be used to guide aspiration of difficult-to-localize effusions of the hip and, occasionally, the shoulder and other joints. Normal synovial fluid contains<180 cells (predominantly mononuclear cells) per microliter. Synovial cell counts averaging 100,000/uL (range, 25,000 to 250,000/uL), with >90% neutrophils, are characteristic of acute bacterial infections. Crystal-induced, rheumatoid, and other noninfectious inflammatory arthritides are usually associated with<30,000 to 50,000 cells/uL; cell counts of 10,000 to 30,000/uL, with 50 to 70% neutrophils and the remainder lymphocytes, are common in mycobacterial and fungal infections. Definitive diagnosis of an infectious process relies on identification of the pathogen in stained smears of synovial fluid, isolation of the pathogen from cultures of synovial fluid and blood, or detection of microbial nucleic acids and proteins by polymerase chain reaction (PCR)-based assays and immunologic techniques.

ACUTE BACTERIAL ARTHRITIS

Pathogenesis Bacteria enter the joint from the bloodstream, from a contiguous site of infection in bone or soft tissue, or by direct inoculation during surgery, injection, or trauma. In hematogenous infection, bacteria escape from synovial capillaries, which have no limiting basement membrane, and within hours provoke neutrophilic infiltration of the synovium. Neutrophils and bacteria enter the joint space; later, bacteria adhere to articular cartilage. Degradation of cartilage begins within 48 h as a result of increased intraarticular pressure, release of proteases and cytokines from chondrocytes and synovial macrophages, and invasion of the cartilage by bacteria and inflammatory cells. Histologic studies reveal bacteria lining the synovium and cartilage as well as abscesses extending into the synovium, cartilage, and -- in severe cases -- subchondral bone. Synovial proliferation results in the formation of a pannus over the cartilage, and thrombosis of inflamed synovial vessels develops. Bacterial factors that appear important in the pathogenesis of infective arthritis include various surface-associated adhesins in *S. aureus* that permit adherence to cartilage and endotoxins that promote chondrocyte-mediated breakdown of cartilage.

Microbiology The hematogenous route of infection is the most common route in all age groups. In infants, group B streptococci, gram-negative enteric bacilli, and *S. aureus* are the usual pathogens. Since the advent of the *Haemophilus influenzae* vaccine, *S. aureus*, *Streptococcus pyogenes* (group A *Streptococcus*), and (in some centers) *Kingella kingae* have predominated among children <5 years of age. Among young adults and adolescents, *N. gonorrhoeae* is the most commonly implicated organism. *S. aureus* accounts for most nongonococcal isolates in adults of all ages; gram-negative bacilli, pneumococci, andb-hemolytic streptococci -- particularly groups A and B, but also groups C, G, and F -- are involved in up to one-third of cases in older adults, especially those with underlying comorbid illnesses.

Infections following surgical procedures or penetrating injuries are due most often to *S. aureus* and occasionally to other gram-positive bacteria or gram-negative bacilli. Infections with coagulase-negative staphylococci are unusual except after the implantation of prosthetic joints or arthroscopy. Anaerobic organisms, often in association with aerobic or facultative bacteria, are found after human bites and when decubitus ulcers or intraabdominal abscesses spread into adjacent joints. Polymicrobial infections complicate traumatic injuries with extensive contamination. Cat bites or scratches may introduce *Pasteurella multocida* into joints.

Nongonococcal Bacterial Arthritis

Epidemiology Although hematogenous infections with virulent organisms such as *S. aureus*, *H. influenzae*, and pyogenic streptococci occur in healthy persons, there is an underlying host predisposition in many cases of septic arthritis. Patients with rheumatoid arthritis have the highest incidence of infective arthritis, most often secondary to *S. aureus*, because of chronically inflamed joints, glucocorticoid therapy, and frequent breakdown of rheumatoid nodules, vasculitic ulcers, and skin overlying deformed joints. Diabetes mellitus, glucocorticoid therapy, hemodialysis, and malignancy all carry an increased risk of infection with *S. aureus* and gram-negative bacilli. Pneumococcal infections complicate alcoholism, deficiencies of humoral immunity, and hemoglobinopathies. Pneumococci, *Salmonella*, and *H. influenzae* cause septic arthritis in persons infected with HIV. Persons with primary immunoglobulin deficiency are at risk for mycoplasmal arthritis, which results in permanent joint damage if treatment with tetracycline and intravenous immunoglobulin replacement is not administered promptly. Intravenous drug users acquire staphylococcal and streptococcal infections from their own flora and acquire pseudomonal and other gram-negative infections from drugs and injection paraphernalia.

Clinical Manifestations Some 90% of patients present with involvement of a single joint: most commonly the knee, less frequently the hip, and still less often the shoulder, wrist, or elbow. Small joints of the hands and feet are more likely to be affected after direct inoculation or a bite. Among intravenous drug users, infections of the spine, sacroiliac joints, or sternoclavicular joints are more common than infections of the appendicular skeleton. Polyarticular infection is most common among patients with rheumatoid arthritis and may resemble a flare of the underlying disease.

The usual presentation consists of moderate to severe pain that is uniform around the joint, effusion, muscle spasm, and decreased range of motion. Fever in the range of 38.3 to 38.9°C (101 to 102°F) and sometimes higher is common but may be lacking, especially in persons with rheumatoid arthritis, renal or hepatic insufficiency, or conditions requiring immunosuppressive therapy. The inflamed, swollen joint is usually evident on examination except in the case of a deeply situated joint, such as the hip, shoulder, or sacroiliac joint. Cellulitis, bursitis, and acute osteomyelitis, which may produce a similar clinical picture, should be distinguished from septic arthritis by their greater range of motion and less-than-circumferential swelling. A focus of extraarticular infection, such as a boil or pneumonia, should be sought. Peripheral-blood leukocytosis with a left shift and elevation of the erythrocyte sedimentation rate or C-reactive protein are common findings.

Plain radiographs show evidence of soft tissue swelling, joint-space widening, and displacement of tissue planes by the distended capsule. Narrowing of the joint space and bony erosions indicate advanced infection and a poor prognosis. Ultrasound is useful for detecting effusions in the hip, and computed tomography or magnetic resonance imaging can demonstrate infections of the sacroiliac joint, sternoclavicular joint, and the spine very well.

Laboratory Findings Specimens of peripheral blood and synovial fluid should be obtained before antibiotics are administered. Blood cultures are positive in up to 50% of *S. aureus* infections but are less frequently positive in infections due to other organisms. The synovial fluid is turbid, serosanguineous, or frankly purulent. Gram-stained smears confirm the presence of large numbers of neutrophils. Levels of total protein and lactate dehydrogenase in synovial fluid are elevated, and the glucose level is depressed; however, these findings are not specific for infection, and measurement of these levels is not necessary to make the diagnosis. The synovial fluid should be examined for crystals, because gout and pseudogout can resemble septic arthritis clinically, and infection and crystal-induced disease occasionally occur together. Organisms are seen on synovial fluid smears in nearly three-quarters of infections with *S. aureus* and streptococci and in 30 to 50% of infections due to gram-negative and other bacteria. Cultures of synovial fluid are positive in>90% of cases. Inoculation of synovial fluid into bottles containing liquid media for blood cultures increases the yield of culture, especially if the pathogen is a fastidious organism or the patient is taking an antibiotic. Although not yet widely available,PCR-based assays for bacterial DNA will also be useful for the diagnosis of partially treated or culture-negative bacterial arthritis.

TREATMENT

Prompt administration of systemic antibiotics and drainage of the involved joint can prevent destruction of cartilage, postinfectious degenerative arthritis, joint instability, or deformity. Once samples of blood and synovial fluid have been obtained for culture, empirical antibiotics should be given that are directed against bacteria visualized on smears or against the pathogens that are likely, given the patient's age and risk factors.

Initial therapy should consist of the intravenous administration of bactericidal agents; direct instillation of antibiotics into the joint is not necessary to achieve adequate levels in synovial fluid and tissue. An intravenous third-generation cephalosporin such as cefotaxime (1 g every 8 h) or ceftriaxone (1 to 2 g every 24 h) will provide adequate empirical coverage for most community-acquired infections in adults when smears show no organisms. Either oxacillin or nafcillin (2 g every 4 h) is used if there are gram-positive cocci on the smear. If methicillin-resistant *S. aureus* is a possible pathogen, as in hospitalized patients, intravenous vancomycin (1 g every 12 h) should be given. In addition, an aminoglycoside should be given to intravenous drug users or other patients in whom *Pseudomonas aeruginosa* may be the responsible agent.

Definitive therapy is based on the identity and antibiotic susceptibility of the bacteria isolated in culture. Infections due to staphylococci are treated with oxacillin, nafcillin, or vancomycin for 4 weeks. Pneumococcal and streptococcal infections due to penicillin-susceptible organisms respond to 2 weeks of therapy with penicillin G (2 million units intravenously every 4 h); infections caused by *H. influenzae* and by strains of *S. pneumoniae* that are resistant to penicillin are treated with cefotaxime or ceftriaxone for 2 weeks. Most enteric gram-negative infections can be cured in 3 to 4 weeks by a second- or third-generation cephalosporin given intravenously or by a fluoroquinolone, such as levofloxacin (500 mg intravenously or orally every 24 h). *P. aeruginosa* infection should be treated for at least 2 weeks with a combination regimen of an aminoglycoside plus either an extended-spectrum penicillin, such as mezlocillin (3 g intravenously every 4 h), or an antipseudomonal cephalosporin, such as ceftazidime (1 g intravenously every 8 h). If tolerated, this regimen is continued for an additional 2 weeks; alternatively, a fluoroquinolone, such as ciprofloxacin (750 mg orally bid), is given by itself or with the penicillin or cephalosporin in place of the aminoglycoside.

Timely drainage of pus and necrotic debris from the infected joint is required for a favorable outcome. Needle aspiration of readily accessible joints such as the knee may be adequate if loculations or particulate matter in the joint does not prevent its thorough decompression. Arthroscopic drainage and lavage may be employed initially or within several days if repeated needle aspiration fails to relieve symptoms, decrease the volume of the effusion and the synovial white cell count, and clear bacteria from smears and cultures. In some cases, arthrotomy is necessary to remove loculations and debride infected synovium, cartilage, or bone. Septic arthritis of the hip is best managed with arthrotomy, particularly in young children, in whom infection threatens the viability of the femoral head. Septic joints do not require immobilization except for pain control before symptoms are alleviated by treatment. Weight bearing should be avoided until signs of inflammation have subsided, but frequent passive motion of the joint is indicated to maintain full mobility. While addition of glucocorticoids to antibiotic treatment improves the outcome of *S. aureus* arthritis in experimental animals, no clinical trials have yet evaluated this approach in humans.

Gonococcal Arthritis

Epidemiology Gonococcal arthritis, accounting for 70% of episodes of infectious arthritis in persons<40 years of age, results from bacteremia arising from gonococcal infection or, more frequently, from asymptomatic gonococcal mucosal colonization of the urethra, cervix, or pharynx. Women are at greatest risk during menses or during pregnancy and overall are two to three times more likely than men to develop disseminated gonococcal infection and arthritis. Persons with complement deficiencies, especially of the terminal components, are prone to recurrent episodes of gonococcemia. Strains of gonococci that are most likely to cause disseminated infection include those that produce transparent colonies in culture, have the type IA outer-membrane protein, or are of the AUH-auxotroph type.

Clinical Manifestations and Laboratory Findings The most common manifestation of disseminated gonococcal infection is a syndrome of fever, chills, rash, and articular symptoms. Small numbers of papules that progress to hemorrhagic pustules develop on the trunk and the extensor surfaces of the distal extremities. Migratory arthritis and tenosynovitis of the knees, hands, wrists, feet, and ankles are prominent. The cutaneous lesions and articular findings are believed to be the consequence of an immune reaction to circulating gonococci and immune-complex deposition in tissues. Thus, cultures of synovial fluid are consistently negative, and blood cultures are positive in <45% of patients. Synovial fluid may be difficult to obtain from inflamed joints and usually contains only 10,000 to 20,000 leukocytes/uL.

True gonococcal septic arthritis is less common than the disseminated gonococcal infection syndrome and always follows disseminated infection, which is unrecognized in one-third of patients. A single joint, such as the hip, knee, ankle, or wrist, is usually involved. Synovial fluid, which contains >50,000 leukocytes/uL, can be obtained with ease; the gonococcus is only occasionally evident in gram-stained smears, and cultures of synovial fluid are positive in<40% of cases. Blood cultures are almost always negative.

Because it is difficult to isolate gonococci from synovial fluid and blood, specimens for culture should be obtained from potentially infected mucosal sites. Cultures and gram-stained smears of skin lesions occasionally are positive. All specimens for culture should be plated onto Thayer-Martin agar directly or in special transport media at the bedside and transferred promptly to the microbiology laboratory in an atmosphere of 5% CO_2, as generated in a candle jar.PCR-based assays are extremely sensitive in detecting gonococcal DNA in synovial fluid. A dramatic alleviation of symptoms within 12 to 24 h after the initiation of appropriate antibiotic therapy supports a clinical diagnosis of the disseminated gonococcal infection syndrome if cultures are negative.

TREATMENT

Initial treatment consists of ceftriaxone (1 g intravenously or intramuscularly every 24 h) to cover possible penicillin-resistant organisms. Once local and systemic signs are clearly resolving, the 7-day course of therapy can be completed with an oral agent such as cefixime (400 mg bid) or ciprofloxacin (500 mg bid) or, if penicillin-susceptible organisms are isolated, amoxicillin (500 mg tid). Suppurative arthritis usually responds to needle aspiration of involved joints and 7 to 14 days of antibiotic treatment. Arthroscopic lavage or arthrotomy is rarely required.

It is noteworthy that arthritis symptoms similar to those seen in disseminated gonococcal infections occur in meningococcemia. A dermatitis-arthritis syndrome, purulent monarthritis, and reactive polyarthritis have been described. All respond to treatment with intravenous penicillin.

SPIROCHETAL ARTHRITIS

Lyme Disease Lyme disease due to infection with the spirochete *Borrelia burgdorferi* causes arthritis in up to 70% of persons who are not treated. Intermittent arthralgias and myalgias, but not arthritis, occur within days or weeks of inoculation of the spirochete by the *Ixodes* tick. Later, there are three patterns of joint disease: (1) Fifty percent of untreated persons experience intermittent episodes of monarthritis or oligoarthritis involving the knee and/or other large joints. The symptoms wax and wane without treatment over months, and each year 10 to 20% of patients report loss of joint symptoms. (2) Twenty percent of untreated persons develop a pattern of waxing and waning arthralgias. (3) Ten percent of patients develop chronic inflammatory synovitis resulting in erosive lesions and destruction of the joint. Serologic tests for IgG antibodies to *B. burgdorferi* are positive in >90% of persons with Lyme arthritis, and aPCR-based assay detects *Borrelia* DNA in 85%.

TREATMENT

Lyme arthritis generally responds well to therapy. A regimen of oral doxycycline (100 mg bid for 30 days), oral amoxicillin (500 mg qid for 30 days), or parenteral ceftriaxone (2 g/d for 2 to 4 weeks) is recommended. Patients who do not respond to a total of 2 months of oral therapy or 1 month of parenteral therapy are unlikely to benefit from additional antibiotic therapy and are treated with anti-inflammatory agents or synovectomy. Failure of therapy is associated with host features such as the HLA-DR4 genotype, persistent reactivity to OspA (outer-surface protein A), and the presence of hLFA-1 (human leukocyte function-associated antigen-1), which cross-reacts with OspA.

Syphilitic Arthritis Articular manifestations occur in different stages of syphilis. In early congenital syphilis, periarticular swelling and immobilization of the involved limbs (Parrot's pseudoparalysis) complicate osteochondritis of long bones. Clutton's joint, a late manifestation of congenital syphilis that typically develops between the ages of 8 and 15 years, is caused by chronic painless synovitis with effusions of large joints, particularly the knees and elbows. Secondary syphilis may be associated with arthralgias; symmetric arthritis of the knees and ankles and occasionally of the shoulders and wrists; and sacroiliitis. The arthritis follows a subacute to chronic course with a mixed mononuclear and neutrophilic synovial-fluid pleocytosis (typical cell counts, 5000 to 15,000/uL). Immunologic mechanisms may contribute to the arthritis, and symptoms usually improve rapidly with penicillin therapy. In tertiary syphilis, Charcot's joint is a result of sensory loss due to tabes dorsalis. Penicillin is not helpful in this setting.

MYCOBACTERIAL ARTHRITIS

Tuberculous arthritis accounts for ~1% of all cases of tuberculosis and for 10% of extrapulmonary cases. The most common presentation is chronic granulomatous monarthritis. An unusual syndrome, Poncet's disease, is a reactive symmetric form of polyarthritis that affects persons with visceral or disseminated tuberculosis. No mycobacteria are found in the joints, and symptoms resolve with antituberculous therapy.

Unlike tuberculous osteomyelitis, which typically involves the thoracic and lumbar spine (50% of cases), tuberculous arthritis primarily involves the large weight-bearing joints, in particular the hips, knees, and ankles, and only occasionally involves smaller non-weight-bearing joints. Progressive monarticular swelling and pain develop over months to years, and systemic symptoms are seen in only half of all cases. Tuberculous arthritis occurs as part of a disseminated primary infection or through late reactivation, often in persons with HIV infection or other immunocompromised hosts. Coexistent active pulmonary tuberculosis is unusual.

Aspiration of the involved joint yields fluid with an average cell count of 20,000/uL, with ~50% neutrophils. Acid-fast staining of the fluid yields positive results in fewer than one-third of cases, and cultures are positive in 80%. Culture of synovial tissue taken at biopsy is positive in ~ 90% of cases and shows granulomatous inflammation in most. DNA amplification methods such asPCR can shorten the time to diagnosis to 1 or 2 days. Radiographs reveal peripheral erosions at the points of synovial attachment, periarticular osteopenia, and eventually joint-space narrowing. Therapy for tuberculous arthritis is the same as that for tuberculous pulmonary disease, requiring the administration of multiple agents for 6 to 9 months. Therapy is more prolonged in immunosuppressed individuals, such as those infected with HIV.

Various atypical mycobacteria found in water and soil may cause chronic indolent arthritis. Such disease results from trauma and direct inoculation associated with farming, gardening, or aquatic activities. Smaller joints, such as the digits, wrists, and knees, are usually involved. Involvement of tendon sheaths and bursae is typical. The mycobacterial species involved include *Mycobacterium marinum*, *M. avium-intracellulare*, *M. terrae*, *M. kansasii*, *M. fortuitum*, and *M. chelonae*. In persons who have HIV infection or are receiving immunosuppressive therapy, hematogenous spread to the joints has been reported for *M. kansasii*, *M. avium-intracellulare*, and *M. haemophilum*. Diagnosis usually requires biopsy and culture, and therapy is based on antimicrobial susceptibility patterns.

FUNGAL ARTHRITIS

Fungi are an unusual cause of chronic monarticular arthritis. Granulomatous articular infection with the endemic dimorphic fungi *Coccidioides immitis*, *Blastomyces dermatitidis*, and (less commonly) *Histoplasma capsulatum* results from hematogenous seeding or direct extension from bony lesions in persons with disseminated disease. Joint involvement is an unusual complication of sporotrichosis (infection with *Sporothrix schenckii*) among gardeners and other persons who work with soil or sphagnum moss. Articular sporotrichosis is six times more common among men than among women, and alcoholics and other debilitated hosts are at risk for polyarticular infection.

Candida infection involving a single joint, usually the knee, hip, or shoulder, results from surgical procedures, intraarticular injections, or (among critically ill patients with debilitating illnesses such as diabetes mellitus or hepatic or renal insufficiency and patients receiving immunosuppressive therapy) hematogenous spread. *Candida* infections in intravenous drug users typically involve the spine, sacroiliac joints, or other fibrocartilaginous joints. Unusual cases of arthritis due to *Aspergillus* species, *Cryptococcus neoformans*, *Pseudallescheria boydii*, and the dematiaceous fungi have also resulted from direct inoculation or disseminated hematogenous infection in immunocompromised persons.

The synovial fluid in fungal arthritis usually contains 10,000 to 40,000 cells/uL, with ~70% neutrophils. Stained specimens and cultures of synovial tissue often confirm the diagnosis of fungal arthritis when studies of synovial fluid give negative results. Treatment consists of drainage and lavage of the joint and systemic administration of amphotericin B, fluconazole, or itraconazole (the exact drug depending on the species involved). The doses and duration of therapy are the same as for disseminated disease (see Part Seven, Section 15). Intraarticular instillation of amphotericin B has been used in addition to intravenous therapy.

VIRAL ARTHRITIS

Viruses produce arthritis by infecting synovial tissue during systemic infection or by provoking an immunologic reaction that involves joints. As many as 50% of women report persistent arthralgias and 10% frank arthritis within 3 days of the rash that follows natural infection with *rubella virus* and within 2 to 6 weeks after receipt of live virus vaccine. Episodes of symmetric inflammation of fingers, wrists, and knees uncommonly recur for longer than a year, but a syndrome of chronic fatigue, low-grade fever, headaches, and myalgias can persist for months or years. Intravenous immunoglobulin has been helpful in selected cases. Self-limited monarticular or migratory polyarthritis may develop within 2 weeks of the parotitis of *mumps*; this sequela is more common in men than in women. Approximately 10% of children and 60% of women develop arthritis after infection with *parvovirus B19*. In adults, arthropathy sometimes occurs without fever or rash. Pain and stiffness, with less prominent swelling (primarily of the hands but also of the knees, wrists, and ankles), usually resolve within weeks, although a small proportion of patients develop chronic arthropathy.

About 2 weeks before the onset of jaundice, up to 10% of persons with acute *hepatitis B* develop an immune complex-mediated, serum sickness-like reaction with maculopapular rash, urticaria, fever, and arthralgias. Less common developments include symmetric arthritis involving the hands, wrists, elbows, or ankles and morning stiffness that resembles a flare of rheumatoid arthritis. Symptoms resolve at the time jaundice develops. Approximately one-third of persons with chronic hepatitis C infection report persistent arthralgia or arthritis, both in the presence and in the absence of cryoglobulinemia. Painful arthritis involving larger joints often accompanies the fever and rash of several arthropod-borne viral infections, including those caused by *chikungunya*, *O'nyong-nyong*, *Ross River*, *Mayaro*, and *Barmah Forest* viruses. Symmetric arthritis involving the hands and wrists may occur during the convalescent phase of infection with *lymphocytic choriomeningitis virus*. Patients infected with an *enterovirus* frequently report arthralgias, and *echovirus* has been isolated from patients with acute polyarthritis.

Several arthritis syndromes are associated with *HIV* infection. Reiter's syndrome with painful lower-extremity oligoarthritis often follows an episode of urethritis in HIV-infected persons. HIV-associated Reiter's syndrome appears to be extremely common among persons with the HLA-B27 haplotype, but sacroiliac joint disease is unusual and is seen mostly in the absence of HLA-B27. Up to one-third of HIV-infected persons with psoriasis develop psoriatic arthritis. Painless monarthropathy and persistent symmetric polyarthropathy occasionally complicate HIV infection. Chronic persistent oligoarthritis of the shoulders, wrists, hands, and knees occurs in women infected with human T lymphotropic virus type I. Synovial thickening, destruction of articular cartilage, and leukemic-appearing atypical lymphocytes in synovial fluid are characteristic, but progression to T cell leukemia is unusual.

PARASITIC ARTHRITIS

Arthritis due to parasitic infection is rare. The guinea worm *Dracunculus medinensis* may cause destructive joint lesions in the lower extremities as migrating gravid female worms invade joints or cause ulcers in adjacent soft tissues that become secondarily infected. Hydatid cysts infect bones in 1 to 2% of cases of infection with *Echinococcus granulosus*. The expanding destructive cystic lesions may spread to and destroy adjacent joints, particularly the hip and pelvis. In rare cases, chronic synovitis has been associated with the presence of schistosomal eggs in synovial biopsies. Monarticular arthritis in children with lymphatic *filariasis* appears to respond to therapy with diethylcarbamazine even in the absence of microfilariae in synovial fluid. Reactive arthritis has been attributed to *hookworm*, *Strongyloides*, *Cryptosporidium*, and *Giardia* infection in case reports, but confirmation is required.

POSTINFECTIOUS OR REACTIVE ARTHRITIS

Reiter's syndrome, a reactive polyarthritis, develops several weeks after ~1% of cases of nongonococcal urethritis and 2% of enteric infections, particularly those due to *Yersinia enterocolitica*, *Shigella flexneri*, *Campylobacter jejuni*, and *Salmonella* species. Only a minority of these patients have the other findings of classic Reiter's syndrome, including urethritis, conjunctivitis, uveitis, oral ulcers, and rash. Studies have identified microbial DNA or antigen in synovial fluid or blood, but the pathogenesis of this condition is poorly understood.

Reiter's syndrome is most common among young men (except after *Yersinia* infection) and has been linked to the HLA-B27 locus as a potential genetic predisposing factor. Patients report painful, asymmetric oligoarthritis affecting mainly the knees, ankles, and feet. Low-back pain is common, and radiographic evidence of sacroiliitis is found in patients with long-standing disease. Most patients recover within 6 months, but prolonged recurrent disease is more common in cases following chlamydial urethritis. Anti-inflammatory agents help to relieve symptoms, but the role of prolonged antibiotic therapy in eliminating microbial antigen from the synovium is controversial.

Migratory polyarthritis and fever constitute the usual presentation of acute rheumatic fever in adults. This presentation is distinct from that of poststreptococcal reactive arthritis, which also follows infections with group A b-hemolytic *Streptococcus* but is not migratory, lasts beyond the typical 3-week maximum of acute rheumatic fever, and responds poorly to aspirin.

INFECTIONS IN PROSTHETIC JOINTS

Infection complicates 1 to 4% of total joint replacements. The majority of infections are acquired intraoperatively or immediately postoperatively as a result of wound breakdown or infection; less commonly, these joint infections develop later after joint replacement and are the result of hematogenous spread or direct inoculation. The presentation may be acute, with fever, pain, and local signs of inflammation, especially in infections due to *S. aureus*, pyogenic streptococci, and enteric bacilli. Alternatively, infection may persist for months or years without causing constitutional symptoms when less virulent organisms, such as coagulase-negative staphylococci or diphtheroids, are involved. Such indolent infections are usually acquired during joint implantation and are discovered during evaluation of chronic unexplained pain or after a radiograph shows loosening of the prosthesis; the erythrocyte sedimentation rate and C-reactive protein are usually elevated in such cases.

The diagnosis is best made by needle aspiration of the joint; accidental introduction of organisms during aspiration must be meticulously avoided. Synovial fluid pleocytosis with a predominance of polymorphonuclear leukocytes is highly suggestive of infection, since other inflammatory processes uncommonly affect prosthetic joints. Culture and Gram's stain usually yield the responsible pathogen. Use of special media for unusual pathogens such as fungi, atypical mycobacteria, and *Mycoplasma* may be necessary if routine and anaerobic cultures are negative.

TREATMENT

Treatment includes surgery and high doses of parenteral antibiotics, which are given for 4 to 6 weeks because bone is usually involved. In most cases, the prosthesis must be replaced to cure the infection. Implantation of a new prosthesis is best delayed for several weeks or months because relapses of infection occur most commonly within this time frame. In some cases, reimplantation is not possible, and the patient must manage without a joint, with a fused joint, or even with amputation. Cure of infection without removal of the prosthesis is occasionally possible in cases that are due to streptococci or pneumococci and that lack radiologic evidence of loosening of the prosthesis. In these cases, antibiotic therapy must be initiated within several days of the onset of infection, and the joint should be drained vigorously either by open arthrotomy or arthroscopically. A high cure rate with retention of the prosthesis has been reported when the combination of oral rifampin and ciprofloxacin is given for 3 to 6 months to persons with staphylococcal prosthetic joint infection of short duration. This approach, which is based on the ability of rifampin to kill organisms adherent to foreign material and in the stationary growth phase, requires confirmation in prospective trials.

Prevention To avoid the disastrous consequences of infection, candidates for joint replacement should be selected with care. Rates of infection are particularly high among patients with rheumatoid arthritis, persons who have undergone previous surgery on the joint, and persons with medical conditions requiring immunosuppressive therapy. Perioperative antibiotic prophylaxis, usually with cefazolin, and measures to decrease intraoperative contamination, such as laminar flow, have lowered the rates of perioperative infection to <1% in many centers. After implantation, measures should be taken to prevent or rapidly treat extraarticular infections that might give rise to hematogenous spread to the prosthesis. The effectiveness of prophylactic antibiotics for the prevention of hematogenous infection following dental procedures has not been demonstrated; in fact, viridans streptococci and other components of the oral flora are extremely unusual causes of prosthetic joint infection. Accordingly, the American Dental Association and the American Academy of Orthopaedic Surgeons do not recommend antibiotic prophylaxis for most dental patients with total joint replacements. They do, however, recommend prophylaxis for patients who may be at high risk of hematogenous infection, including those with inflammatory arthropathies, immunosuppression, type 1 diabetes mellitus, joint replacement within 2 years, previous prosthetic joint infection, malnourishment, or hemophilia. The recommended regimen is amoxicillin (2 g orally) 1 h before dental procedures associated with a high incidence of bacteremia. Clindamycin (600 mg orally) is suggested for patients allergic to penicillin.

(Bibliography omitted in Palm version)

Back to Table of Contents

324. PSORIATIC ARTHRITIS AND ARTHRITIS ASSOCIATED WITH GASTROINTESTINAL DISEASE - *Peter H. Schur*

PSORIATIC ARTHRITIS

Psoriatic arthritis (PsA) is a chronic inflammatory arthritis that affects 5 to 42% of people with psoriasis.

ETIOLOGY AND PATHOGENESIS

To date, the cause and pathogenesis ofPsA are unknown. Indirect evidence has suggested that interactions of infections, trauma, increased humoral and cellular immunity (e.g., to streptococci), cytokines (including TH1 and TH2), adhesion molecules, and abnormal fibroblast, dendritic cell, keratinocyte, and polymorphonuclear leukocyte (PMN) function are involved. Polyarthritis has developed in patients with psoriasis and hepatitis treated with interferon a. Most studies have observed an increased frequency of HLA-B17, CW6, and/or B27 in patients with psoriatic spondylitis, while B27, B38, B39, and DR7 have been noted in association with peripheral arthritis in different studies. Fulminant disease should make one suspect HIV disease (Chap. 309).

CLINICAL MANIFESTATIONS

Three major types ofPsA are generally recognized: asymmetric inflammatory arthritis, symmetric arthritis, and psoriatic spondylitis. A mean of 47% of patients (range, 16 to 70%) have an asymmetric inflammatory arthritis. Disease appears equally in men and women. Psoriasis tends to precede the arthritis by years. Many patients complain of morning stiffness. The proximal interphalangeal (PIP) and distal interphalangeal (DIP) joints are commonly involved [with characteristic sausage-shaped digits (dactylitis)], while knees, hips, ankles, temporomandibular joints, and wrists are less frequently involved. Most patients have onychodystrophy (onycholysis, ridging and pitting of nails), the course of which does not parallel that of the synovitis. The prognosis is good, with only one-fourth of the patients developing progressive destructive disease; one-third develop inflammatory ocular complications (conjunctivitis, iritis, episcleritis).

A mean of 25% of patients (range, 15 to 39%) develop symmetric arthritis resembling rheumatoid arthritis (Chap. 312). This disease occurs twice as frequently in women. Psoriasis and inflammatory arthritis usually develop simultaneously; most patients experience morning stiffness. TheDIP,PIP, metacarpophalangeal (MCP), metatarsophalangeal (MTP), sternoclavicular, and, in particular, large peripheral joints are involved. Practically all patients have onychodystrophy, which helps distinguish them from patients with rheumatoid arthritis. Over half of the patients in this group go on to develop destructive arthritis, including arthritis mutilans. Eye complications are uncommon. Subcutaneous nodules are not present, but one-fourth of patients have rheumatoid factors. Unilateral upper limb edema has been described.

A mean of 23% of the patients (range 5 to 40%) have psoriatic "spondylitis," with or without peripheral joint involvement. Psoriasis tends to precede the arthritis by a few years, and low back pain with morning stiffness is common. Psoriatic spondylitis is more common in men. About half the patients in this group have spondylitis and the other half have sacroiliitis. The back disease is usually slowly progressive, with little clinical deterioration as compared with ankylosing spondylitis; the peripheral disease also tends not to be destructive except for the occasional patient with arthritis mutilans. Enthesopathy, i.e., inflammation of tendons and ligamentous attachments to bone, is characteristic, for example, of the Achilles tendon or of the plantar fascia causing heel pain. Many patients have onychodystrophy, but few have inflammatory ocular complications. Gut inflammation occurs in 30% (no gut inflammation was noted in patients with only peripheral arthritis).

Some authors have described additional subsets of psoriatic arthritis: predominantDIPjoint involvement, arthritis mutilans, peripheral enthesitis, juvenilePsA, and SAPHO (synovitis, acne, pustulosis, hyperostosis, osteomyelitis).

The pathology ofPsA is similar to that seen in rheumatoid arthritis: synoviocytic hyperplasia, earlyPMNinfiltration and later mononuclear cell infiltration, cartilage erosion, and pannus formation. However, in PsA, the synovium is more vascular and there are fewer macrophages and less expression of endothelial cell leukocyte adhesion molecule-1 (ELAM-1). Fibrosis of the joint capsule and marrow is prominent in many patients.

LABORATORY FINDINGS

There are few laboratory abnormalities. Elevated erythrocyte sedimentation rates, C-reactive proteins, and complement levels reflect inflammation. Rheumatoid factors are uncommon and are more likely to be observed in those with symmetric arthritis. Immunoglobulin levels, especially IgA levels, may be elevated (IgA antibodies to cytokeratins and antienterobacteria antibodies are elevated). Uric acid levels may be elevated; sodium urate crystals in joint fluids suggest gout.

Radiologic investigation reveals findings similar to those of rheumatoid arthritis: soft tissue swelling, loss of the cartilage space, erosions, bony ankylosis of fingers, subluxations, and subchondral cysts; of note, there is less demineralization. However, more unique and suggestive of psoriatic arthritis are erosions atDIPjoints, expansion of the base of the terminal phalanx, tapering of the proximal phalanx and cuplike erosions and bony proliferation of the distal terminal phalanx ("pencil-in-cup" appearance), proliferation of bone near osseous erosions, terminal phalangeal osteolysis, bone proliferation and periostitis (especially of phalanges), and telescoping of one bone into its neighbor, leading to the "opera-glass" deformity (Fig. 324-1). The axial skeleton shows asymmetric or unilateral sacroiliitis, often asymptomatic paravertebral ossification, including cervical involvement, and large asymmetric nonmarginal syndesmophytes. Echocardiographic abnormalities resemble those of ankylosing spondilitis.

DIAGNOSIS

The diagnosis ofPsAshould be considered in individuals with arthritis and psoriasis. Psoriasis should be distinguished from seborrheic dermatitis and eczema. Psoriatic lesions may be quite small peripherally and are often hidden in the scalp, umbilicus, and gluteal folds. Fungal infection of nails can be distinguished from psoriasis, for the latter will demonstrate pitting and onycholysis. Furthermore, onychodystrophy is uncommon (20% of cases) in uncomplicated psoriasis. It is often difficult to distinguish Reiter's syndrome (Chap. 315) from PsA, since both manifest dactylitis. Reiter's syndrome usually presents in younger individuals, especially males; is less frequently progressive or destructive; and is more likely to be associated with characteristic skin lesions (keratoderma blenorrhagica -- which may, however, resemble pustular psoriasis), urethritis, and conjunctivitis. Gout can be distinguished by the presence of intraarticular sodium urate crystals (Chap. 322). Psoriasis in association with Heberden's nodes or Bouchard's nodes of theDIP andPIPjoints, respectively, rather suggests osteoarthritis (Chap. 321). PsA differs from rheumatoid arthritis by the relative lack of rheumatoid factors; the tendency to asymmetry, dactylitis, iritis, enthesopathy, and onychodystrophy; the high frequency of HLA-B27, especially in patients with axial skeletal involvement; and characteristic radiologic features.

TREATMENT

The treatment ofPsAbegins with patient education and physical and occupational therapy to maintain muscle strength and joint and muscle function. Orthotics and occasional intraarticular glucocorticoids for isolated acutely and severely inflamed joints may be added as needed (Fig. 324-2). The mainstay, however, is the use of nonsteroidal anti-inflammatory drugs (NSAIDs), which reduce inflammation and alleviate pain for most patients. For patients with more severe involvement, a disease-modifying antirheumatic drug should be used. While hydroxychloroquine is often successful in producing either amelioration or remission, it carries a significant risk of exacerbation of psoriasis and exfoliation. Sulfasalizine (2 to 4 g/d) has well-demonstrated efficacy for PsA. For more severe cases, especially with extensive skin involvement, 5 to 25 mg methotrexate per week is recommended. Most patients respond well with respect to both skin lesions and arthritis. Patients who are resistant to oral therapy may respond to parenteral therapy. Folic acid (1 mg/d) is recommended to prevent hematologic complications. Renal and liver function tests and a complete blood count should be performed every 6 to 8 weeks, and any abnormalities should suggest modification of the dosage. Liver biopsies are recommended after a total of 1.5 g methotrexate have been given and then every 2 years to identify the rare patient with fibrosis and cirrhosis, which necessitate withdrawal of the drug. Patients are advised to avoid nephrotoxic and hepatotoxic (e.g., ethanol) drugs. Patients with HIV infection may have worsening of their disease when treated with methotrexate. Intramuscular gold, cyclosporine (2 to 5 mg/kg per day), etretinate (0.5 mg/kg per day), and azathioprine have also proved successful. The arthritis may also respond to heliotherapy.

ARTHRITIS ASSOCIATED WITH GASTROINTESTINAL DISEASE

INFLAMMATORY BOWEL DISEASE

Peripheral arthritis occurs in 9 to 30% of patients with inflammatory bowel disease (IBD) (e.g., ulcerative colitis or Crohn's disease; Chap. 287), and arthralgia is more common. Arthritis is somewhat more likely to occur in patients with large-bowel disease and in those patients with complications such as abscesses, pseudomembranous polyposis, perianal disease, massive hemorrhage, erythema nodosum, stomatitis, uveitis, and pyoderma gangrenosum. Males and females are affected equally. The arthritis tends to be acute, is associated with a flare-up of the bowel disease, occurs early in the course of the bowel disease, is self-limiting (90% of cases resolve within 6 months), and does not result in destruction. Most patients have a symmetric, migratory polyarthritis affecting primarily large joints of the lower extremity. Rheumatoid factors are not present. There is some association with HLA-BW62. Synovial fluids have 5000 to 12,000 white blood cells per microliter, mostlyPMNs. Radiographs demonstrate soft tissue swelling and effusions without erosions or destruction. Pathologic examination of synovial biopsy specimens reveals only nonspecific inflammation. The peripheral arthritis responds to successful treatment of the bowel disease, such as colectomy (for ulcerative colitis), or administration of glucocorticoids, anti-tumor necrosis factora therapy, or sulfasalazine.NSAIDsrelieve pain and inflammation but should be used with caution because of possible gastrointestinal side effects.

Spondylitis occurs in 1.1 to 43% of patients withIBD(while gut inflammation develops in 68% of patients with spondyloarthropathies). Spondylitis often precedes IBD; their clinical courses are often strongly related. Males are affected more frequently. Patients typically complain of stiffness in the back and/or buttocks in the morning or after rest. The stiffness and associated pain are often relieved by exercise. Gastrointestinal infection/inflammation is thought to play a role in exacerbation of spondylitis. Physical examination reveals limitation of spinal flexion and reduced chest expansion. Some patients may have peripheral arthritis, especially of the hips and/or shoulders. Uveitis is a frequent complication. Radiographs of the back show the typical findings of ankylosing spondylitis and bilateral sacroiliitis. HLA-B27 is found in 53 to 75% of these patients. Treatment includes physical therapy,NSAIDs, glucocorticoids, and sulfasalazine. NSAIDs should be used with caution lest they exacerbate the IBD. The axial disease progresses slowly in a manner akin to that of ankylosing spondylitis.

Asymptomatic sacroiliitis detected by radiography occurs in 4 to 32% of patients withIBD. By contrast, 52% of patients with IBD have abnormalities on technetium pyrophosphate bone scans of the sacroiliac joint. There is no increased frequency of HLA-B27 in these patients. This "disease" does not necessarily progress to spondylitis.

Other complications of chronicIBDinclude (1) finger clubbing (observed in 4 to 13% of patients with Crohn's disease, especially those with small-bowel involvement), which may regress after surgery; (2) development of amyloid, especially in association with Crohn's disease; and (3) osteoporosis resulting from inactivity, malabsorption, and/or treatment with glucocorticoids. Osteomalacia can result from malabsorption. In this setting, with acutely increased back pain, one should suspect compression fracture.

INTESTINAL BYPASS ARTHRITIS

Intestinal bypass surgery was developed for the treatment of obesity in 1952; 11 years later arthritis was recognized as a postoperative complication. Polyarthralgia, tenosynovitis, and sometimes arthritis occur weeks, even years, after surgery in 8 to 36% of patients. There is often an associated urticarial, vesicular, pustular, macular, or nodular eruption. X-rays generally show no joint damage. Tests for rheumatoid factors, antinuclear antibodies, and HLA-B27 are usually negative, while immune complexes (and cryoglobulins) are often present. They contain bacterial antigens, the corresponding antibodies, IgA secretory component, and complement components.

These observations suggest that the syndrome has the following pathogenesis: Bacteria proliferate in intestinal blind loops; bacterial antigens are absorbed; and antibodies to these antigens develop and combine with them to form immune complexes, which deposit in synovial tissue to cause arthritis.NSAIDsand glucocorticoids can relieve the joint symptoms, but more lasting results can be achieved by tetracycline therapy to decrease the bacterial load; even better is reanastomosis of the bowel or resection of the blind loop.

WHIPPLE'S DISEASE (INTESTINAL LIPODYSTROPHY)

Whipple's disease is rare and occurs mostly in middle-aged Caucasian males, who develop arthritis, prolonged diarrhea, malabsorption, and weight loss as a result of infection with the actinomyocete, *Tropheryma whippelii*. The organism has been found in waste water. Up to 90% of patients with the disease develop arthritis, usually prior to other symptoms. Knees and ankles and, to a lesser extent, fingers, hips, shoulders, elbows, and wrists are involved. The arthritis is acute in onset, migratory, usually lasts just a few days, and is rarely chronic or a cause of permanent joint damage. Associated symptoms may include fever (54%), edema, serositis (pleurisy, pericarditis, endocarditis), pneumonia, hypotension, lymphadenopathy (54%), hyperpigmentation (54%), subcutaneous nodules, clubbing, and uveitis. Central nervous system involvement may develop (80%), with cognitive changes, headache, diplopia, and papilledema, and may be appreciated by abnormalities in magnetic resonance images of the brain. Oculomasticatory myorhythmia and oculo-facial-skeletal myorhythmia are felt to be pathognomonic and are found in 20% of patients; they are always accompanied by supranuclear vertical gaze palsy. Laboratory abnormalities include anemia (75%), low serum levels of carotene (95%), and albumin (93%). The presence of HLA-B27 (8 to 30%) occurs in those patients with axial arthritis. Synovial fluids have been reported to contain 450 to 36,000 white blood cells per microliter (30 to 95% neutrophils) or a mild monocytosis. Joint x-rays rarely show erosions but may show a sacroiliitis in the occasional patients who have axial skeletal symptoms; abdominal computed tomographic scans may reveal lymphadenopathy. The lamina propria and/or foamy macrophages in small intestine contain PAS-staining bacterial remnants, presumably of *T. whippelii*. These inclusion-containing foamy macrophages have also been detected in the synovium, lymph nodes, and other tissues. Diagnosis is often established by polymerase chain reaction of the 165 ribosomal gene sequences of these bacteria in biopsied tissue, usually the duodenum. The syndrome responds best to therapy with penicillin (or ceftriaxone) and streptomycin for 2 weeks followed by trimethoprim-sulfamethoxazole for 1 to 2 years. However, central nervous system relapse may develop, which has been treated with cefixime.

REACTIVE ARTHRITIS

A Reiter's-like syndrome of arthritis can develop 2 to 3 weeks following diarrhea caused by *Shigella*, *Salmonella*, *Yersinia*, *Chlamydia trachomatis*, or *Campylobacter* organisms.**This condition is described in Chap. 315.*

(Bibliography omitted in Palm version)

Back to Table of Contents

325. RELAPSING POLYCHONDRITIS AND OTHER ARTHRITIDES- *Bruce C. Gilliland*

RELAPSING POLYCHONDRITIS

Relapsing polychondritis is an uncommon inflammatory disorder of unknown cause characterized by an episodic and generally progressive course affecting predominantly the cartilage of the ears, nose, and laryngotracheobronchial tree. Other manifestations include scleritis, neurosensory hearing loss, polyarthritis, vasculitis, cardiac abnormalities, skin lesions, and glomerulonephritis. The peak age of onset is between the ages of 40 to 50 years but relapsing polychondritis may affect children and the elderly. It is found in all races, and both sexes are equally affected. No familial tendency is apparent. A significantly higher frequency of HLA-DR4 has been found in patients with relapsing polychondritis than in normal individuals. A predominant subtype allele(s) of HLA-DR4 was not found. Approximately 30% of patients with relapsing polychondritis will have another rheumatologic disorder, the most frequent being systemic vasculitis, followed by rheumatoid arthritis, systemic lupus erythematosus (SLE), or Sjogren's syndrome. Nonrheumatic disorders associated with relapsing polychondritis include inflammatory bowel disease, primary biliary cirrhosis, and myelodysplastic syndrome.

Diagnostic criteria were suggested several years ago by McAdam et al. and modified by Damiani and Levine a few years later. These criteria continue to be generally used in clinical practice. McAdam et al. proposed the following: (1) recurrent chondritis of both auricles; (2) nonerosive inflammatory arthritis; (3) chondritis of nasal cartilage; (4) inflammation of ocular structures including conjunctivitis, keratitis, scleritis/episcleritis, and/or uveitis; (5) chondritis of the laryngeal and/or tracheal cartilages; and (6) cochlear and/or vestibular damage manifested by neurosensory hearing loss, tinnitus, and/or vertigo. The diagnosis is certain when three or more of these features were present with biopsy confirmation. Damiana and Levine later suggested that the diagnosis could be made when one or more of the above features and a positive biopsy were present, when two or more separate sites of cartilage inflammation were present that responded to glucocorticoids or dapsone, or when three or more of the above features were present. A biopsy is not necessary in most patients with clinically evident disease.

PATHOLOGY AND PATHOPHYSIOLOGY

The earliest abnormality of cartilage noted histologically is a focal or diffuse loss of basophilic staining indicating depletion of proteoglycan from the cartilage matrix. Inflammatory infiltrates are found adjacent to involved cartilage and consist predominantly of mononuclear cells and occasional plasma cells. In acute disease, polymorphonuclear white cells may also be present. Destruction of cartilage begins at the outer edges and advances centrally. There is lacunar breakdown and loss of chondrocytes. Degenerating cartilage is replaced by granulation tissue and later by fibrosis and focal areas of calcification. Small loci of cartilage regeneration may be present. Immunofluorescence studies have shown immunoglobulins and complement at sites of involvement. Fine granular material observed in the degenerating cartilage matrix by electron microscopy has been interpreted to be enzymes or immunoglobulins.

Immunologic mechanisms play a role in the pathogenesis of relapsing polychondritis.

Immunoglobulin and complement deposits are found at sites of inflammation. In addition, antibodies to type II collagen and to matrilin-1 and immune complexes are detected in the sera of some patients. The possibility that an immune response to type II collagen may be important in the pathogenesis is supported experimentally by the occurrence of auricular chondritis in rats immunized with type II collagen. Antibodies to type II collagen are found in the sera of these animals, and immune deposits are detected at sites of ear inflammation. Cell-mediated immunity may also be operative in causing tissue injury, since lymphocyte transformation can be demonstrated when lymphocytes of patients are exposed to cartilage extracts. Humoral and cellular immune responses to type IX and type XI collagen have been demonstrated in some patients. In a recent study, rats immunized with cartilage matrix protein (matrilin-1) were found to develop severe inspiratory stridor and swelling of the nasal septum. Cartilage matrix protein is a noncollagenous protein present in the extracellular matrix in cartilage. It is present in high concentrations in the trachea and is also present in the nasal septum but not in articular cartilage. The immunized rats had severe inflammation in the larynx close to the epiglottis, which was characterized by increased numbers of CD 4+ and CD 8+ T cells. All had IgG antibodies to cartilage matrix protein. The inflammation was believed to have been largely mediated by T cells. The results of the study suggest that immune responses to various cartilage proteins play a role in the pathogenesis of relapsing polychondritis.

Dissolution of cartilage matrix can be induced by the intravenous injection of crude papain, a proteolytic enzyme, into young rabbits, which results in collapse of their normally rigid ears within 4 h. Reconstitution of the matrix occurs in about 7 days. In relapsing polychondritis, loss of cartilage matrix also most likely results from action of proteolytic enzymes released from chondrocytes, polymorphonuclear white cells, and monocytes that have been activated by inflammatory mediators.

CLINICAL MANIFESTATIONS

The onset of relapsing polychondritis is frequently abrupt with the appearance of one or two sites of cartilagenous inflammation. Fever, fatigue, and weight loss occur and may precede the clinical signs of relapsing polychondritis by several weeks. Relapsing polychondritis may go unrecognized for several months or even years in patients who only initially manifest intermittent joint pain and/or swelling, or who have unexplained eye inflammation, hearing loss, valvular heart disease, or pulmonary symptoms. The pattern of cartilagenous involvement and the frequency of episodes vary widely among patients.

Auricular chondritis is the most frequent presenting manifestation of relapsing polychondritis in 40% of patients and eventually affects about 85% of patients (Table 325-1). One or both ears are involved, either sequentially or simultaneously. Patients experience the sudden onset of pain, tenderness, and swelling of the cartilaginous portion of the ear. Earlobes are spared because they do not contain cartilage. The overlying skin has a beefy red or violaceous color. Prolonged or recurrent episodes result in a flabby or droopy ear as a sequela of cartilage destruction. Swelling may close off the eustachian tube (causing otitis media) or the external auditory meatus, either of which can impair hearing. Inflammation of the internal auditory artery or its cochlear branch produces hearing loss, vertigo, ataxia, nausea, and vomiting. The cartilage of the nose becomes inflamed during the first or subsequent attacks. Approximately 50% of patients will eventually have nose involvement. Patients may experience nasal stuffiness, rhinorrhea, and epistaxis. The bridge of the nose becomes red, swollen, and tender and may collapse, producing a saddle deformity. In some patients, the saddle deformity develops insidiously without overt inflammation. Saddle nose is observed more frequently in younger patients, especially in women.

Arthritis is the presenting manifestation in relapsing polychondritis in approximately one-third of patients and may be present for several months before other features appear. Eventually, more than half the patients will have arthritis. The arthritis is usually asymmetric and oligo- or polyarticular, and involves both large and small peripheral joints. An episode of arthritis lasts from a few days to several weeks and resolves spontaneously without residual joint deformity. Attacks of arthritis may not be temporally related to other manifestations of relapsing polychondritis. The joints are warm, tender, and swollen. Joint fluid has been reported to be noninflammatory. In addition to peripheral joints, inflammation may involve the costochondral, sternomanubrial, and sternoclavicular cartilages. Destruction of these cartilages may result in a pectus excavatum deformity or even a flail anterior chest wall. Relapsing polychondritis may occur in patients with preexisting rheumatoid arthritis, Reiter's syndrome, psoriatic arthritis, or ankylosing spondylitis.

Eye manifestations occur in more than half of patients and include conjunctivitis, episcleritis, scleritis, iritis, and keratitis. Ulceration and perforation of the cornea may occur and cause blindness. Other manifestations include eyelid and periorbital edema, proptosis, cataracts, optic neuritis, extraocular muscle palsies, retinal vasculitis, and renal vein occlusion.

Laryngotracheobronchial involvement occurs in ~50% of patients. Symptoms include hoarseness, a nonproductive cough, and tenderness over the larynx and proximal trachea. Mucosal edema, strictures, and/or collapse of laryngeal or tracheal cartilage may cause stridor and life-threatening airway obstruction necessitating tracheostomy. Collapse of cartilage in bronchi leads to pneumonia and, when extensive, to respiratory insufficiency.

Aortic regurgitation occurs in about 5% of patients and is due to progressive dilation of the aortic ring or to destruction of the valve cusps. Mitral and other heart valves are less often affected. Other cardiac manifestations include pericarditis, myocarditis, and conduction abnormalities. Aneurysms of the proximal, thoracic, or abdominal aorta may occur and occasionally rupture.

Systemic vasculitis may occur in association with relapsing polychondritis. Vasculitides include leukocytoclastic vasculitis, polyarteritis, temporal arteritis, and Takayasu's arteritis (Chap. 317). Neurologic abnormalities usually occur as a result of underlying vasculitis, manifesting as seizures, strokes, ataxia, and peripheral and cranial nerve neuropathies. Cranial nerves VI and VII are most often involved. Approximately 25% of patients have skin lesions, none of which is characteristic for relapsing polychondritis. These include purpura, erythema nodosum, erythema multiforme, angioedema/urticaria, livedo reticularis, and panniculitis. Segmental necrotizing glomerulonephritis with crescent formation has been noted in some patients in the absence of systemic vasculitis.

The course of disease is highly variable, with episodes lasting from a few days to several weeks and then subsiding spontaneously. Attacks may recur at intervals varying from weeks to months. In other patients, the disease has a chronic, smoldering course. In a few patients, the disease may be limited to one or two episodes of cartilage inflammation. In one study, the 5-year estimated survival rate was 74% and the 10-year survival rate 55%. In contrast to earlier series, only about half the deaths could be attributed to relapsing polychondritis or complications of treatment. Pulmonary complications accounted for only 10% of all fatalities. In general, patients with more widespread disease have a worse prognosis.

LABORATORY FINDINGS

Mild leukocytosis and normocytic, normochromic anemia are often present. The erythrocyte sedimentation rate is usually elevated. Rheumatoid factor and antinuclear antibody tests are occasionally positive in low titers. Antibodies to type II collagen are present in most patients, but they are not specific. Circulating immune complexes may be detected, especially in patients with early active disease. Elevated levels ofg globulin may be present. Antineutrophil cytoplasmic antibodies (ANCA), either cytoplasmic (C-ANCA) or perinuclear (P-ANCA), are found in some patients with active disease. The upper and lower airways can be evaluated by imaging techniques such as linear tomography, laryngotracheography, and computed tomography, and by bronchoscopy. Bronchography is performed to demonstrate bronchial narrowing. Intrathoracic airway obstruction can also be evaluated by inspiratory-expiratory flow studies. The chest film may show narrowing of the trachea and/or the main bronchi, widening of the ascending or descending aorta due to an aneurysm, and cardiomegaly when aortic insufficiency is present. Radiographs may show calcification at previous sites of cartilage damage involving ear, nose, larynx, or trachea.

DIAGNOSIS

Diagnosis is based on recognition of the typical clinical features. Biopsies of the involved cartilage from the ear, nose, or respiratory tract will confirm the diagnosis but are only necessary when clinical features are not typical. Patients with Wegener's granulomatosis may have a saddle nose and pulmonary involvement but can be distinguished by the absence of auricular involvement and the presence of granulomatous lesions in the tracheobronchial tree. Patients with Cogan's syndrome have interstitial keratitis and vestibular and auditory abnormalities, but this syndrome does not involve the respiratory tract or ears. Reiter's syndrome may initially resemble relapsing polychondritis because of oligoarticular arthritis and eye involvement, but it is distinguished in time by the appearance of urethritis and typical mucocutaneous lesions and the absence of nose or ear cartilage involvement. Rheumatoid arthritis may initially suggest relapsing polychondritis because of arthritis and eye inflammation. The arthritis in rheumatoid arthritis, however, is erosive and symmetric. In addition, rheumatoid factor titers are usually high compared with those in relapsing polychondritis. Bacterial infection of the pinna may be mistaken for relapsing polychondritis but differs by usually involving only one ear, including the earlobe. Auricular cartilage may also be damaged by trauma or frostbite.

Relapsing polychondritis may develop in patients with a variety of autoimmune disorders, includingSLE, rheumatoid arthritis, Sjogren's syndrome, and vasculitis. In most cases, these disorders antedate the appearance of polychondritis, usually by months or years. It is likely that these patients have an immunologic abnormality that predisposes them to development of this group of autoimmune disorders.

TREATMENT

In patients with active chondritis or associated vasculitis, prednisone, 40 to 60 mg/d, is often effective in suppressing disease activity; it is tapered gradually once disease is controlled. In some patients, prednisone can be stopped, while in others low doses in the range of 10 to 15 mg/d are required for continued suppression of disease. Immunosuppressive drugs such as methotrexate, cyclophosphamide, or azathioprine should be reserved for patients who fail to respond to prednisone or who require high doses for control of disease activity. Methotrexate has been found by some investigators to be very effective in treating relapsing polychondritis. Dapsone and cyclosporine have been reported to be beneficial in a few patients. Patients with significant ocular inflammation often require intraocular steroids as well as high doses of prednisone. Heart valve replacement or repair of an aortic aneurysm may be necessary. In patients with early subglottic disease, intralesional injection of glutocorticoids may be beneficial. When obstruction is severe, tracheostomy is required. Stents may be necessary in patients with tracheobronchial collapse.

OTHER ARTHRITIDES

NEUROPATHIC JOINT DISEASE

Neuropathic joint disease (Charcot's joint) is a progressive destructive arthritis associated with loss of pain sensation, proprioception, or both. In addition, normal muscular reflexes that modulate joint movement are decreased. Without these protective mechanisms, joints are subjected to repeated trauma, resulting in progressive cartilage and bone damage. Neuropathic arthropathy was first described by Jean-Martin Charcot in 1868 in patients with tabes dorsalis. The term *Charcot joint* is commonly used interchangeably with *neuropathic joint*. Today, diabetes mellitus is the most frequent cause of neuropathic joint disease. A variety of other disorders are associated with neuropathic arthritis including leprosy, yaws, syringomyelia, meningomyelocoele, congenital indifference to pain, peroneal muscular atrophy (Charcot-Marie-Tooth disease), and amyloidosis. An arthritis resembling neuropathic joint disease is seen in patients who have received frequent intraarticular glucocorticoid injections into a weight-bearing joint and in patients with calcium pyrophosphate dihydrate crystal deposition disease. The distribution of joint involvement depends on the underlying neurologic disorder (Table 325-2). In tabes dorsalis, knees, hips, and ankles are most commonly affected; in syringomyelia, the glenohumeral joint, elbow, and wrist; and in diabetes mellitus, the tarsal and tarsometatarsal joints.

Pathology and Pathophysiology The pathologic changes in the neuropathic joint are similar to those found in the severe osteoarthritic joint. There is fragmentation and eventual loss of articular cartilage with eburnation of the underlying bone. Osteophytes are found at the joint margins. With more advanced disease, erosions are present on the joint surface. Fractures, devitalized bone, and intraarticular loose bodies may be present. Microscopic fragments of cartilage and bone are seen in the synovial tissue.

At least two underlying mechanisms are believed to be involved in the pathogenesis of neuropathic arthritis. An abnormal autonomic nervous system is thought to be responsible for the increased blood flow to the joint and subsequent resorption of bone. Loss of bone, particularly in the diabetic foot, may be the initial manifestation. With the loss of deep pain, proprioception, and protective neuromuscular reflexes, the joint is subjected to repeated injuries including ligamental tears and bone fractures. The mechanism of injury that occurs following frequent intraarticular glucocorticoid injections is thought to be due to the analgesic effect of glucocorticoids leading to overuse of an already damaged joint, which results in accelerated cartilage damage. It is not understood why only a few patients with neuropathies develop neuropathic arthritis.

Clinical Manifestations Neuropathic joint disease usually begins in a single joint and then progresses to involve other joints, depending on the underlying neurologic disorder. The involved joint progressively becomes enlarged from bony overgrowth and synovial effusion. Loose bodies may be palpated in the joint cavity. Joint instability, subluxation, and crepitus occur as the disease progresses. Neuropathic joints may develop rapidly, and a totally disorganized joint with multiple bony fragments may evolve in a patient within weeks or months. The amount of pain experienced by the patient is less than would be anticipated based on the degree of joint involvement. Patients may experience sudden joint pain from intraarticular fractures of osteophytes or condyles.

Neuropathic arthritis is encountered most often in patients with diabetes mellitus, with the incidence estimated in the range of 0.5%. The usual age of onset is³50 years following several years of diabetes, but exceptions occur. The tarsal and tarsometatarsal joints are most often affected, followed by the metatarsophalangeal and talotibial joints. The knees and spine are occasionally involved. In about 20%, neuropathic arthritis may be present in both feet. Patients often attribute the onset of foot pain to antecedent trauma such as twisting their foot. Neuropathic changes may develop rapidly following a foot fracture or dislocation. Swelling of the foot and ankle are often present. Downward collapse of the tarsal bones leads to convexity of the sole, referred to as a "rocker foot." Large osteophytes may protrude from the top of the foot. Calluses frequently form over the metatarsal heads and may lead to infected ulcers and osteomyelitis. Radiographs may show resorption and tapering of the distal metatarsal bones. The term *Lisfranc fracture-dislocation* is sometimes used to describe the destructive changes at the tarsometatarsal joints.

Diagnosis The diagnosis of neuropathic arthritis is based on the clinical features and characteristic radiographic findings in a patient with an underlying sensory neuropathy. The differential diagnosis of neuropathic arthritis includes osteomyelitis, osteonecrosis, advanced osteoarthritis, stress fractures, and calcium pyrophosphate dihydrate (CPDD) deposition disease. Radiographs in neuropathic arthritis initially show changes of osteoarthritis with joint space narrowing, subchondral bone sclerosis, osteophytes, and joint effusions followed later by marked destructive and hypertrophic changes. Soft tissue swelling, bone resorption, fractures, large osteophytes, extraarticular bone fragments, and subluxation are present with advanced arthropathy. The radiographic findings of neuropathic arthritis may be difficult to differentiate from those of osteomyelitis, especially in the diabetic foot. The joint margins in a neuropathic joint tend to be distinct, while in osteomyelitis, they are blurred. Imaging studies and cultures of fluid and tissue from the joint are often required to exclude osteomyelitis. Magnetic resonance imaging is helpful in differentiating these disorders. Another useful study is a bone scan using indium 111-labeled white blood cells or indium 111-labeled immunoglobulin G, which will show an increased uptake in osteomyelitis but not in a neuropathic joint. A technetium bone scan will not distinguish osteomyelitis from neuropathic arthritis as increased uptake is observed in both. The joint fluid in neuropathic arthritis is noninflammatory; may be xanthochromic or even bloody; and may contain fragments of synovium, cartilage, and bone. The finding of CPPD crystals suggests the diagnosis of a crystal associated neuropathic-like arthropathy. In the absence of CPPD crystals, the presence of increased number of leukocytes may indicate osteomyelitis.

TREATMENT

The primary focus of treatment is to provide stabilization of the joint. Treatment of the underlying disorder, even if successful, does not usually alter the joint disease. Braces and splints are helpful. Their use requires close surveillance, since patients may be unable to appreciate pressure from a poorly adjusted brace. In the diabetic patient, early recognition and treatment of a Charcot's foot by prohibiting weight bearing of the foot for at least 8 weeks may possibly prevent severe disease from developing. Fusion of a very unstable joint may improve function, but nonunion is frequent, especially when immobilization of the joint is inadequate.

HYPERTROPHIC OSTEOARTHROPATHY AND CLUBBING

Hypertrophic osteoarthropathy (HOA) is characterized by clubbing of digits and, in more advanced stages, by periosteal new bone formation and synovial effusions. HOA occurs in primary and familial forms and usually begins in childhood. The secondary form of HOA is associated with intrathoracic malignancies, suppurative lung disease, congenital heart disease, and a variety of other disorders and is more common in adults. Clubbing is almost always a feature of HOA but can occur as an isolated manifestation (Fig. 325-1). The presence of clubbing in isolation is generally considered to represent either an early stage or an element in the spectrum of HOA. The presence of only clubbing in a patient usually has the same clinical significance as HOA.

Pathology and Pathophysiology InHOA, the bone changes in the distal extremities begin as periostitis followed by new bone formation. At this stage, a radiolucent area may be observed between the new periosteal bone and subjacent cortex. As the process progresses, multiple layers of new bone are deposited, which become contiguous with the cortex and result in cortical thickening. The outer portion of bone is laminated in appearance, with an irregular surface. Initially, the process of periosteal new bone formation involves the proximal and distal diaphyses of the tibia, fibula, radius, and ulna and, less frequently, the femur, humerus, metacarpals, metatarsals, and phalanges. Occasionally, scapulae, clavicles, ribs, and pelvic bones are also affected. In long-standing disease, these changes extend to involve metaphyses and musculotendinous insertions. The adjacent interosseous membranes may become ossified. The distribution of the bone manifestations is usually bilateral and symmetric. The soft tissue overlying the distal third of the arms and legs may be thickened. Mononuclear cell infiltration may be present in the adjacent soft tissue. Proliferation of connective tissue occurs in the nail bed and volar pad of digits, giving the distal phalanges a clubbed appearance. Small blood vessels in the clubbed digits are dilated and have thickened walls. In addition, the number of arteriovenous anastomoses is increased. The synovium of involved joints shows edema, varying degrees of synovial cell proliferation, thickening of the subsynovium, vascular congestion, vascular obliteration with thrombi, and small numbers of lymphocyte infiltrates.

Several theories have been suggested for the pathogenesis ofHOA. Most have either been disproved or have not explained the development in all clinical disorders associated with HOA. Previously proposed neurogenic and humoral theories are no longer considered likely explanations for HOA. The neurogenic theory was based on the observation that vagotomy resulted in symptomatic improvement in a small number of patients with lung tumors and HOA. It was postulated that vagal stimuli from the tumor site led via a neural reflex to efferent nerve impulses to the distal extremities, resulting in HOA. This theory, however, did not explain HOA in conditions where vagal stimulation did not occur, as in cyanotic congenital heart disease or arterial aneurysms. The humoral theory postulated that soluble substances that are normally inactivated or removed during passage through the lung reached the systemic circulation in an active form and stimulated the changes of HOA. Substances proposed included prostaglandins, ferritin, bradykinin, estrogen, and growth hormone. These substances seemed unlikely candidates, since their blood levels in HOA patients overlapped those in individuals without HOA. Furthermore, these substances did not explain the development of localized HOA associated with arterial aneurysms or infected arterial grafts.

Recent studies have suggested a role for platelets in the development ofHOA. It has been observed that megakaryocytes and large platelet particles, present in venous circulation, were fragmented in their passage through normal lung. In patients with cyanotic congenital heart disease and in other disorders associated with right-to-left shunts, these large platelet particles may bypass the lung and reach the distal extremities, where they can interact with endothelial cells. Platelet clumps have been demonstrated to form on an infected heart valve in bacterial endocarditis, in the wall of arterial aneurysms, and on infected arterial grafts. These platelet particles may also reach the distal extremities and interact with endothelial cells. Platelet-endothelial activation in the distal portion of extremities would then result in the release of platelet-derived growth factor (PDGF) and other factors leading to the proliferation of connective tissue and periosteum. Stimulation of fibroblasts by PDGF and transforming growth factor b(TGF-b) results in cell growth and collagen synthesis. Elevated plasma levels of von Willebrand factor antigen have been found in patients with both primary and secondary forms of HOA, indicating endothelial activation or damage. Abnormalities of collagen synthesis have been demonstrated in the involved skin of patients with primary HOA. Fibroblasts from affected skin were shown to have increased collagen synthesis, increaseda1(I) procollagen mRNA, and evidence for upregulation of collagen transcription. Other factors are undoubtedly involved in the pathogenesis of HOA, and further studies are needed to better understand this disorder.

Clinical Manifestations PrimaryHOA, also referred to as *pachydermoperiostitis* or *Touraine-Solente-Gole syndrome*, usually begins insidiously at puberty. In a smaller number of patients, the onset is in the first year of life. The disorder is inherited as an autosomal dominant trait with variable expression and is nine times more common in boys than in girls. Approximately one-third of patients have a family history of primary HOA.

PrimaryHOA is characterized by clubbing, periostitis, and unusual skin features. A small number of patients with this syndrome do not express clubbing. The skin changes and periostitis are prominent features of this syndrome. The skin becomes thickened and coarse. Deep nasolabial folds develop, and the forehead may become furrowed. Patients may have heavy-appearing eyelids and ptosis. The skin is often greasy, and there may be excessive sweating of the hands and feet. Patients may also experience acne vulgaris, seborrhea, and folliculitis. In a few patients, the skin over the scalp becomes very thick and corrugated, a feature that has been descriptively termed *cutis verticis gyrata*. The distal extremities, particularly the legs, become thickened owing to proliferation of new bone and soft tissue; when the process is extensive, the distal lower extremities resemble those of an elephant. The periostitis is usually not painful, as it may be in secondary HOA. Clubbing of the fingers may be extensive, producing large, bulbous deformities and clumsiness. Clubbing also affects the toes. Patients may experience articular and periarticular pain, especially in the ankles and knees, and joint motion may be mildly restricted owing to periarticular bone overgrowth. Noninflammatory effusions occur in the wrists, knees, and ankles. Synovial hypertrophy is not found. Associated abnormalities observed in patients with primary HOA include hypertrophic gastropathy, bone marrow failure, female escutcheon, gynecomastia, and cranial suture defects. In patients with primary HOA, the symptoms disappear when adulthood is reached.

HOAsecondary to an underlying disease occurs more frequently than primary HOA. It accompanies a variety of disorders and may precede clinical features of the associated disorder by months. Clubbing is more frequent than the full syndrome of HOA in patients with associated illnesses. Because clubbing evolves over months and is usually asymptomatic, it is often recognized first by the physician and not the patient. Patients may experience a burning sensation in their fingertips. Clubbing is characterized by widening of the fingertips, enlargement of the distal volar pad, convexity of the nail contour, and the loss of the normal 15° angle between the proximal nail and cuticle. The thickness of the digit at the base of the nail is greater than the thickness at the distal interphalangeal joint. An objective measurement of finger clubbing can be made by determining the diameter at the base of the nail and at the distal interphalangeal joint of all 10 digits. Clubbing is present when the sum of the individual digit ratios is >10. At the bedside, clubbing can be appreciated by having the patient place the dorsal surface of the fourth fingers together. Normally, an open area is visible between the opposing fingers; when clubbing is present, this open space is no longer visible. The base of the nail feels spongy when compressed, and the nail can be easily rocked on its bed. Marked periungual erythema is usually present. When clubbing is advanced, the finger may have a drumstick appearance, and the distal interphalangeal joint can be hyperextended. Periosteal involvement in the distal extremities may produce a burning or deep-seated aching pain. The pain can be quite incapacitating and is aggravated by dependency and relieved by elevation of the affected limbs. The overlying soft tissue may be swollen, and the skin slightly erythematous. Pressure applied over the distal forearms and legs may be quite painful.

Patients may also experience joint pain, most often in the ankles, wrists, and knees. Joint effusions may be present; usually they are small and noninflammatory. The small joints of the hands are rarely affected. Severe joint or bone pain may be the presenting symptom of an underlying lung malignancy and may precede the appearance of clubbing. In addition, the progression ofHOAtends to be more rapid when associated with malignancies, most notably bronchogenic carcinoma. Unlike primary HOA, excessive sweating and oiliness of the skin and thickening of the facial skin are uncommon in secondary HOA.

HOAoccurs in 5 to 10% of patients with intrathoracic malignancies, the most common being bronchogenic carcinoma and pleural tumors (Table 325-3). Lung metastases infrequently cause HOA. HOA is also seen in patients with intrathoracic infections, including lung abscesses, empyema, bronchiectasis, chronic obstructive lung disease, and, uncommonly, pulmonary tuberculosis. HOA may also accompany chronic interstitial pneumonitis, sarcoidosis, and cystic fibrosis. In the latter, clubbing is more common than the full syndrome of HOA. Other causes of clubbing include congenital heart disease with right-to-left shunts, bacterial endocarditis, Crohn's disease, ulcerative colitis, sprue, and neoplasms of the esophagus, liver, and small and large bowel. In patients with congenital heart disease with right-to-left shunts, clubbing alone occurs more often than the full syndrome of HOA.

Unilateral clubbing has been found in association with aneurysms of major extremity arteries, infected arterial grafts, and with arteriovenous fistulas of brachial vessels. Clubbing of the toes but not fingers has been associated with an infected abdominal aortic aneurysm and patent ductus arteriosus. Clubbing of a single digit may follow trauma and has been reported in tophaceous gout and sarcoidosis. While clubbing occurs more commonly than the full syndrome in most diseases, periostitis in the absence of clubbing has been observed in the affected limb of patients with infected arterial grafts.

Hyperthyroidism (Graves' disease), treated or untreated, is occasionally associated with clubbing and periostitis of the bones of the hands and feet. This condition is referred to as *thyroid acropachy*. Periostitis is asymptomatic and occurs in the midshaft and diaphyseal portion of the metacarpal and phalangeal bones. The long bones of the extremities are seldom affected. Elevated levels of long-acting thyroid stimulator (LATS) are found in the serum of these patients.

Laboratory Findings The laboratory abnormalities reflect the underlying disorder. The synovial fluid of involved joints has <500 white cells per microliter, and the cells are predominantly mononuclear. Radiographs show a faint radiolucent line beneath the new periosteal bone along the shaft of long bones at their distal end. These changes are observed most frequently at the ankles, wrists, and knees. The ends of the distal phalanges may show osseous resorption. Radionuclide studies show pericortical linear uptake along the cortical margins of long bones that may be present before any radiographic changes.

TREATMENT

The treatment ofHOA is to identify the associated disorder and treat it appropriately. The symptoms and signs of HOA may disappear completely with removal or effective chemotherapy of a tumor or with antibiotic therapy and drainage of a chronic pulmonary infection. Vagotomy or percutaneous block of the vagus nerve leads to symptomatic relief in some patients. Aspirin, other nonsteroidal anti-inflammatory drugs (NSAIDs), or analgesics may help control symptoms of HOA.

FIBROMYALGIA

Fibromyalgia is a commonly encountered disorder characterized by widespread musculoskeletal pain, stiffness, paresthesia, nonrestorative sleep, and easy fatigability along with multiple tender points which are widely and symmetrically distributed. Fibromyalgia affects predominantly women in a ratio of 8 or 9 to 1 compared to men. This disorder is found in most countries, in most ethnic groups, and in all types of climates. The prevalence of fibromyalgia in the general population of a community in the United States using the 1990 American College of Rheumatology (ACR) classification criteria was reported to be 3.4%in women and 0.5% in men. Contrary to some previous reports, fibromyalgia was not found to be present mainly in young women but, rather, to be most prevalent in women ³50 years. The prevalence increased with age, being 7.4% in women between the ages of 70 and 79. Although not common, fibromyalgia also occurs in children. The reported prevalence of fibromyalgia in some rheumatology clinics has been as high as 20%.

Pathogenesis Several causative mechanisms for fibromyalgia have been postulated. Disturbed sleep has been implicated as a factor in the pathogenesis. Nonrestorative sleep or awakening unrefreshed has been observed in most patients with fibromyalgia. Sleep electroencephalographic studies in patients with fibromyalgia have shown disruption of normal stage 4 sleep [non-rapid eye movement (NREM) sleep] by many repeated a-wave intrusions. The idea that stage 4 sleep deprivation has a role in causing this disorder was supported by the observation that symptoms of fibromyalgia developed in normal subjects whose stage 4 sleep was disrupted artificially by induceda-wave intrusions. This sleep disturbance, however, has been demonstrated in healthy individuals; in emotionally distressed individuals; and in patients with sleep apnea, fever, osteoarthritis, or rheumatoid arthritis. Low levels of serotonin metabolites have been reported in the cerebrospinal fluid of patients with fibromyalgia, suggesting that a deficiency of serotonin, a neurotransmitter that regulates pain and NREM sleep, might also be involved in the pathogenesis of fibromyalgia. Drugs that affect serotonin metabolism have not had a dramatic effect on fibromyalgia, however. Since patients experience pain from muscle and musculotendinous sites, many studies have been done to examine muscle, both structurally and physiologically. Inflammation or diagnostic muscle abnormalities have not been found. Evidence indicates deconditioning of muscles, and patients experience a greater degree of postexertional pain than do unaffected persons. Fibromyalgia patients as a group have been reported by some investigators to have reduced levels of growth hormone, which is important for muscle repair and strength. Growth hormone is secreted normally during stage 4 sleep, which is disturbed in patients with fibromyalgia. The reduction of growth hormone may explain the extended periods of muscle pain following exertion in these patients. The level of the neurotransmitter substance P has been reported to be increased in the cerebrospinal fluid of fibromyalgia patients and may play a role in spreading muscle pain. Patients with fibromyalgia have a decreased cortisol response to stress. Low urinary free cortisol and a diminished cortisol response to corticotropin-releasing hormone suggest an abnormal hypothalamic-pituitary-adrenal axis. Disturbances of the autonomic and peripheral nervous systems may account for the cold sensitivity and Raynaud's-like symptoms seen in patients with fibromyalgia.

Many patients with fibromyalgia have psychological abnormalities; there has been disagreement as to whether some of these abnormalities represent reactions to the chronic pain or whether the symptoms of fibromyalgia are a reflection of psychiatric disturbance. Many patients fit a psychiatric diagnosis, the most common being depression, anxiety, somatization, and hypochondriases. Studies have also shown a high prevalence of sexual and physical abuse and eating disorders. However, fibromyalgia also occurs in patients without significant psychiatric problems. Patients with fibromyalgia may have a lower pain threshold than usual, although not all investigators in the field agree on this point. A better understanding of fibromyalgia awaits further studies.

Clinical Manifestations Symptoms are generalized aching and stiffness of the trunk, hip, and shoulder girdles. Other patients complain of generalized muscle aching and weakness. Patients may complain of low back pain, which may radiate into the buttocks and legs. Others complain of pain and tightness in the neck and across the upper posterior shoulders. Patients complain of muscle pain after even mild exertion. Some degree of pain is always present. The pain has been described as a burning or gnawing pain or as soreness, stiffness, or aching. While pain may begin in one region, such as the shoulders, neck, or lower back, it eventually becomes widespread. Patients may complain of joint pain and perceive that their joints are swollen; however, joint examination yields normal findings. Stiffness is usually present on arising in the morning; usually it improves during the day, but in some patients it lasts all day. Patients may complain of numbness of their hands and feet. They may also feel colder overall than others in the home, and some may experience Raynaud's-like phenomena or actual Raynaud's phenomenon. Patients complain of feeling fatigued and exhausted and wake up tired. They also awaken frequently at night and have trouble falling back to sleep. Symptoms are made worse by stress or anxiety, cold, damp weather, and overexertion. Patients often feel better during warmer weather and vacations.

The characteristic feature on physical examination is the demonstration of specific tender points, which are exclusively more tender or painful than adjacent areas. TheACRCriteria for Fibromyalgia defines 18 tender points (Fig. 325-2). These points of tenderness are remarkably constant in location. A moderate degree of pressure should be used in digital palpation of these tender points. Some workers recommend that the tender site be palpated using a rolling motion, which may be more effective in eliciting the tenderness. The tender sites can also be examined using a dolorimeter, which is a spring-loaded pressure gauge. Digital palpation appears to be as effective and accurate for the diagnosis of fibromyalgia as dolorimetry. The amount of pressure applied by the examiner introduces variability in the interpretation, however. If too much pressure is applied, the pain will be produced even in normal subjects. Likewise, tenderness will not be appreciated if too little pressure is applied or the site is missed on palpation. Some investigators have quantitated their response, but the number of tender point sites is more diagnostic. Some patients are tender all over and not just at the specific tender point sites. These patients are still more tender over the specific tender point sites, however. Sites where there is usually no tenderness and which can be used as controls are the dorsum of the third digit between the proximal interphalangeal and distal interphalangeal joints, the medial third of the clavicle, the medial malleolus, and the forehead. If tenderness at these sites is also present, the diagnosis of fibromyalgia should be questioned and possible psychiatric disorders investigated. Whether such patients can be diagnosed as also having fibromyalgia is debatable.

Skinfold tenderness may be present, particularly over the upper scapular region. Subcutaneous nodules may be felt at sites of tenderness. Nodules in similar locations are present in normal persons but are not tender.

Fibromyalgia may be triggered by emotional stress, medical illness, surgery, hypothyroidism, and trauma. It has appeared in some patients with human immunodeficiency virus (HIV) infection, parvovirus B19 infection, or Lyme disease. In the latter situation, fibromyalgia persisted despite adequate antibiotic treatment for Lyme disease. Disorders commonly associated with fibromyalgia include irritable bowel syndrome, irritable bladder, headaches (including migraine headaches), dysmenorrhea, premenstrual syndrome, restless legs syndrome, temporomandibular joint pain, and sicca syndrome.

The course of fibromyalgia is variable. Symptoms wax and wane in some patients, while in others pain and fatigue are persistent regardless of therapy. Studies from tertiary medical centers indicate a poor prognosis for most patients. The prognosis may be better in community-treated patients. In a community-based study reported after 2 years of treatment, 24% of patients were in remission, and 47% no longer fulfilled theACRcriteria for fibromyalgia.

Diagnosis Fibromyalgia is diagnosed by a history of widespread pain and the demonstration of at least 11 of the 18 tender point sites on digital palpation (Fig. 325-2). TheACRcriteria are useful for standardizing the diagnosis; however, not all patients with fibromyalgia meet these criteria (Table 325-4). Some patients have fewer tender sites and more regional pain and may be considered to have probable fibromyalgia.

Results of joint and muscle examinations are normal in fibromyalgia patients, and there are no laboratory abnormalities. Fibromyalgia may occur in patients with rheumatoid arthritis, other connective tissue diseases, or other medical illness. A distinction is no longer made between primary and secondary fibromyalgia (concomitant with other disease), as the signs and symptoms are similar. Fibromyalgia and chronic fatigue syndrome have many similarities (Chap. 384). Both are associated with fatigue, abnormal sleep, musculoskeletal pain, and psychiatric conditions such as less severe forms of depression and anxiety. Patients with chronic fatigue syndrome, however, are more likely to have symptoms suggesting a viral illness. These include mild fever, sore throat, and pain in the axillary and anterior and posterior cervical lymph nodes. The onset of chronic fatigue syndrome is usually sudden; patients are usually able to date the onset. Patients also have impaired memory and concentration. While many patients with chronic fatigue syndrome have tender points, the diagnosis does not require their presence. Polymyalgia rheumatica is distinguished from fibromyalgia in an elderly patient by the presence of more proximal muscle stiffness and pain and an elevated erythrocyte sedimentation rate. Patients should be evaluated for hypothyroidism, which may have symptoms similar to fibromyalgia or may accompany fibromyalgia.

The diagnosis of fibromyalgia has taken on a more complex significance in regard to labor and industry issues. This has become a significant issue since it has been reported that 10 to 25% of patients are not able to work in any capacity, while others require modification of their work. Disability evaluation in fibromyalgia is controversial. The diagnosis of fibromyalgia is not accepted by all. It is hard to evaluate patients' perceptions of their inability to function. The determination of tender points can also be subjective, on the part of both the physician and the patient, particularly when issues of compensation are pending. Patients also encounter difficulty in having their illness recognized as a disability. Physicians have been placed in the inappropriate role of assessing the patient's disability. Physicians are not in a position to quantitate disability at the workplace; that is better done by a work evaluation specialist. Better instruments are clearly needed for measuring disability, particularly in patients with fibromyalgia.

TREATMENT

Patients should be informed that they have a condition that is not crippling, deforming, or degenerative, and that treatment is available. Salicylates or otherNSAIDsonly partially improve symptoms. Glucocorticoids have been of little benefit and should not be used in these patients. Opiate analgesics should be avoided. Local measures such as heat, massage, injection of tender sites with steroids or lidocaine, and acupuncture provide only temporary relief of symptoms. Other therapies that may help to varying degrees including biofeedback, behavioral modification, hypnotherapy, and stress management and relaxation response training. The use of tricyclics such as amitriptyline (10 to 50 mg) and doxepin (10 to 25 mg) or a pharmacologically similar drug, cyclobenzaprine (10 to 40 mg), 1 to 2 h before bedtime will give the patient restorative sleep (stage 4 sleep), resulting in clinical improvement. Patients should be started on a low dose, which is increased gradually as needed. Side effects of these tricyclics and cyclobenzoprine limit their use; these include constipation, dry mouth, weight gain, drowsiness, and difficulty thinking. Depression and anxiety should be treated with appropriate drugs and, when indicated, with psychiatric counseling. Alprazolam and lorazepam can be used for anxiety, while trazodone, sertraline, fluoxetine, paroxetine or other newer selective serotonin reuptake inhibitors can be used as antidepressants. Patients may also benefit by regular aerobic exercises. Exercise should be of a low-impact type and begun at a low level. Eventually, the patient should be exercising 20 to 30 min 3 to 4 days a week. Regular stretching exercises are also very important. Life stresses should be identified and discussed with the patient, and the patient should be provided with help on how to cope with these stresses. Patients may benefit from a multidisciplinary team approach involving a mental health professional, a physical therapist, and a physical medicine and rehabilitation specialist. Group therapy may be beneficial. Patients should be well educated about their disorder and taught the importance of self help. There are patient support groups in many communities. While treatment of fibromyalgia is effective in some patients, others continue to have chronic disease, which is relieved only partially if at all.

MYOFASCIAL PAIN SYNDROME

Myofascial pain syndrome is characterized by localized musculoskeletal pain and tenderness in association with trigger points. The pain is deep and aching and may be accompanied by a burning sensation. Myofascial pain may follow trauma, overuse, or prolonged static contraction of a muscle or muscle group, which may occur when reading or writing at a desk or working at a computer. In addition, this syndrome may be associated with underlying osteoarthritis of the neck or low back. Trigger points are a diagnostic feature of this syndrome. Pain is referred from trigger points to defined areas distant from the original tender points. Palpation of the trigger point reproduces or accentuates the pain. The trigger points are usually located in the center of a muscle belly, but they can occur at other sites, such as costosternal junctions, the xyphoid process, ligamentous and tendinous insertions, fascia, and fatty areas. Trigger point sites in muscle have been described as feeling indurated and taut, and palpation may cause the muscle to twitch. These findings, however, have been shown not to be unique for myofascial pain syndrome, since in a controlled study they were also present in fibromyalgia patients and normal subjects. Myofascial pain most often involves the posterior neck, low back, shoulders, and chest. Chronic pain in the muscles of the posterior neck may involve referral of pain from the trigger point in the erector neck muscle or upper trapezius to the head, leading to persistent headaches which may last for days. Trigger points in the paraspinal muscles of the low back may refer pain to the buttock. Pain may be referred down the leg from a trigger point in the gluteus medius and can mimic sciatica. A trigger point in the infraspinatus muscle may produce local and referred pain over the lateral deltoid and down the outside of the arm into the hand. Injection of a local anesthetic such as 1% lidocaine into the trigger point site often results in pain relief. Another useful technique is first to spray from the trigger point toward the area of referred pain with an agent such as ethyl chloride and then to stretch the muscle. This maneuver may need to be repeated several times. Massage and application of ultrasound to the affected area may also be beneficial. Patients should be instructed in methods to prevent muscle stresses related to work and recreation. Posture and resting positions are important in preventing muscle tension. The prognosis in most patients is good. In some patients, myofascial pain syndrome may evolve into fibromyalgia. Patients at risk for developing fibromyalgia are thought to be those with anxiety, depression, nonrestorative sleep, and fatigue.

PSYCHOGENIC RHEUMATISM

Patients may experience severe joint pain involving a few to several joints without physical findings of arthritis. These patients are often convinced that they have rheumatoid arthritis,SLE, or another connective tissue disease. This disorder is recognized by the inconsistencies, exaggerations, and emotional lability of the patient during the history and physical examination. Results of laboratory studies are normal. Organic disease needs to be excluded, which necessitates seeing the patient at regular intervals. This condition also needs to be distinguished from fibromyalgia. Anti-inflammatory or other drugs are not helpful.

REFLEX SYMPATHETIC DYSTROPHY SYNDROME

The reflex sympathetic dystrophy syndrome (RSDS) is now referred to as *complex* *regional pain syndrome, type 1*, by the new Classification of the International Association for the Study of Pain. It is characterized by pain and swelling, usually of a distal extremity, accompanied by vasomotor instability, trophic skin changes, and the rapid development of bony demineralization. RSDS occasionally involves an isolated site such as a knee, hip, or one or two digits of a foot or hand. The contralateral side is affected clinically in ~25% of patients and may be involved in virtually all patients with RSDS, as shown by scintigraphic studies. A precipitating event can be identified in at least two-thirds of cases. These events include trauma, such as fractures and crush injuries; myocardial infarction; strokes; peripheral nerve injury; and use of certain drugs, including barbiturates, anti-tuberculous drugs, and, more recently, cyclosporine administered to patients undergoing renal transplantation. The pathogenesis of RSDS is poorly understood and is thought to involve abnormal activity of the sympathetic nervous system following a precipitating event.

RSDSevolves through three clinical phases. The first phase is characterized by an intense burning pain and swelling of a distal extremity. The involved extremity is warm, edematous, and very tender, especially around joints. Sweating and hair growth are increased. Light touch causes pain, which may continue after the stimulus is removed. Passive or active motion of joints is very painful, and the joints are stiff. In the first phase, especially when both sides are involved, the clinical findings may suggest early rheumatoid arthritis. Redness and swelling over a distal extremity such as an ankle or wrist may also mimic inflammatory arthritis, or even an infectious arthritis. In 3 to 6 months, the skin gradually becomes thin, shiny, and cool. This is the second phase of the disease. The clinical features of the first and second phases often overlap. In another 3 to 6 months (third phase), the skin becomes atrophic and dry, and irreversible flexion contractures, palmar fibromatosis, and Dupuytren's contractures develop, resulting in a clawlike hand deformity. Similar changes occur in the feet. When RSDS occurs in the upper extremity, motion of the shoulder on the affected side may be painful and restricted, a condition referred to as *shoulder-hand syndrome* (see "Adhesive Capsulitis," below).**Reflex sympathetic dystrophy syndrome, including its treatment, is covered in greater detail in Chap. 366.*

TIETZE'S SYNDROME AND COSTOCHONDRITIS

Tietze's syndrome is manifested by painful swelling of one or more costochondral articulations. The age of onset is usually before 40, and both sexes are affected equally. In most patients only one joint is involved, usually the second or third costochondral joint. The onset of anterior chest pain may be sudden or gradual. The pain may radiate to the arms or shoulder and is aggravated by sneezing, coughing, deep inspirations, or twisting motions of the chest. The term *costochondritis* is often used interchangeably with *Tietze's syndrome*, but some workers restrict the former term to pain of the costochondral articulations without swelling. Costochondritis is observed in patients over age 40; tends to affect the third, fourth, and fifth costochondral joints; and occurs more often in women. Both syndromes may mimic cardiac or upper abdominal causes of pain. Rheumatoid arthritis, ankylosing spondylitis, and Reiter's syndrome may involve costochondral joints but are distinguished easily by their other clinical features. Other skeletal causes of anterior chest wall pain are xiphoidalgia and the slipping rib syndrome, which usually involves the tenth rib. Malignancies such as breast cancer, prostate cancer, plasma cell cytoma, and sarcoma can invade the ribs, thoracic spine, or chest wall and produce symptoms suggesting Tietze's syndrome. They should be easily distinguishable by radiographs and biopsy. Analgesics, anti-inflammatory drugs, and local glucocorticoid injections usually relieve symptoms.

MUSCULOSKELETAL DISORDERS ASSOCIATED WITH HYPERLIPIDEMIA (See alsoChap. 344)

Musculoskeletal manifestations may be the first indication of a hereditary disorder of lipoprotein metabolism. Patients with familial hypercholesterolemia (previously referred to as type II hyperlipoproteinemia) may have recurrent migratory polyarthritis involving knees and other large peripheral joints and, to a lesser degree, peripheral small joints. In a few patients, the arthritis is monarticular. Fever may accompany the arthritis. Pain ranges from moderate to very severe to incapacitating. The involved joints can be warm, erythematous, swollen, and tender. Arthritis usually has a sudden onset, lasts from a few days to 2 weeks, and does not cause joint damage. Episodes may suggest acute gout attacks. Several attacks occur a year. Synovial fluid from involved joints is not inflammatory and contains few white cells and no crystals. Joint involvement may actually represent inflammatory periarthritis or peritendinitis and not intraarticular disease. The recurrent, transient nature of the arthritis may suggest rheumatic fever, especially since patients with lipoproteinemia have an elevated erythrocyte sedimentation rate, and a falsely elevated antistreptolysin O titer. Patients may also experience Achilles tendinitis, which can be very painful. Attacks of tendinitis come on gradually and last only a few days. Fever is not present. Patients may be asymptomatic between attacks. During an attack the Achilles tendon is warm, erythematous, swollen, and tender to palpation. Achilles tendinitis and other joint manifestations often precede the appearance of xanthomas and may be the first clinical indication of hyperlipoproteinemia. Attacks of tendinitis may occur following treatment with a lipid-lowering drug. Patients also have tendinous xanthomas in the Achilles, patellar, and extensor tendons of the hands over the knuckles and feet. Xanthomas have also been reported in the peroneal tendon, the plantar aponeurosis, and the periosteum overlying the distal tibia. These xanthomas are located within tendon fibers. Tuberous xanthomas are soft subcutaneous masses located over the extensor surfaces of the elbows, knees, and hands, as well as on the buttocks. They appear in childhood in homozygous patients and after the age of 30 in heterozygous patients. Patients with elevated plasma levels of very low density lipoprotein (VLDL) and triglyceride (previously referred to as type IV hyperlipoproteinemia) may also have a mild inflammatory arthritis affecting large and small peripheral joints, usually in an asymmetric pattern with only a few joints involved at a time. The onset of arthritis is usually in middle age. Arthritis may be persistent or recurrent, with episodes lasting a few days to weeks. Joint pain is severe in some patients. Patients may experience morning stiffness. Joint tenderness and periarticular hyperesthesia may also be present, as may synovial thickening. Joint fluid is usually noninflammatory and without crystals, but may have increased white blood cell counts with predominantly mononuclear cells. The fluid is occasionally lactescent. Radiographs may show juxtaarticular osteopenia and cystic lesions. Large bone cysts have been noted in a few patients. Xanthoma and bone cysts are also observed in other lipoprotein disorders. The pathogenesis of arthritis in patients with familial hypercholesterolemia or with elevated levels of VLDL and triglyceride is not well understood. Salicylates, otherNSAIDs, or analgesics usually provide relief of symptoms. Clinical improvement also may occur in patients treated with lipid lowering agents. Patients, however, treated with a HMG CoA reductase agent may experience myalgias and a few patients may develop polymysitis or even rhabdomyolysis (Chap. 382).

ARTHROPATHY OF ACROMEGALY

Acromegaly is the result of excessive production of growth hormone by an adenoma in the anterior pituitary gland (Chap. 328). Middle-aged persons are most often affected. The excessive secretion of growth hormone along with insulin-like growth factor I stimulates proliferation of cartilage, periarticular connective tissue, and bone, resulting in several musculoskeletal abnormalities, including osteoarthritis, back pain, muscle weakness, and carpal tunnel syndrome.

An arthropathy resembling osteoarthritis is a common feature, affecting most often the knees, shoulders, hips, and hands. Single or multiple joints may be affected. The overgrowth of cartilage initially produces widening of the joint space. The newly synthesized cartilage is not developed in an organized manner, making it susceptible to fissuring, ulceration, and destruction. Ligamental laxity of the joint resulting from the growth of connective tissue also contributes to the development of osteoarthritis. With breakdown and loss of cartilage, the joint space narrows, and subchondral sclerosis and osteophytes appear on radiographs. Joint examination reveals marked crepitus and hypermobility. Joint fluid is noninflammatory. Calcium pyrophosphate dihydrate crystals are found in the cartilage in some cases of acromegaly arthropathy and, when shed into the joint, can produce attacks of pseudogout. Chondrocalcinosis may also be observed radiographically. Approximately half of the patients with acromegaly experience back pain, which is predominantly lumbosacral. Hypermobility of the spine may be a contributing factor in back pain. Radiograph of the spine shows normal or increased intervertebral disk spaces, hypertrophic anterior osteophytes, and ligamental calcification. These changes are similar to those observed in patients with diffuse idiopathic skeletal hyperostosis. Dorsal kyphosis in conjunction with elongation of the ribs contributes to the development of the barrel chest seen in acromegalic patients. The hands and feet become enlarged owing to soft tissue proliferation. The fingers are thickened and have spadelike distal tufts. One-third of patients have a thickened heel pad. Approximately 25% of patients have Raynaud's phenomenon.

Carpal tunnel syndrome occurs in about half of patients. The median nerve is compressed by the excessive growth of connective tissue in the carpal tunnel. The median nerve also becomes enlarged. Patients with acromegaly also develop proximal muscle weakness, which is thought to be caused by the effect of growth hormone on muscle. Results of muscle enzyme assays and electromyography are normal. Muscle biopsy specimens show muscle fibers of varying size and no inflammatory changes.

ARTHROPATHY OF HEMOCHROMATOSIS

Hemochromatosis is a disorder of iron storage. Excessive amounts of iron are absorbed from the intestine, leading to iron deposition in parenchymal cells, which results in tissue damage and impairment of organ function (Chap. 345). Symptoms of hemochromatosis usually begin between the ages of 40 and 60 but can occur earlier. Arthritis, which occurs in 20 to 40% of patients, usually begins after the age of 50 and may be the first clinical feature of hemochromatosis. The arthropathy is an inflammatory osteoarthritis-like disorder affecting the small joints of the hands, followed later by larger joints such as knees, ankles, shoulders, and hips. The second and third metacarpophalangeal joints of both hands are often the first joints affected; they can provide an important clue to the possibility of hemochromatosis. Patients experience stiffness and pain. Morning stiffness usually lasts less than half an hour. The affected joints are enlarged and mildly tender. Synovial tissue is not appreciatively increased. Radiographs show irregular narrowing of the joint space, subchondral sclerosis, and subchondral cysts. There is juxtaarticular proliferation of bone, with frequent hooklike osteophytes. The synovial fluid is noninflammatory. The synovium shows mild to moderate proliferation of lining cells, fibrosis, and a low number of inflammatory cells, which are mononuclear. In approximately half of patients, there is evidence of calcium pyrophosphate deposition disease. Iron can be demonstrated in the lining cells of the synovium and also in chondrocytes.

Iron may damage the articular cartilage in several ways. Promotion by iron of superoxide-dependent lipid peroxidation may play a role in joint damage. In animal models, ferric iron has been shown to interfere with collagen formation. Iron has also been shown to increase the release of lysosomal enzymes from cells in the synovial membrane. Iron may also play a role in the development of chondrocalcinosis. Iron inhibits synovial tissue pyrophosphatase in vitro and, therefore, may inhibit pyrophosphatase in vivo, resulting in chondrocalcinosis. Iron in synovial cells may also inhibit the clearance of calcium pyrophosphate from the joint.

TREATMENT

The treatment of hemochromatosis is repeated phlebotomy. Unfortunately, this treatment has little effect on the arthritis, which, along with chondrocalcinosis, usually continues to progress. Treatment of the arthritis consists of administration of acetaminophen andNSAIDs. Placement of a hip or knee prosthesis has been successful in advanced disease.

HEMOPHILIC ARTHROPATHY

Hemophilia is a sex-linked recessive genetic disorder characterized by the absence or deficiency of factor VIII (hemophilia A, or classic hemophilia) or factor IX (hemophilia B, or Christmas disease) (Chap. 117). Hemophilia A is by far the more common type, constituting 85% of cases. Spontaneous hemarthrosis is a common problem with both types of hemophilia and can lead to a chronic deforming arthritis. The frequency and severity of hemarthrosis are related to the degree of clotting factor deficiency. Hemarthrosis is not common in other inherited disorders of coagulation, such as von Willebrand's disease or factor V deficiency.

Hemarthrosis becomes evident after 1 year of age, when the child begins to walk and run. In order of frequency, the joints most commonly affected are the knees, ankles, elbows, shoulders, and hips. Small joints of the hands and feet are occasionally involved.

In the initial stage of arthropathy, hemarthrosis produces a warm, tensely swollen, and painful joint. The patient holds the affected joint in flexion and guards against any movement. Blood in the joint remains liquid because of the absence of intrinsic clotting factors and the absence of tissue thromboplastin in the synovium. The blood in the joint space is resorbed over a period of a week or longer, depending on the size of the hemarthrosis. Joint function usually returns to normal or baseline in about 2 weeks.

Recurrent hemarthrosis leads to the development of a chronic arthritis. The involved joints remain swollen, and flexion deformities develop. In the later stages of arthropathy, joint motion is restricted and function is severely limited. Joint ankylosis, subluxation, or laxity are features of end-stage disease.

Bleeding into muscle and soft tissue also causes musculoskeletal disorders. When bleeding into the iliopsoas muscle occurs, the hip is held in flexion because of the pain, resulting in a hip flexion contracture. Rotation of the hip is preserved, which distinguishes this problem from intraarticular hemorrhage. Expansion of the hematoma may place pressure on the femoral nerve, resulting in a femoral neuropathy. Another problem is shortening of the heel cord secondary to bleeding into the gastrocnemius. Hemorrhage into a closed compartment space, such as the volar compartment in the forearm, can result in muscle necrosis and flexion deformities of the wrist and fingers. When bleeding involves periosteum or bone, a pseudotumor forms. These occur distal to the elbows or knees in children and improve with treatment of the hemophilia. Surgical removal is indicated if the pseudotumor continues to enlarge. In adults, they occur in the femur and pelvis and are usually refractory to treatment. When bleeding occurs in muscle, cysts may develop within the muscle. Needle aspiration of a cyst is contraindicated because it can induce bleeding.

Septic arthritis can occur in hemophilia and is difficult at times to distinguish from acute hemarthrosis. Whenever there is suspicion of an infected joint, the joint should be aspirated immediately, the fluid cultured, and the patient started on a broad-spectrum antibiotic. The patient should be infused with the deficient clotting factor before the joint is tapped to decrease the risk of further bleeding.

Radiographs of joints reflect the stage of disease. In early stages there is only capsule distention; later, juxtaarticular osteopenia, marginal erosions, and subchondral cysts develop. In late disease, the joint space is narrowed and there is bony overgrowth. The changes are similar to those observed in osteoarthritis. Unique features of hemophilic arthropathy are widening of the femoral intercondylar notch, enlargement of the proximal radius, and squaring of the distal end of the patella.

Recurrent hemarthrosis produces synovial hyperplasia and hypertrophy. A pannus covers the cartilage. Cartilage is damaged by collagenase and other degradative enzymes released by mononuclear cells in the overlying synovium. Hemosiderin is found in synovial lining cells, the subsynovium, and chondrocytes and may also play a role in cartilage destruction.

TREATMENT

The treatment of hemarthrosis is initiated with the immediate infusion of factor VIII or IX at the first sign of joint or muscle hemorrhage. The patient is placed at bed rest, with the involved joint in as much extension as the patient can tolerate. AnalgesicNSAIDsand local icing may help with the pain. NSAIDs can be given safely for short periods even though they have a stabilizing effect on platelets. Studies have shown no significant abnormalities in platelet function or bleeding time in hemophiliacs receiving ibuprofen. The new cyclooxygenase-2 inhibitors celecoxib and rofecoxib do not interfere with platelet function and can be safely given for pain. Synovectomy, open or arthroscopic, may be indicated in patients with chronic synovial proliferation and recurrent hemarthrosis. Hypertrophied synovium is very vascular and subject to bleeding. Both types of synovectomy reduce the number of hemarthroses and slow the roentgenographic progression of hemophilic arthropathy. Open surgical synovectomy, however, is associated with some loss of range of motion. Radiosynovectomy with either yttrium 90 silicate or phosphorus 31 colloid has also been effective and may be a useful alternative when surgical synovectomy is not practical. Total joint replacement is indicated for severe joint destruction and incapacitating pain. Because of the young age of hemophilic patients, total-joint prostheses may need to be replaced more than once during their lives.

ARTHROPATHIES ASSOCIATED WITH HEMOGLOBINOPATHIES

Sickle Cell Disease Sickle cell disease (Chap. 106) is associated with several musculoskeletal abnormalities (Table 325-5). Children under the age of 5 may develop diffuse swelling, tenderness, and warmth of the hands and feet lasting from 1 to 3 weeks. The condition, referred to as *sickle cell dactylitis* or *hand-foot syndrome* has also been observed in sickle cell disease and sickle cell thalassemia. Dactylitis is believed to result from infarction of the bone marrow and cortical bone leading to periostitis and soft tissue swelling. Radiographs show periosteal elevation, subperiosteal new bone formation, and areas of radiolucency and increased density involving the metacarpals, metatarsals, and proximal phalanges. These bone changes disappear after several months. The syndrome leaves little or no residual damage. Because hematopoiesis ceases in the small bones of hands and feet with age, the syndrome is rarely seen after age 4 or 5 and does not occur in adults.

Sickle cell crisis is often associated with periarticular pain and joint effusions. The joint and periarticular area are warm and tender. Knees and elbows are most often affected, but other joints can be involved. Joint effusions are noninflammatory, with white cell counts<1000/uL; mononuclear cells predominate. There have been a few reports of sterile inflammatory effusion with high cell counts consisting of mostly polymorphonuclear white cells. Synovial biopsies have shown mild lining cell proliferation and microvascular thrombosis. Scintigraphic studies have shown decreased marrow uptake adjacent to the involved joint. The joint effusion and periarticular pain are considered to be the result of ischemia and infarction of the synovium and adjacent bone and bone marrow. The treatment is that for sickle cell crisis (Chap. 106).

Patients with sickle cell disease may also develop osteomyelitis, which commonly involves the long tubular bones (Chap. 129). These patients are particularly susceptible to bacterial infections, especially *Salmonella* infections, which are found in more than half of cases (Chap. 156). Radiographs of the involved site show periosteal elevation initially, followed by disruption of the cortex. Treatment of the infection results in healing of the bone lesion. Sickle cell disease is also associated with bone infarction resulting from thrombosis secondary to the sickling of red cells. Bone infarction also occurs in hemoglobin S-C disease and sickle cell thalassemia (Chap. 106). The bone pain in sickle cell crisis is due to bone and bone marrow infarction. In children, infarction of the epiphyseal growth plate interferes with normal growth of the affected extremity. Radiographically, infarction of the bone cortex results in periosteal elevation and irregular thickening of the bone cortex. Infarction in the bone marrow leads to lysis, fibrosis, and new bone formation.

Avascular necrosis of the head of the femur is seen in ~5% of patients. It also occurs in the humeral head and less commonly in the distal femur, tibial condyles, distal radius, vertebral bodies, and other juxtaarticular sites. The mechanism for avascular necrosis is most likely the same as for bone infarction. Subchondral hemorrhage may play a role in the deterioration of articular cartilage. Irregularity of the femoral head or of other bone surfaces affected by avascular necrosis eventually results in degenerative joint disease. Radiograph of the affected joint may show patchy radiolucency and density followed by flattening of the bone. Magnetic resonance imaging is a sensitive technique for detecting early avascular necrosis as well as bone infarction elsewhere. Total hip replacement and placement of prostheses in other joints may improve function and relieve pain in those patients with severe joint destruction.

Septic arthritis is occasionally encountered in sickle cell disease (Chap. 323). Multiple joints may be infected. Joint infection may result from hematogenous spread or from spread of contiguous osteomyelitis. Microorganisms identified include staphylococcus, *Streptococcus*, *Escherichia coli*, and *Salmonella*. The latter is not seen as frequently in septic arthritis as it is in osteomyelitis. Acute gouty arthritis is uncommon in sickle cell disease, even though 40% of patients are hyperuremic. Hyperuricemia is due to overproduction of uric acid secondary to increased red cell turnover. Attacks may be polyarticular.

The bone marrow hyperplasia in sickle cell disease results in widening of the medullary cavities, thinning of the cortices, and coarse trabeculations and central cupping of the vertebral bodies. These changes are also seen to a lesser degree in hemoglobin S-C disease and sickle cell thalassemia. In normal individuals, red marrow is located mostly in the axial skeletal, but in sickle cell disease, red marrow is found in the bones of the extremities and even in the tarsal and carpal bones. Vertebral compression may lead to dorsal kyphosis, and softening of the bone in the acetabulum may result in protrusio acetabuli.

Thalassemia b-Thalassemia is a congenital disorder of hemoglobin synthesis characterized by impaired production of bchains (Chap. 106). Bone and joint abnormalities occur inb-thalassemia, being most common in the major and intermedia groups. In one study, approximately 50% of patients with b-thalassemia had evidence of symmetric ankle arthropathy, characterized by a dull aching pain aggravated by weight bearing. The onset was most often in the second or third decade of life. The degree of ankle pain in these patients varied. Some patients experienced self-limited ankle pain, which occurred only after strenuous physical activity and lasted several days to weeks. Other patients had chronic ankle pain, which became worse with walking. Symptoms eventually abated in a few patients. Compression of the ankle, calcaneus, or forefoot was painful in some patients. Synovial fluid from two patients was noninflammatory. Radiographs of ankle showed osteopenia, widened medullary spaces, thin cortices, and coarse trabeculations. These findings were largely the result of bone marrow expansion. The joint space was preserved. Specimens of bone from three patients revealed osteomalacia, osteopenia, and microfractures. Increased osteoblasts as well as increased foci of bone resorption were present on the bone surface. Iron staining was found in the bone trabeculae, in osteoid, and in the cement line. Synovium showed hyperplasia of lining cells which contained deposits of hemosiderin. This arthropathy was considered to be related to the underlying bone pathology. The role of iron overload or abnormal bone metabolism in the pathogenesis of this arthropathy is not known. The arthropathy was treated with analgesics and splints. Patients were also transfused to decrease hematopoiesis and bone marrow expansion.

Patients withb-thalassemia major and intermedia also have involvement of other joints, including the knees, hips, and shoulders. Acquired hemochromatosis with arthropathy has been described in a patient with thalassemia. Gouty arthritis and septic arthritis can occur. Avascular necrosis is not a feature of thalassemia because there is no sickling of red cells leading to thrombosis and infarction.

b-Thalassemia minor (trait) is also associated with joint manifestations. Chronic seronegative oligoarthritis affecting predominantly ankles, wrists, and elbows has been described. These patients had mild persistent synovitis without large effusions. Joint erosions were not seen. Recurrent episodes of an acute asymmetric arthritis also have been reported; episodes last less than a week and may affect knees, ankles, shoulders, elbows, wrists, and metacarpal phalangeal joints. The mechanism for this arthropathy is unknown. Treatment with nonsteroidal drugs was not particularly effective.

TUMORS OF JOINTS

Primary tumors and tumor-like disorders of synovium are uncommon but should be considered in the differential diagnosis of monarticular joint disease. In addition, metastases to bone and primary bone tumors adjacent to a joint may produce joint symptoms.**For further discussion, see Chap. 98.*

Pigmented villonodular synovitis is characterized by the slowly progressive, exuberant, benign proliferation of synovial tissue, usually involving a single joint. The most common age of onset is in the third decade, and women are affected slightly more often than men. The cause of this disorder is unknown.

The synovium has a brownish color and numerous large, finger-like villi that fuse to form pedunculated nodules. There is marked hyperplasia of synovial cells in the stroma of the villi. Hemosiderin granules and lipids are found in the cytoplasm of macrophages and in the interstitial tissue. Multinucleated giant cells may be present. The proliferative synovium grows into the subsynovial tissue and invades adjacent cartilage and bone.

The clinical picture of pigmented villonodular synovitis is characterized by the insidious onset of swelling and pain in one joint, most commonly the knee. Other joints affected include the hips, ankles, calcaneocuboid joints, elbows, and small joints of the fingers or toes. The disease may also involve the common flexor sheath of the hand or fingers.

Less commonly, tendon sheaths in the wrist, ankle, or foot may be involved. Symptoms may be mild and intermittent and may be present for years before the patient seeks medical attention. Radiographs may show joint space narrowing, erosions, and subchondral cysts. The joint fluid contains blood and is dark red or almost black in color. Lipid-containing macrophages may be present in the fluid. The joint fluid may be clear if hemorrhages have not occurred.

The treatment of pigmented villonodular synovitis is complete synovectomy. With incomplete synovectomy, the villonodular synovitis recurs, and the rate of tissue growth may be faster than originally. Irradiation of the involved joint has been successful in some patients.

Synovial chondromatosis is a disorder characterized by multiple focal metaplastic growths of normal-appearing cartilage in the synovium or tendon sheath. Segments of cartilage break loose and continue to grow as loose bodies. When calcification and ossification of loose bodies occur, the disorder is referred to as *synovial osteochondromatosis*. The disorder is usually monarticular and affects young to middle-aged individuals. The knee is most often involved, followed by hip, elbow, and shoulder. Symptoms are pain, swelling, and decreased motion of the joint. Radiographs may show several rounded calcifications within the joint cavity. Treatment is synovectomy; however, the tumor may recur.

Hemangiomas occur in synovium and in tendon sheaths. The knee is affected most commonly. Recurrent episodes of joint swelling and pain usually begin in childhood. The joint fluid is bloody. Treatment is excision of the lesion. *Lipomas* occur most often in the knee, originating in the subsynovial fat on either side of the patellar tendon. Lipomas also appear in tendon sheaths of the hands, wrists, feet, and ankles. In some instances, surgical removal is necessary.

Synovial sarcoma is a malignant neoplasm often found near a large joint of both upper and lower extremities, being more common in the lower extremity. It seldom arises within the joint itself. Synovial sarcomas comprise 10% of sarcomas. The tumor is believed to arise from primitive mesenchymal tissue which differentiates into epithelial cells and/or spindle cells. Small foci of calcification may be present in the tumor mass. It occurs most often in young adults and is more common in men. The tumor presents as a slowly growing deep seated mass near a joint, without much pain. The area of the knee is the most common site, followed by the foot, ankle, elbow, and shoulder. Other primary sites include the buttocks, abdominal wall, retioperitoneum and mediastinum. The tumor spreads along tissue planes. The most common site of visceral metastasis is lung. The diagnosis is made by biopsy. Treatment is wide resection of the tumor including adjacent muscle and regional lymph nodes, followed by chemotherapy and radiation therapy. Currently used chemotherapeutic agents are doxorubicin, ifosfamide, and cisplatin. Amputation of the involved distal extremity may be required. Chemotherapy may be beneficial in some patients with metastatic disease. Isolated pulmonary mitostasis can be surgically removed. The 5-year survival with treatment has been reported as high as 88%.

(Bibliography omitted in Palm version)

Back to Table of Contents

326. PERIARTICULAR DISORDERS OF THE EXTREMITIES - *Bruce Gilliland*

A number of periarticular disorders have become increasingly common over the past two to three decades, due in part to greater participation in recreational sports by individuals of a wide range of ages. This chapter discusses some of the more common periarticular disorders of the extremities.

BURSITIS

Bursitis is inflammation of a bursa, which is a thin-walled sac lined with synovial tissue. The function of the bursa is to facilitate movement of tendons and muscles over bony prominences. Excessive frictional forces, trauma, systemic disease (e.g., rheumatoid arthritis, gout), or infection may cause bursitis. *Subacromial bursitis* (subdeltoid bursitis) is the most common form of bursitis. The subacromial bursa, which is contiguous with the subdeltoid bursa, is located between the undersurface of the acromion and the humeral head, and is covered by the deltoid muscle. Bursitis is caused by repetitive overhead motion and often accompanies rotator cuff tendinitis. Another frequently encountered form is *trochanteric bursitis*, which involves the bursa around the insertion of the gluteus medius onto the greater trochanter of the femur. Patients experience pain over the lateral aspect of the hip and upper thigh and have tenderness over the posterior aspect of the greater trochanter. External rotation and resisted abduction of the hip elicit pain. *Olecranon bursitis* occurs over the posterior elbow, and when the area is acutely inflamed, infection should be excluded by aspirating and culturing fluid from the bursa. *Achilles bursitis* involves the bursa located above the insertion of the tendon to the calcaneus and results from overuse and wearing tight shoes. *Retrocalcaneal bursitis* involves the bursa that is located between the calcaneus and posterior surface of the Achilles tendon. The pain is experienced at the back of the heel, and swelling appears on the medial and/or lateral side of the tendon. It occurs in association with spondyloarthropathies, rheumatoid arthritis, gout, or trauma. *Ischial bursitis* (weaver's bottom) affects the bursa separating the gluteus medius from the ischial tuberosity and develops from prolonged sitting and pivoting on hard surfaces. *Iliopsoas bursitis* affects the bursa that lies between the iliopsoas muscle and hip joint and is lateral to the femoral vessels. Pain is experienced over this area and is made worse by hip extension and flexion. Bursitis results from trauma or overuse but can also be seen in patients with rheumatoid arthritis. *Anserine bursitis* is an inflammation of the sartorius bursa located over the medial side of the tibia just below the knee and under the conjoint tendon and is manifested by pain on climbing stairs. Tenderness is present over the insertion of the conjoint tendon of the sartorius, gracilis, and semitendinosus. *Prepatellar bursitis* (housemaid's knee) occurs in the bursa situated between the patella and overlying skin and is caused by kneeling on hard surfaces. Treatment of bursitis consists of prevention of the aggravating situation, rest of the involved part, administration of a nonsteroidal anti-inflammatory drug (NSAID), or local glucocorticoid injection.

ROTATOR CUFF TENDINITIS AND IMPINGEMENT SYNDROME

Tendinitis of the rotator cuff is the major cause of a painful shoulder and is currently thought to be caused by inflammation of the tendon(s). The rotator cuff consists of the tendons of the supraspinatus, infraspinatus, subscapularis, and teres minor muscles, and inserts on the humeral tuberosities. Of the tendons forming the rotator cuff, the supraspinatus tendon is the most often affected, probably because of its repeated impingement (impingement syndrome) between the humeral head and the undersurface of the anterior third of the acromion and coracoacromial ligament above as well as the reduction in its blood supply that occurs with abduction of the arm (Fig. 326-1). The tendon of the infraspinatus or the long head of the biceps is less commonly involved. The process begins with edema and hemorrhage of the rotator cuff, which evolves to fibrotic thickening and eventually to rotator cuff degeneration with tendon tears and bone spurs. Subacromial bursitis also accompanies this syndrome. Symptoms usually appear after injury or overuse, especially with activities involving elevation of the arm with some degree of forward flexion. Impingement syndrome occurs in persons participating in baseball, tennis, swimming, or occupations that require repeated elevation of the arm. Those over age 40 are particularly susceptible. Patients complain of a dull aching in the shoulder, which may interfere with sleep. Severe pain is experienced when the arm is actively abducted into an overhead position. The arc between 60 and 120° is especially painful. Tenderness is present over the lateral aspect of the humeral head just below the acromion.NSAIDs, local glucocorticoid injection, and physical therapy may relieve symptoms.

Patients may tear the supraspinatus tendon acutely by falling on an outstretched arm or lifting a heavy object. Symptoms are pain, along with weakness of abduction and external rotation of the shoulder. Atrophy of the supraspinatus muscles develops. The diagnosis is established by arthrogram or ultrasound. Surgical repair may be necessary in patients who fail to respond to conservative measures. In patients with moderate to severe tears and functional loss, surgery is indicated.

CALCIFIC TENDINITIS

This condition is characterized by deposition of calcium salts, primarily hydroxyapatite, within a tendon. The exact mechanism of calcification is not known but may be initiated by ischemia or degeneration of the tendon. The supraspinatus tendon is most often affected because it is frequently impinged on and has a reduced blood supply when the arm is abducted. The condition usually develops after age 40. Calcification within the tendon may evoke acute inflammation, producing sudden and severe pain in the shoulder. However, it may be asymptomatic or not related to the patient's symptoms.

BICIPITAL TENDINITIS AND RUPTURE

Bicipital tendinitis, or tenosynovitis, is produced by friction on the tendon of the long head of the biceps as it passes through the bicipital groove. When the inflammation is acute, patients experience anterior shoulder pain that radiates down the biceps into the forearm. Abduction and external rotation of the arm are painful and limited. The bicipital groove is very tender to palpation. Pain may be elicited along the course of the tendon by resisting supination of the forearm with the elbow at 90° (Yergason's supination sign). Acute rupture of the tendon may occur with vigorous exercise of the arm and is often painful. In a young patient, it should be repaired surgically. Rupture of the tendon in an older person may be associated with little or no pain and is recognized by the presence of persistent swelling of the biceps ("Popeye" muscle) produced by the retraction of the long head of the biceps. Surgery is usually not necessary in this setting.

ADHESIVE CAPSULITIS

Often referred to as "frozen shoulder," adhesive capsulitis is characterized by pain and restricted movement of the shoulder, usually in the absence of intrinsic shoulder disease. Adhesive capsulitis, however, may follow bursitis or tendinitis of the shoulder or be associated with systemic disorders such as chronic pulmonary disease, myocardial infarction, and diabetes mellitus. Prolonged immobility of the arm contributes to the development of adhesive capsulitis, and reflex sympathetic dystrophy is thought to be a pathogenic factor. The capsule of the shoulder is thickened, and a mild chronic inflammatory infiltrate and fibrosis may be present.

Adhesive capsulitis occurs more commonly in women after age 50. Pain and stiffness usually develop gradually over several months to a year but progress rapidly in some patients. Pain may interfere with sleep. The shoulder is tender to palpation, and both active and passive movement are restricted. Radiographs of the shoulder show osteopenia. The diagnosis is confirmed by arthrography, in that only a limited amount of contrast material, usually <15 mL, can be injected under pressure into the shoulder joint.

In most patients, the condition improves spontaneously 1 to 3 years after onset, but some have permanent restriction of movement. Early mobilization of the arm following an injury to the shoulder may prevent the development of this disease. Slow but forceful injection of contrast material into the joint may lyse adhesions and stretch the capsule, resulting in improvement of shoulder motion. Manipulation under anesthesia may be helpful in some patients. Once the disease is established, therapy may have little effect on its natural course. Local injections of glucocorticoids,NSAIDs, and physical therapy may provide relief of symptoms.

LATERAL EPICONDYLITIS (TENNIS ELBOW)

Lateral epicondylitis, or tennis elbow, is a painful condition involving the soft tissue over the lateral aspect of the elbow. The pain originates at or near the site of attachment of the common extensors to the lateral epicondyle and may radiate into the forearm and dorsum of the wrist. This painful condition is thought to be caused by small tears of the extensor aponeurosis resulting from repeated resisted contractions of the extensor muscles. The pain usually appears after work or recreational activities involving repeated motions of wrist extension and supination against resistance. Most patients with this disorder injure themselves in activities other than tennis, such as pulling weeds, carrying suitcases or briefcases, or using a screwdriver. The injury in tennis usually occurs when hitting a backhand with the elbow flexed. Shaking hands and opening doors can reproduce the pain. Striking the lateral elbow against a solid object may also induce pain.

The treatment is usually rest along with administration of anNSAID. Ultrasound, icing, and friction massage may also help relieve pain. When pain is severe, the elbow is placed in a sling or splinted at 90° of flexion. When the pain is acute and well localized, injection of a glucocorticoid using a small-gauge needle may be effective. Following injection, the patient should be advised to rest the arm for at least 1 month and avoid activities that would aggravate the elbow. Once symptoms have subsided, the patient should begin rehabilitation to strengthen and increase flexibility of the extensor muscles before resuming physical activity involving the arm. A forearm band placed 2.5 to 5.0 cm (1 to 2 in) below the elbow may help to reduce tension on the extensor muscles at their attachment to the lateral epicondyle. The patient should be advised to restrict activities requiring forcible extension and supination of the wrist. Improvement may take several months. The patient may continue to experience mild pain but, with care, can usually avoid the return of debilitating pain. In an occasional patient, surgical release of the extensor aponeurosis may be necessary.

MEDIAL EPICONDYLITIS

Medial epicondylitis is an overuse syndrome resulting in pain over the medial side of the elbow with radiation into the forearm. The cause of this syndrome is considered to be repetitive resisted motions of wrist flexion and pronation, which lead to microtears and granulation tissue at the origin of the pronator teres and forearm flexors, particularly the flexor carpi radialis. This overuse syndrome is usually seen in patients >35 years and is much less common than lateral epicondylitis. It occurs most often in work-related repetitive activities but also occurs with recreational activities such as swinging a golf club (golfer's elbow) or throwing a baseball. On physical examination, there is tenderness just distal to the medial epicondyle over the origin of the forearm flexors. Pain can be reproduced by resisting wrist flexion and pronation with the elbow extended. Radiographs are usually normal. The differential diagnosis of patients with medial elbow symptoms include tears of the pronator teres, acute medial collateral ligament tear, and medial collateral ligament instability. Ulnar neuritis has been found in 25 to 50% of patients with medial epicondylitis and is associated with tenderness over the ulnar nerve at the elbow as well as hypesthesia and paresthesia on the ulnar side of the hand.

The initial treatment of medial epicondylitis is conservative, involving rest,NSAIDs, friction massage, ultrasound, and icing. Some patients may require splinting. Injections of glucocorticoids at the painful site may also be effective. Patients should be instructed to rest at least 1 month. Also, patients should be started on physical therapy once the pain has subsided. In patients with chronic debilitating medial epicondylitis that remains unresponsive after at least a year of treatment, surgical release of the flexor muscle at its origin may be necessary and is often successful.

(Bibliography omitted in Palm version)

Back to Table of Contents

PART THIRTEEN -ENDOCRINOLOGY AND METABOLISM

SECTION 1 -ENDOCRINOLOGY

327. PRINCIPLES OF ENDOCRINOLOGY - *J. Larry Jameson*

The management of endocrine disorders requires an understanding of such disparate areas as intermediary metabolism, reproductive physiology, bone metabolism, and growth. Accordingly, the practice of endocrinology is intimately linked to a conceptual framework for understanding hormone secretion, hormone action, and principles of feedback control systems. The endocrine system is investigated primarily by measuring hormone concentrations, thereby arming the clinician with valuable diagnostic information. Most disorders of the endocrine system are amenable to effective treatment, once the correct diagnosis is determined. Endocrine deficiency disorders are treated with physiologic hormone replacement; hormone excess conditions, usually due to benign glandular adenomas, are managed by removing tumors surgically or by reducing hormone levels medically.

SCOPE OF ENDOCRINOLOGY

The specialty of endocrinology encompasses the study of glands and the hormones they produce. The term *endocrine* was coined by Starling to contrast the actions of hormones secreted internally (endocrine) with those secreted externally (*exocrine*) or into a lumen, such as the gastrointestinal tract. The term *hormone*, derived from a Greek phrase meaning "to set in motion," aptly describes the dynamic actions of these circulating substances as they elicit cellular responses and regulate physiologic processes through feedback mechanisms.

Unlike certain other specialties in medicine, it is not possible to define endocrinology strictly along anatomic lines. The classic endocrine glands -- pituitary, thyroid, parathyroid, pancreatic islets, adrenal, and gonads -- communicate broadly with other organs through the nervous system, hormones, cytokines, and growth factors. In addition to its traditional synaptic functions, the brain produces a vast array of peptide hormones, spawning the discipline of neuroendocrinology. Through the production of hypothalamic releasing factors, the central nervous system exerts a major regulatory influence over pituitary hormone secretion (Chap. 328). The peripheral nervous system modulates adrenal medulla and pancreatic islet hormone production. The immune and endocrine systems are also intimately intertwined. The adrenal glucocorticoid, cortisol, is a powerful immunosuppressant. Cytokines and interleukins (ILs) have profound effects on the functions of the pituitary, adrenal, thyroid, and gonads. Common endocrine diseases, such as autoimmune thyroid disease and type 1 diabetes mellitus, are caused by dysregulation of immune surveillance and tolerance. Less common diseases such as polyglandular failure, Addison's disease, and lymphocytic hypophysitis also have an immunologic basis.

The interdigitation of endocrinology with physiologic processes in other specialties sometimes blurs the role of hormones. For example, hormones play an important role in maintenance of blood pressure, intravascular volume, and peripheral resistance in the cardiovascular system. The heart is the principal source of atrial natriuretic peptide, which acts in classic endocrine fashion to induce natriuresis at a distant target organ (the kidney). Vasoactive substances such as catecholamines, angiotensin II, endothelin, and nitric oxide are involved in dynamic changes of vascular tone, in addition to their multiple roles in other tissues. Erythropoietin, a traditional circulating hormone, is made in the kidney and stimulates erythropoiesis in the bone marrow (Chap. 104). The kidney is also integrally involved in the renin-angiotensin axis (Chap. 331) and is a primary target of several hormones including parathyroid hormone (PTH), mineralocorticoids, and vasopressin. The gastrointestinal tract produces a surprising number of peptide hormones such as cholecystokinin, gastrin, secretin, and vasoactive intestinal peptide, among many others. Carcinoid and islet tumors can secrete excessive amounts of these hormones, leading to specific clinical syndromes (Chap. 93). Many of these gastrointestinal hormones are also produced in the central nervous system, where their functions remain poorly understood. As new hormones such as inhibin, ghrelin, and leptin are discovered, they become integrated into the science and practice of medicine on the basis of their functional roles rather than through their structures or mechanisms of action.

Characterization of hormone receptors frequently reveals unexpected relationships to factors in nonendocrine disciplines. The growth hormone (GH) receptor, for example, is a member of the cytokine receptor family. The G protein-coupled receptors (GPCRs), which mediate the actions of many peptide hormones, are used in numerous physiologic processes including vision, smell, and neurotransmission.

It is apparent that hormones and growth factors play an important functional role in all organ systems. Though endocrinologists are not usually involved in the administration of the hormones or growth factors used to treat diseases in other specialties (e.g., cardiology, hematology), the principles of endocrinology can be applied in these cases, thus emphasizing the impact of endocrinology across multiple disciplines.

NATURE OF HORMONES

Hormones can be divided into five major classes: (1) *amino acid derivatives* such as dopamine, catecholamines, and thyroid hormone; (2) *small neuropeptides* such as gonadotropin-releasing hormone (GnRH), thyrotropin-releasing hormone (TRH), somatostatin, and vasopressin; (3) *large proteins* such as insulin, luteinizing hormone (LH), and PTH produced by classic endocrine glands; (4) *steroid hormones* such as cortisol and estrogen that are synthesized from cholesterol-based precursors; and (5) *vitamin derivatives* such as retinoids (vitamin A) and vitamin D. A variety of *peptide growth factors*, most of which act locally, share actions with hormones. As a rule, amino acid derivatives and peptide hormone interact with cell-surface membrane receptors. Steroids, thyroid hormones, vitamin D, and retinoids are lipid-soluble and interact with intracellular nuclear receptors.

HORMONE AND RECEPTOR FAMILIES

Many hormones and receptors can be grouped into families, reflecting their structural similarities (Table 327-1). The evolution of these families generates diverse but highly selective pathways of hormone action. Recognizing these relationships allows extrapolation of information gleaned from one hormone or receptor to other family members.

The glycoprotein hormone family, consisting of thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), LH, and human chorionic gonadotropin (hCG), illustrates many features of related hormones. The glycoprotein hormones are heterodimers that share the a subunit in common; theb subunits are distinct and confer specific biologic actions. The overall three-dimensional architecture of theb subunits is similar, reflecting the locations of conserved disulfide bonds that restrain protein conformation. The cloning of theb-subunit genes from multiple species suggests that this family arose from a common ancestral gene, probably by gene duplication and subsequent divergence to evolve new biologic functions.

As the hormone families enlarge and diverge, their receptors must co-evolve, if new biologic functions are to be derived. RelatedGPCRs, for example, have evolved for each of the glycoprotein hormones. These receptors are structurally similar, and each is coupled to the G$_{s}$asignaling pathway. However, there is minimal overlap of hormone binding. For example, TSH binds with high specificity to the TSH receptor but interacts weakly with the LH or the FSH receptor. Nonetheless, there can be subtle physiologic consequences of hormone cross-reactivity with other receptors. Very high levels ofhCGduring pregnancy stimulate the TSH receptor and increase thyroid hormone levels.

Insulin, insulin-like growth factor (IGF) I, and IGF-II share structural similarities that are most apparent when precursor forms of the proteins are compared. In contrast to the high degree of specificity seen with the glycoprotein hormones, there is moderate cross-talk among the members of the insulin/IGF family. High concentrations of an IGF-II precursor produced by certain tumors (e.g., sarcomas) can cause hypoglycemia, partly because of binding to insulin and IGF-I receptors (Chap. 334). High concentrations of insulin also bind to the IGF-I receptor, perhaps accounting for some of the clinical manifestations seen in severe insulin resistance.

Another important example of receptor cross-talk is seen with PTH and parathyroid hormone-related peptide (PTHrP) (Chap. 341). PTH is produced by the parathyroid glands, whereas PTHrP is expressed at high levels during development and by a variety of tumors. These hormones share amino acid sequence similarity, particularly in their amino-terminal regions. Both hormones bind to a single PTH receptor that is expressed in bone and kidney. Hypercalcemia and hypophosphatemia may therefore result from excessive production of either hormone, making it difficult to distinguish hyperparathyroidism from hypercalcemia of malignancy solely on the basis of serum chemistries. However, sensitive and specific assays for PTH now allow these disorders to be separated more readily.

Based on their specificities for DNA binding sites, the nuclear receptor family can be subdivided into type 1 receptors (GR, MR, AR, ER, PR) that bind steroids and type 2 receptors (TR, VDR, RAR, PPAR) that bind thyroid hormone, vitamin D, retinoic acid, or lipid derivatives. Certain functional domains in nuclear receptors, such as the zinc finger DNA-binding domains, are highly conserved. However, selective amino acid differences within this domain confer DNA sequence specificity. The hormone-binding domains are more variable, providing great diversity in the array of small molecules that can bind to different nuclear receptors. With few exceptions, hormone binding is highly specific for a single type of nuclear receptor. One exception involves the highly related glucocorticoid and mineralocorticoid receptors. Because the mineralocorticoid receptor also binds glucocorticoids with high affinity, an enzyme (11b-hydroxysteroid dehydrogenase) located in renal tubular cells inactivates glucocorticoids, allowing selective responses to mineralocorticoids such as aldosterone. However, when very high glucocorticoid concentrations occur, as in Cushing's syndrome, the glucocorticoid degradation pathway becomes saturated, allowing excessive cortisol levels to exert mineralocorticoid effects (sodium retention, potassium wasting). This phenomenon is particularly pronounced in ectopic adrenocorticotropic hormone (ACTH) syndromes (Chap. 331). Another example of relaxed nuclear receptor specificity involves the estrogen receptor, which can bind an array of compounds, some of which share little structural similarity to the high-affinity ligand estradiol. This feature of the estrogen receptor makes it susceptible to activation by "environmental estrogens" such as resveratrol, octylphenol, and many other aromatic hydrocarbons. On the other hand, this lack of specificity provides an opportunity to synthesize a remarkable series of clinically useful antagonists (e.g., tamoxifen) and selective estrogen response modulators (SERMs), such as raloxifene. These compounds generate distinct conformations that alter receptor interactions with components of the transcription machinery (see below), thereby conferring their unique actions.

HORMONE SYNTHESIS AND PROCESSING

The synthesis of peptide hormones and their receptors occurs through a classic pathway of gene expression: transcription ® mRNA® protein ®posttranslational protein processing ®intracellular sorting, membrane integration, or secretion (Chap. 65). Though endocrine genes contain regulatory DNA elements similar to those found in many other genes, their exquisite control by other hormones also necessitates the presence of specific hormone response elements. For example, the TSH genes are repressed directly by thyroid hormones acting through the thyroid hormone receptor, a member of the nuclear receptor family. Steroidogenic enzyme gene expression requires specific transcription factors such as steroidogenic factor-1 (SF-1), acting in conjunction with signals transmitted by trophic hormones (e.g., ACTH or LH). For some hormones, substantial regulation occurs at the level of translational efficiency. Insulin biosynthesis, while requiring ongoing gene transcription, is regulated primarily at the translational level in response to elevated levels of glucose or amino acids.

Many hormones are embedded within larger precursor polypeptides that are proteolytically processed to yield the biologically active hormone. Examples include: proopiomelanocortin (POMC)® ACTH; proglucagon® glucagon; proinsulin® insulin; pro-PTH ®PTH, among others. In many cases, such as POMC and proglucagon, these precursors generate multiple biologically active peptides. It is provocative that hormone precursors are typically inactive, presumably adding an additional level of regulatory control. This is true not only for peptide hormones but also for certain steroids (testosterone ®dihydrotestosterone) and thyroid hormone (T_4®T_3).

Hormone precursor processing is intimately linked to intracellular sorting pathways that transport proteins to appropriate vesicles and enzymes, resulting in specific cleavage steps, followed by protein folding and translocation to secretory vesicles. Hormones destined for secretion are translocated across the endoplasmic reticulum under the guidance of an amino-terminal signal sequence that is subsequently cleaved. Cell-surface receptors are inserted into the membrane via short segments of hydrophobic amino acids that remain embedded within the lipid bilayer. During translocation through the Golgi and endoplasmic reticulum, hormones and receptors are also subject to a variety of posttranslational modifications, such as glycosylation and phosphorylation, which can alter protein conformation, modify circulating half-life, and alter biologic activity.

Synthesis of most steroid hormones is based on modifications of the precursor, cholesterol. Multiple regulated enzymatic steps are required for the synthesis of testosterone (Chap. 335), estradiol (Chap. 336), cortisol (Chap. 331), and vitamin D (Chap. 340). This large number of synthetic steps predisposes to multiple genetic and acquired disorders of steroidogenesis (see below).

HORMONE SECRETION, TRANSPORT, AND DEGRADATION

The circulating level of a hormone is determined by its rate of secretion and its circulating half-life. After protein processing, peptide hormones (GnRH, insulin, GH) are stored in secretory granules. As these granules mature, they are poised beneath the plasma membrane for imminent release into the circulation. In most instances, the stimulus for hormone secretion is a releasing factor or neural signal that induces rapid changes in intracellular calcium concentrations, leading to secretory granule fusion with the plasma membrane and release of its contents into the extracellular environment and blood stream. Steroid hormones, in contrast, diffuse into the circulation as they are synthesized. Thus, their secretory rates are closely aligned with rates of synthesis. For example, ACTH and LH induce steroidogenesis by stimulating the activity of *st*eroidogenic *a*cute *r*egulatory (StAR) protein (transports cholesterol into the mitochondrion) along with other rate-limiting steps (e.g., cholesterol side-chain cleavage enzyme, CYP11A1) in the steroidogenic pathway.

Hormone transport and degradation dictate the rapidity with which a hormonal signal decays. Some hormonal signals are evanescent (e.g., somatostatin), whereas others are longer lived (e.g., TSH). Because somatostatin exerts effects in virtually every tissue, a short half-life allows it concentrations and actions to be controlled locally. Structural modifications that impair somatostatin degradation have been useful for generating long-acting therapeutic analogues, such as octreotide (Chap. 328). On the other hand, the actions of TSH are highly specific for the thyroid gland. Its prolonged half-life accounts for relatively constant serum levels, even though TSH is secreted in discrete pulses.

An understanding of circulating hormone half-life is important for achieving physiologic hormone replacement, as the frequency of dosing and the time required to reach steady state are intimately linked to rates of hormone decay. T_4, for example, has a plasma half-life of 7 days. Consequently,>1 month is required to reach a new steady state, but single daily doses are sufficient to achieve constant hormone levels. T_3, in contrast, has a half-life of 1 day. Its administration is associated with more dynamic serum levels and it must be administered two to three times per day. Similarly, synthetic glucocorticoids vary widely in their half-lives; those with longer half-lives (e.g., dexamethasone) are associated with greater suppression of the hypothalamic-pituitary-adrenal (HPA) axis. Most protein hormones [e.g., ACTH, GH, prolactin (PRL); PTH, LH] have relatively short half-lives (<20 min), leading to sharp peaks of secretion and decay. The only accurate way to profile the pulse frequency and amplitude of these hormones is to measure levels in frequently sampled blood (every 10 min) over long durations (8 to 24 h). Because this is not practical in a clinical setting, an alternative strategy is to pool three to four samples drawn at about 30-min intervals, recognizing that pulsatile secretion makes it difficult to establish a narrow normal range. Rapid hormone decay is useful in certain clinical settings. For example, the short half-life of PTH allows the use of intraoperative PTH determinations to confirm successful removal of an adenoma. This is particularly valuable diagnostically when there is a possibility of multicentric disease or parathyroid hyperplasia, as occurs with multiple endocrine neoplasia (MEN) or renal insufficiency.

Many hormones circulate in association with serum-binding proteins. Examples include: (1) T_4 and T_3binding to thyroxine-binding globulin (TBG), albumin, and thyroxine-binding prealbumin (TBPA); (2) cortisol binding to cortisol-binding globulin (CBG); (3) androgen and estrogen binding to sex hormone-binding globulin (SHBG) (also called testosterone-binding globulin, TeBG); (4)IGF-I and -II binding to multiple IGF-binding proteins (IGFBPs); (5) GH interactions with GH-binding protein (GHBP), a circulating fragment of the GH receptor extracellular domain; and (6) activin binding to follistatin. These interactions provide a hormonal reservoir, prevent otherwise rapid degradation of unbound hormones, restrict hormone access to certain sites (e.g., IGFBPs), and modulate the unbound, or "free," hormone concentrations. Although a variety of binding protein abnormalities have been identified, most have little clinical consequence, aside from creating diagnostic problems. For example, TBG deficiency can greatly reduce total thyroid hormone levels, but the free concentrations of T_4 and T_3remain normal. Liver disease and certain medications can also influence binding protein levels (e.g., estrogen increases TBG) or cause displacement of hormones from binding proteins (e.g., salsalate displaces T_4 from TBG). Only free hormone is available to bind receptors and thereby elicit a biologic response. Short-term perturbations in binding proteins change the free hormone concentration, which in turn induces compensatory adaptations through feedback loops. SHBG changes in women are an exception to this self-correcting mechanism. When SHBG decreases because of insulin resistance or androgen excess, the free testosterone concentration is increased, potentially leading to hirsutism (Chap. 53). The increased free testosterone levels does not result in an adequate compensatory feedback correction because estrogen, and not testosterone, is the primary regulator of the reproductive axis.

HORMONE ACTION THROUGH RECEPTORS

Receptors for hormones are divided into two major classes -- membrane and nuclear. *Membrane receptors* primarily bind peptide hormones and catecholamines. *Nuclear receptors* bind small molecules that can diffuse across the cell membrane, such as thyroid hormone, steroids, and vitamin D. Certain general principles apply to hormone-receptor interactions, regardless of the class of receptor. Hormones bind to receptors with specificity and a high affinity that generally coincides with the dynamic range of circulating hormone concentrations. Low concentrations of free hormone (usually $10^{-12}$to 10^{-9}*M*) rapidly associate and dissociate from receptors in a bimolecular reaction, such that the occupancy of the receptor at any given moment is a function of hormone concentration and the receptor's affinity for the hormone. Receptor numbers vary greatly in different target tissues, providing one of the major determinants of specific cellular responses to circulating hormones. For example, ACTH receptors are located almost exclusively in the adrenal cortex, and FSH receptors are found only in the gonads. In contrast, insulin and thyroid hormone receptors are widely distributed, reflecting the need for metabolic responses in all tissues.

MEMBRANE RECEPTORS

Membrane receptors for hormones can be divided into several major groups: (1) seven transmembraneGPCRs, (2) tyrosine kinase receptors, (3) cytokine receptors, and (4) serine kinase receptors (Fig. 327-1). The *seven transmembrane GPCR* family binds a remarkable array of hormones including large proteins (e.g., LH, PTH), small peptides (e.g.,TRH, somatostatin), catecholamines (epinephrine, dopamine), and even minerals (e.g., calcium). The extracellular domains of GPCRs vary widely in size and are the major binding site for large hormones. The transmembrane-spanning regions are composed of hydrophobic a-helical domains that traverse the lipid bilayer. Like some channels, these domains are thought to circularize and form a hydrophobic pocket into which certain small ligands fit. Hormone binding induces conformational changes in these domains, transducing structural changes to the intracellular domain, which is a docking site for G proteins.

The large family of *G proteins*, so named because they bind guanine nucleotides (GTP, GDP), provides great diversity for coupling to different receptors. G proteins form a heterotrimeric complex that is composed of various aand bg subunits. The asubunit contains the guanine nucleotide-binding site and hydrolyzes GTP® GDP. The bg subunits are tightly associated and modulate the activity of thea subunit, as well as mediating their own effector signaling pathways. G protein activity is regulated by a cycle that involves GTP hydrolysis and dynamic interactions between the a andbg subunits. Hormone binding to the receptor induces GDP dissociation, allowing Ga to bind GTP and dissociate from the bg complex. Under these conditions, the Ga subunit is activated and mediates signal transduction through various enzymes such as adenylate cyclase or phospholipase C. GTP hydrolysis to GDP allows reassociation with thebg subunits and restores the inactive state. As described below, a variety of endocrinopathies result from G protein mutations or from mutations in receptors that modify their interactions with G proteins.

There are more than a dozen isoforms of the Ga subunit. $G_s$astimulates, whereas $G_i$ainhibits adenylate cyclase, an enzyme that generates the second messenger, cyclic AMP, leading to activation of protein kinase A (Table 327-1). $G_q$subunits couple to phospholipase C, generating diacylglycerol and inositol triphosphate, leading to activation of protein kinase C and the release of intracellular calcium.

The *tyrosine kinase receptors* transduce signals for insulin and a variety of growth factors, such asIGF-I, epidermal growth factor (EGF), nerve growth factor, platelet-derived growth factor, and fibroblast growth factor. The cysteine-rich extracellular ligand-binding domains contain growth factor binding sites. After ligand binding, this class of receptors undergoes autophosphorylation, inducing interactions with intracellular adaptor proteins such as Shc and insulin receptor substrates 1 to 4. In the case of the insulin receptor, multiple kinases are activated including the Raf-Ras-MAPK and the Akt/protein kinase B pathways. The tyrosine kinase receptors play a prominent role in cell growth and differentiation as well as in intermediary metabolism.

The GH and PRL receptors belong to the *cytokine receptor* family (Chap. 305). Analogous to the tyrosine kinase receptors, ligand binding induces receptor binding to intracellular kinases -- the Janus kinases (JAKs), which phosphorylate members of the signal transduction and activators of transcription (STAT) family -- as well as other signaling pathways (Ras, PI3-K, MAPK). The activated STAT proteins translocate to the nucleus and stimulate expression of target genes (Chap. 328).

The *serine kinase receptors* mediate the actions of activins, transforming growth factor b, mullerian-inhibiting substance (MIS, also known as anti-mullerian hormone, AMH), and bone morphogenic proteins (BMPs). This family of receptors (consisting of type I and II subunits) signal through proteins termed *smads* (fusion of terms for *Caenorhabditis elegans* sma + mammalian mad). Like theSTATproteins, the smads serve a dual role of transducing the receptor signal and acting as transcription factors. The pleomorphic actions of these growth factors dictate that they act primarily in a local (paracrine or autocrine) manner. Binding proteins, such as follistatin (which binds activin and other members of this family), function to inactivate the growth factors and restrict their distribution.

NUCLEAR RECEPTORS

The family of nuclear receptors has grown to nearly 100 members, many of which are still classified as orphan receptors because their ligands, if they exist, remain to be identified (Fig. 327-2). Otherwise, most nuclear receptors are classified based on the nature of their ligands. Though all nuclear receptors ultimately act to increase or decrease gene transcription, some (e.g., glucocorticoid receptor) reside primarily in the cytoplasm, whereas others (e.g., thyroid hormone receptor) are always located in the nucleus. After ligand binding, the cytoplasmically localized receptors translocate to the nucleus.

The structures of nuclear receptors have been extensively studied, including by x-ray crystallography. The DNA binding domain, consisting of two zinc fingers, contacts specific DNA recognition sequences in target genes. Most nuclear receptors bind to DNA as dimers. Consequently, each monomer recognizes an individual DNA motif, referred to as a "half-site." The steroid receptors, including the glucocorticoid, estrogen, progesterone, and androgen receptors, bind to DNA as homodimers. Consistent with this twofold symmetry, their DNA recognition half-sites are palindromic. The thyroid, retinoid, PPAR, and vitamin D receptors bind to DNA preferentially as heterodimers in combination with retinoid X receptors (RXRs). Their DNA half-sites are arranged as direct repeats. Receptor specificity for DNA sequences is determined by (1) the sequence of the half-site, (2) the orientation of the half-sites (palindromic, direct repeat), and (3) the spacing between the half-sites. For example, vitamin D, thyroid and retinoid receptors recognize similar tandemly repeated half-sites (TAAGTCA), but these DNA repeats are spaced by three, four, and five nucleotides, respectively.

The carboxy-terminal hormone-binding domain mediates transcriptional control. For type II receptors, such as TR and RAR, co-repressor proteins bind to the receptor in the absence of ligand and silence gene transcription. Hormone binding induces conformational changes, triggering the release of co-repressors and inducing the recruitment of coactivators that stimulate transcription. Thus, these receptors are capable of mediating dramatic changes in the level of gene activity. Certain disease states are associated with defective regulation of these events. For example, mutations in the thyroid hormone receptor prevent co-repressor dissociation, resulting in a dominant form of hormone resistance (Chap. 330). In promyelocytic leukemia, fusion of RARa to other nuclear proteins causes aberrant gene silencing and prevents normal cellular differentiation. Treatment with retinoic acid reverses this repression and allows cellular differentiation and apoptosis to occur (Chap. 111). Type 1 steroid receptors do not interact with co-repressors, but ligand binding still mediates interactions with an array of coactivators. X-ray crystallography shows that variousSERMsinduce distinct receptor conformations. The tissue-specific responses caused by these agents in breast, bone, and uterus appear to reflect distinct interactions with coactivators. The receptor-coactivator complex stimulates gene transcription by several pathways including (1) recruitment of enzymes (histone acetyl transferases) that modify chromatin structure, (2) interactions with additional transcription factors on the target gene, and (3) direct interactions with components of the general transcription apparatus to enhance the rate of RNA polymerase II-mediated transcription.

FUNCTIONS OF HORMONES

The functions of individual hormones are described in detail in subsequent chapters. Nonetheless, it is useful to illustrate how most biologic responses require integration of several different hormonal pathways. The physiologic functions of hormones can be divided into three general areas: (1) growth and differentiation, (2) maintenance of homeostasis, and (3) reproduction.

GROWTH

Multiple hormones and nutritional factors mediate the complex phenomenon of growth (Chap. 328). Short stature may be caused by GH deficiency, hypothyroidism, Cushing's syndrome, precocious puberty, malnutrition or chronic illness, or genetic abnormalities that affect the epiphyseal growth plates (e.g., *FGFR3* or *SHOX* mutations). Many factors (GH,IGF-I, thyroid hormone) stimulate growth, whereas others (sex steroids) lead to epiphyseal closure. Understanding these hormonal interactions is important in the diagnosis and management of growth disorders. For example, delaying exposure to high levels of sex steroids may enhance the efficacy of GH treatment.

MAINTENANCE OF HOMEOSTASIS

Though virtually all hormones affect homeostasis, the most important among these are the following:

1. Thyroid hormone -- controls about 25% of basal metabolism in most tissues (Chap. 330)

2. Cortisol -- exerts a permissive action for many hormones in addition to its own direct effects (Chap. 331)

3. PTH -- regulates calcium and phosphorus levels (Chap. 341)

4. Vasopressin -- regulates serum osmolality by controlling renal free water clearance (Chap. 329)

5. Mineralocorticoids -- control vascular volume and serum electrolyte (Na^+, K^+) concentrations (Chap. 331)

6. Insulin -- maintains euglycemia in the fed and fasted states (Chap. 333)

The defense against hypoglycemia is an impressive example of integrated hormone action (Chap. 334). In response to the fasted state and falling blood glucose, insulin secretion is suppressed, resulting in decreased glucose uptake and enhanced glycogenolysis, lipolysis, proteolysis, and gluconeogenesis to mobilize fuel sources. If hypoglycemia develops (usually from insulin administration or sulfonylureas), an orchestrated counterregulatory response occurs -- glucagon and epinephrine rapidly stimulate glycogenolysis and gluconeogenesis, whereas GH and cortisol act over several hours to raise glucose levels and antagonize insulin action.

Although free water clearance is primarily controlled by vasopressin, cortisol and thyroid hormone are also important for facilitating renal tubular responses to vasopressin effects (Chap. 329). PTH and vitamin D function in an interdependent manner to control calcium metabolism (Chap. 340). PTH stimulates renal synthesis of 1,25 dihydroxyvitamin D, which increases calcium absorption in the gastrointestinal tract and enhances PTH action in bone. Increased calcium, along with vitamin D, feeds back to suppress PTH, thereby maintaining calcium balance.

Depending on the severity of a given stress and whether it is acute or chronic, multiple endocrine and cytokine pathways are activated to mount an appropriate physiologic response (Chap. 328). In severe acute stress such as trauma or shock, the sympathetic nervous system is activated and catecholamines are released, leading to increased cardiac output and a primed musculoskeletal system. Catecholamines also increase mean blood pressure and stimulate glucose production (Chap. 72). Multiple stress-induced pathways converge on the hypothalamus, stimulating several hormones including vasopressin and corticotropin-releasing hormone (CRH). These hormones, in addition to cytokines (tumor necrosis factor a, IL-2, IL-6), increase ACTH and GH production. ACTH stimulates the adrenal gland, increasing cortisol, which in turn helps to sustain blood pressure and dampen the inflammatory response. Increased vasopressin acts to conserve free water.

REPRODUCTION

The stages of reproduction include: (1) sex determination during fetal development (Chap. 338); (2) sexual maturation during puberty (Chap. 8); (3) conception, pregnancy, lactation, and child-rearing (Chap. 336), and (4) cessation of reproductive capability at menopause. Each of these stages involves an orchestrated interplay of multiple hormones, a phenomenon well illustrated by the dynamic hormonal changes that occur during each 28-day menstrual cycle. In the early follicular phase, pulsatile secretion of LH and FSH stimulate the progressive maturation of the ovarian follicle. This results in a gradual increase of estrogen and progesterone leading to enhanced pituitary sensitivity toGnRH, which, when combined with accelerated GnRH secretion, triggers the LH surge and rupture of the mature follicle. Inhibin, a protein produced by the granulosa cells, enhances follicular growth and feeds back to the pituitary to selectively suppress FSH, without affecting LH. Growth factors, such asEGF andIGF-I modulate follicular responsiveness to gonadotropins. Vascular endothelial growth factor and prostaglandins play a role in follicle vascularization and rupture.

During pregnancy, the increased production of prolactin, in combination with placentally derived steroids (e.g., estrogen and progesterone), prepares the breast for lactation (Chap. 337). Estrogens induce the production of progesterone receptors, allowing for increased responsiveness to progesterone. In addition to these and other hormones involved in lactation, the nervous system and oxytocin mediate the suckling response and milk release.

HORMONAL FEEDBACK REGULATORY SYSTEMS

Feedback control, both negative and positive, is a fundamental feature of endocrine systems. Each of the major hypothalamic-pituitary-hormone axes is governed by negative feedback, a process that maintains hormone levels within a relatively narrow range (Chap. 328). Examples of hypothalamic-pituitary negative feedback include (1) thyroid hormones on theTRH-TSH axis, (2) cortisol on theCRH-ACTH axis, (3) gonadal steroids on theGnRH-LH/FSH axis, and (4)IGF-I on the growth hormone-releasing hormone (GHRH)-GH axis (Fig. 327-3). These regulatory loops include both positive (e.g., TRH, TSH) and negative components (e.g., T_4, T_3), allowing for exquisite control of hormone levels. As an example, a small reduction of thyroid hormone triggers a rapid increase of TRH and TSH secretion, resulting in thyroid gland stimulation and increased thyroid hormone production. When the thyroid hormone reaches a normal level, it feeds back to suppress TRH and TSH, and a new steady state is attained. Feedback regulation also occurs for endocrine systems that do not involve the pituitary gland, such as calcium feedback on PTH, glucose inhibition of insulin secretion, and leptin feedback on the hypothalamus. An understanding of feedback regulation provides important insights into endocrine testing paradigms (see below).

Positive feedback control also occurs but is not well understood. The primary example is estrogen-mediated stimulation of the midcycle LH surge. Though chronic low levels of estrogen are inhibitory, gradually rising estrogen levels stimulate LH secretion. This effect, which is illustrative of an endocrine rhythm (see below), involves activation of the hypothalamicGnRHpulse generator. In addition, estrogen-primed gonadotropes are extroardinarily sensitive to GnRH, leading to a 10- to 20-fold amplification of LH release.

PARACRINE AND AUTOCRINE CONTROL

The aforementioned examples of feedback control involve classic endocrine pathways in which hormones are released by one gland and act on a distant target gland.

However, local regulatory systems, often involving growth factors, are increasingly recognized. *Paracrine regulation* refers to factors released by one cell that act on an adjacent cell in the same tissue. For example, somatostatin secretion by pancreatic islet d cells inhibits insulin secretion from nearby b cells. *Autocrine regulation* describes the action of a factor on the same cell from which it is produced.IGF-I acts on many cells that produce it, including chondrocytes, breast epithelium, and gonadal cells. Unlike endocrine actions, paracrine and autocrine control are difficult to document because local growth factor concentrations cannot be readily measured.

Anatomic relationships of glandular systems also greatly influence hormonal exposure -- the physical organization of islet cells enhances their intercellular communication; the portal vasculature of the hypothalamic-pituitary system exposes the pituitary to high concentrations of hypothalamic releasing factors; testicular seminiferous tubules gain exposure to high testosterone levels produced by the interdigitated Leydig cells; the pancreas receives nutrient information from the gastrointestinal tract; and the liver is the proximal target of insulin action because of portal drainage from the pancreas.

HORMONAL RHYTHMS

The feedback regulatory systems described above are superimposed on hormonal rhythms that are used for adaptation to the environment. Seasonal changes, the daily occurrence of the light-dark cycle, sleep, meals, and stress are examples of the many environmental events that affect hormonal rhythms. The *menstrual cycle* is repeated on average every 28 days, reflecting the time required to follicular maturation and ovulation (Chap. 336). Essentially all pituitary hormone rhythms are entrained to sleep and the *circadian cycle*, generating reproducible patterns that are repeated approximately every 24 h. TheHPAaxis, for example, exhibits characteristic peaks of ACTH and cortisol production in the early morning, with a nadir in the afternoon and evening. Recognition of these rhythms is important for endocrine testing and treatment. Patients with Cushing's syndrome characteristically exhibit increased midnight cortisol levels when compared to normal individuals (Chap. 331). In contrast, morning cortisol levels are similar in these groups, as cortisol is normally high at this time of day in normal individuals. The HPA axis is more susceptible to suppression by glucocorticoids administered at night as they blunt the early morning rise of ACTH. Understanding these rhythms allows glucocorticoid replacement that mimics diurnal production by administering larger doses in the morning than in the afternoon (Chap. 331).

Other endocrine rhythms occur on a more rapid time scale. Many peptide hormones are secreted in discrete bursts every few hours. LH and FSH secretion are exquisitely sensitive toGnRHpulse frequency. Intermittent pulses of GnRH are required to maintain pituitary sensitivity, whereas continuous exposure to GnRH causes pituitary gonadotrope desensitization. This feature of the hypothalamic-pituitary-gonadotrope (HPG) axis forms the basis for using long-acting GnRH agonists to treat central precocious puberty or to decrease testosterone levels in the management of prostate cancer.

It is important to be aware of the pulsatile nature of hormone secretion and the rhythmic patterns of hormone production when relating serum hormone measurements to normal values. For some hormones, integrated markers have been developed to circumvent hormonal fluctuations. Examples include 24-h urine collections for cortisol,IGF-I as a biologic marker of GH action, and HbA1c as an index of long-term (weeks to months) blood glucose control.

Often, one must interpret endocrine data only in the context of other hormonal results. For example, parathyroid hormone levels are typically assessed in combination with serum calcium concentrations. A high serum calcium level in association with elevated PTH is suggestive of hyperparathyroidism, whereas a suppressed PTH in this situation is more likely to be caused by hypercalcemia of malignancy or other causes of hypercalcemia. Similarly, TSH should be elevated when T_4 and T_3concentrations are low, reflecting reduced feedback inhibition. When this is not the case, it is important to consider other abnormalities in the hormonal axis, such as secondary hypothyroidism, which is caused by a defect at the level of the pituitary.

PATHOLOGIC MECHANISMS OF ENDOCRINE DISEASE

Endocrine diseases can be divided into three major types of conditions: (1) hormone excess, (2) hormone deficiency, and (3) hormone resistance (Table 327-2).

CAUSES OF HORMONE EXCESS

Syndromes of hormone excess can be caused by neoplastic growth of endocrine cells, autoimmune disorders, and excess hormone administration. Benign endocrine tumors, including parathyroid, pituitary, and adrenal adenomas, often retain the capacity to produce hormones, perhaps reflecting the fact that they are relatively well differentiated. Many endocrine tumors exhibit relatively subtle defects in their "set points" for feedback regulation. For example, in Cushing's disease, impaired feedback inhibition of ACTH secretion is associated with autonomous function. However, the tumor cells are not completely resistant to feedback, as revealed by the fact that ACTH is ultimately suppressed by higher doses of dexamethasone (e.g., high-dose dexamethasone test) (Chap. 331). Similar set point defects are also typical of parathyroid adenomas and autonomously functioning thyroid nodules.

The molecular basis of some endocrine tumors, such as theMENsyndromes (MEN-1, -2A, -2B), have provided important insights into tumorigenesis (Chap. 339). MEN-1 is characterized primarily by the triad of parathyroid, pancreatic islet, and pituitary tumors. MEN-2 predisposes to medullary thyroid carcinoma, pheochromocytoma, and hyperparathyroidism. The *MEN1* gene, located on chromosome 11q13, encodes a putative tumor-suppressor gene. Analogous to the paradigm first described for retinoblastoma, the affected individual inherits a mutant copy of the *MEN1* gene, and tumorigenesis ensues after a somatic "second hit" leads to loss of function of the normal *MEN1* gene (through deletion or point mutations).

In contrast to inactivation of a tumor-suppressor gene, as occurs inMEN-1 and most other inherited cancer syndromes, MEN-2 is caused by activating mutations in a single allele. In this case, activating mutations of the *RET* proto-oncogene, which encodes a receptor tyrosine kinase, leads to thyroid C-cell hyperplasia in childhood before the development of medullary thyroid carcinoma. Elucidation of the pathogenic mechanism has allowed early genetic screening for *RET* mutations in individuals at risk for MEN-2, permitting identification of those who may benefit from prophylactic thyroidectomy and biochemical screening for pheochromocytoma and hyperparathyroidism.

Mutations that activate hormone receptor signaling have been identified in severalGPCRs(Table 327-3). For example, activating mutations of the LH receptor causes a dominantly transmitted form of male-limited precocious puberty, reflecting premature stimulation of testosterone synthesis in Leydig cells (Chap. 335). Activating mutations in these GPCRs are located primarily in the transmembrane domains and induce receptor coupling to G_sa, even in the absence of hormone. Consequently, adenylate cyclase is activated and cyclic AMP levels increase in a manner that mimics hormone action. A similar phenomenon results from activating mutations in G_sa. When these occur early in development, they cause McCune-Albright syndrome. When they occur only in somatotropes, the activating $G_s$amutations cause GH-secreting tumors and acromegaly (Chap. 328).

In autoimmune Graves' disease, antibody interactions with the TSH receptor mimic TSH action, leading to hormone overproduction (Chap. 330). Analogous to the effects of activating mutations of the TSH receptor, these stimulating autoantibodies induce conformational changes that release the receptor from a constrained state, thereby triggering receptor coupling to G proteins.

CAUSES OF HORMONE DEFICIENCY

Most examples of hormone deficiency states can be attributed to glandular destruction caused by autoimmunity, surgery, infection, inflammation, infarction, hemorrhage, or tumor infiltration (Table 327-2). Autoimmune damage to the thyroid gland (Hashimoto's thyroiditis) and pancreatic islet b cells (type 1 diabetes mellitus) are prevalent causes of endocrine disease. Mutations in a number of hormones, hormone receptors, transcription factors, enzymes, and channels can also lead to hormone deficiencies (Table 327-3).

HORMONE RESISTANCE

Most severe hormone resistance syndromes are due to inherited defects in membrane receptors, nuclear receptors, or in the pathways that transduce receptor signals (Table 327-3). These disorders are characterized by defective hormone action, despite the presence of increased hormone levels. In complete androgen resistance, for example, mutations in the androgen receptor cause genetic (XY) males to have a female phenotypic appearance, even though LH and testosterone levels are increased (Chap. 338). In addition to these relatively rare genetic disorders, more common acquired forms of functional hormone resistance include insulin resistance in type 2 diabetes mellitus, leptin resistance in obesity, and GH resistance in catabolic states. The pathogenesis of functional resistance involves receptor downregulation and postreceptor desensitization of signaling pathways; functional forms of resistance are generally reversible.

Approach to the Patient

Because endocrinology interfaces with numerous physiologic systems, there is no standard endocrine history and examination. Moveover, because most glands are relatively inaccessible, the examination usually focuses on the manifestations of hormone excess or deficiency, as well as direct examination of palpable glands, such as the thyroid and gonads. For these reasons, it is important to evaluate patients in the context of their presenting symptoms, review of systems, family and social history, and exposure to medications that may affect the endocrine system. Astute clinical skills are required to detect subtle symptoms and signs suggestive of underlying endocrine disease. For example, a patient with Cushing's syndrome may manifest specific findings, such as central fat redistribution, striae, and proximal muscle weakness, in addition to features seen commonly in the general population, such as obesity, plethora, hypertension, and glucose intolerance. Similarly, the insidious onset of hypothyroidism -- with mental slowing, fatigue, dry skin, and other features -- can be difficult to distinguish from similar, nonspecific findings in the general population. Clinical judgment, based on knowledge of pathophysiology and experience, is required to decide when to embark on more extensive evaluation of these disorders. As described below, laboratory testing plays an essential role in endocrinology by allowing quantitative assessment of hormone levels and dynamics. Radiologic imaging tests, such as CT scan, MRI, thyroid scan, and ultrasound, are also used for the diagnosis of endocrine disorders. However, these tests are generally employed only after a hormonal abnormality has been established by biochemical testing.

Hormone Measurements and Endocrine Testing Radioimmunoassays are the most important diagnostic tool in endocrinology, as they allow sensitive, specific, and quantitative determination of steady-state and dynamic changes in hormone concentrations. Radioimmunoassays use antibodies to detect specific hormones. For many peptide hormones, these measurements are now configured as immunoradiometric assays (IRMAs), which use two different antibodies to increase binding affinity and specificity. There are many variations of these assays -- a common format involves using one antibody to capture the antigen (hormone) onto an immobilized surface and a second antibody, labeled with a fluorescent or radioactive tag, to detect the antigen. These assays are sensitive enough to detect plasma hormone concentrations in the picomolar to nanomolar range, and they can readily distinguish structurally related proteins, such as PTH fromPTHrP. A variety of other techniques are used to measure specific hormones, including mass spectroscopy, various forms of chromatography, and enzymatic methods; bioassays are now used rarely.

Most hormone measurements are based on plasma or serum samples. However, urinary hormone determinations remain useful for the evaluation of some conditions. Urinary collections over 24 h provide an integrated assessment of the production of a hormone or metabolite, many of which vary during the day. It is important to assure complete collections of 24-h urine samples; simultaneous measurement of creatinine provides an internal control for the adequacy of collection and can be used to normalize some hormone measurements. A 24-h urine free cortisol measurement largely reflects the amount of unbound cortisol, thus providing a reasonable index of biologically available hormone. Other commonly used urine determinations include: 17-hydroxycorticosteroids, 17-ketosteroids, vanillylmandelic acid (VMA), metanephrine, catecholamines, 5-hydroxyindoleacetic acid (5-HIAA), and calcium.

The value of quantitative hormone measurements lies in their correct interpretation in a clinical context. The normal range for most hormones is relatively broad, often varying by a factor of two- to tenfold. The normal ranges for many hormones are gender- and age-specific. Thus, using the correct normative database is an essential part of interpreting hormone tests. The pulsatile nature of hormones and factors that can affect their secretion, such as sleep, meals, and medications, must also be considered. Cortisol values increase fivefold between midnight and dawn; reproductive hormone levels vary dramatically during the female menstrual cycle.

For many endocrine systems, much information can be gained from basal hormone testing, particularly when different components of an endocrine axis are assessed simultaneously. For example, low testosterone and elevated LH levels suggest a primarily gonadal problem, whereas a hypothalamic-pituitary disorder is likely if both LH and testosterone are low. Because TSH is a sensitive indicator of thyroid function, it is generally recommended as a first-line test for thyroid disorders. An elevated TSH level is almost always the result of primary hypothyroidism, whereas a low TSH is most often caused by thyrotoxicosis. These predictions can be confirmed by determining the free thyroxine level. Elevated calcium and PTH levels suggest hyperparathyroidism, whereas PTH is suppressed in hypercalcemia caused by malignancy or granulomatous diseases. A suppressed ACTH in the setting of hypercortisolemia, or increased urine free cortisol, is seen with hyperfunctioning adrenal adenomas.

It is not uncommon, however, for baseline hormone levels associated with pathologic endocrine conditions to overlap with the normal range. In this circumstance, dynamic testing is useful to further separate the two groups. There are a multitude of dynamic endocrine tests, but all are based on principles of feedback regulation, and most responses can be remembered based on the pathways that govern endocrine axes. *Suppression tests* are used in the setting of suspected endocrine hyperfunction. An example is the dexamethasone suppression test used to evaluate Cushing's syndrome (Chaps. 328 and331). *Stimulation tests* are generally used to assess endocrine hypofunction. The ACTH stimulation test, for example, is used to assess the adrenal gland response in patients with suspected adrenal insufficiency. Other stimulation tests use hypothalamic-releasing factors such asTRH,GnRH,CRH, andGHRH to evaluate pituitary hormone reserve (Chap. 328). Insulin-induced hypoglycemia evokes pituitary ACTH and GH responses. Stimulation tests based on reduction or inhibition of endogenous hormones are less commonly used. Examples include metyrapone inhibition of cortisol synthesis and clomiphene inhibition of estrogen feedback.

Screening and Assessment of Common Endocrine Disorders Because many endocrine disorders are prevalent in the adult population (Table 327-4), most are diagnosed and managed by general internists, family practitioners, or other primary health care providers. The high prevalence and clinical impact of certain endocrine diseases justifies vigilance for features of these disorders during routine physical examinations; laboratory screening is indicated in selected high-risk populations.

(Bibliography omitted in Palm version)

Back to Table of Contents

328. DISORDERS OF THE ANTERIOR PITUITARY AND HYPOTHALAMUS - *Shlomo Melmed*

The anterior pituitary is often referred to as the "master gland" because, together with the hypothalamus, it orchestrates the complex regulatory functions of multiple other endocrine glands. The anterior pituitary gland produces six major hormones: (1) prolactin (PRL), (2) growth hormone (GH), (3) adrenocorticotropin hormone (ACTH), (4) luteinizing hormone (LH), (5) follicle-stimulating hormone (FSH), and (6) thyroid-stimulating hormone (TSH) (Table 328-1). Pituitary hormones are secreted in a pulsatile manner, reflecting stimulation by an array of specific hypothalamic releasing factors. Each of these pituitary hormones elicits specific responses in peripheral target tissues. The hormonal products of these peripheral glands, in turn, exert feedback control at the level of the hypothalamus and pituitary to modulate pituitary function (Fig. 328-1). Pituitary tumors cause characteristic hormone excess syndromes. Hormone deficiency may be inherited or acquired. Fortunately, efficacious treatments exist for the various pituitary hormone excess and deficiency syndromes. Nonetheless, these diagnoses are often elusive, emphasizing the importance of recognizing subtle clinical manifestations and performing the correct laboratory diagnostic tests.**For discussion of disorders of the posterior pituitary, or neurohypophysis, see Chap. 329.*

ANATOMY AND DEVELOPMENT

Anatomy The pituitary gland weighs ~600 mg and is located within the sella turcica ventral to the diaphragma sella; it comprises anatomically and functionally distinct anterior and posterior lobes. The sella is contiguous to vascular and neurologic structures, including the cavernous sinuses, cranial nerves, and optic chiasm. Thus, expanding intrasellar pathologic processes may have significant central mass effects in addition to their endocrinologic impact.

Hypothalamic neural cells synthesize specific releasing and inhibiting hormones that are secreted directly into the portal vessels of the pituitary stalk. Blood supply of the pituitary gland is derived from the superior and inferior hypophyseal arteries (Fig. 328-2). The hypothalamic-pituitary portal plexus provides the major blood source for the anterior pituitary, allowing reliable transmission of hypothalamic peptide pulses without significant systemic dilution; consequently, pituitary cells are exposed to sharp spikes of releasing factors and in turn release their hormones as discrete pulses (Fig. 328-3).

The posterior pituitary is supplied by the inferior hypophyseal arteries. In contrast to the anterior pituitary, the posterior lobe is directly innervated by hypothalamic neurons (supraopticohypophyseal and tuberohypophyseal nerve tracts) via the pituitary stalk (Chap. 329). Thus, posterior pituitary production of vasopressin (antidiuretic hormone; ADH) and oxytocin is particularly sensitive to neuronal damage by lesions that affect the pituitary stalk or hypothalamus.

Pituitary Development The embryonic differentiation and maturation of anterior pituitary cells have been elucidated in considerable detail. Pituitary development from Rathke's pouch involves a complex interplay of lineage-specific transcription factors expressed in pluripotential stem cells and gradients of locally produced growth factors (Table 328-1). The transcription factor Pit-1 determines cell-specific expression ofGH,PRL, andTSH in somatotropes, lactotropes, and thyrotropes. Expression of high levels of estrogen receptors in cells that contain Pit-1 favors PRL expression, whereas thyrotrope embronic factor (TEF) induces TSH expression. Pit-1 binds to GH, PRL, and TSH gene regulatory elements, as well as to recognition sites on its own promoter, providing a mechanism for perpetuating selective pituitary phenotypic stability. The transcription factor Prop-1 induces the pituitary development of Pit-1-specific lineages, as well as gonadotropes. Gonadotrope cell development is further defined by the cell-specific expression of the nuclear receptors, steroidogenic factor (SF-1) and DAX-1. Development of corticotrope cells, which express the proopiomelanocortin (POMC) gene, requires corticotropin upstream transcription element (CUTE) and the PTX-1 transcription factor. Abnormalities of pituitary development caused by mutations of Pit-1, Prop-1, SF-1, and DAX-1 result in a series of rare, selective or combined, pituitary hormone deficits.

HYPOTHALAMIC AND ANTERIOR PITUITARY INSUFFICIENCY

Hypopituitarism results from impaired production of one or more of the anterior pituitary trophic hormones. Reduced pituitary function can result from inherited disorders; more commonly, it is acquired and reflects the mass effects of tumors or the consequences of inflammation or vascular damage. These processes may also impair synthesis or secretion of hypothalamic hormones, with resultant pituitary failure (Table 328-2).

DEVELOPMENTAL AND GENETIC CAUSES OF HYPOPITUITARISM

Pituitary Dysplasia Pituitary dysplasia may result in aplastic, hypoplastic, or ectopic pituitary gland development. Because pituitary development requires midline cell migration from the nasopharyngeal Rathke's pouch, midline craniofacial disorders, such as cleft lip and palate, basal encephalocele, hypertelorism, and optic nerve hypoplasia, may be associated with pituitary dysplasia. Acquired pituitary failure in the newborn can also be caused by birth trauma, including cranial hemorrhage, asphyxia, and breech delivery.

Septo-optic Dysplasia Hypothalamic dysfunction and hypopituitarism may result from dysgenesis of the septum pellucidum or corpus callosum. Affected children have mutations in the *HESX1* gene, which is involved in early development of the ventral prosencephalon. These children exhibit cleft palate, syndactyly, ear deformities, hypertelorism, optic atrophy, micropenis, and anosmia. Pituitary dysfunction leads to diabetes insipidus,GHdeficiency and short stature, and, occasionally,TSHdeficiency.

Tissue-Specific Factor Mutations Several pituitary cell-specific transcription factors, such as Pit-1 and Prop-1, are critical for determining the development and function of specific anterior pituitary cell lineages. Autosomal dominant or recessive Pit-1 mutations cause combinedGH,PRL, andTSHdeficiencies. These patients present with growth failure and varying degrees of hypothyroidism. The pituitary may appear hypoplastic on magnetic resonance imaging (MRI).

Prop-1 is expressed early in pituitary development and appears to be required for Pit-1 function. Familial and sporadic *PROP1* mutations result in combinedGH,PRL,TSH, and gonadotropin deficiency, with preservation ofACTH. Over 80% of these patients have growth retardation and, by adulthood, all are deficient in TSH and gonadotropins. Because of gonadotropin deficiency, they do not enter puberty spontaneously (Fig. 328-4).

Developmental Hypothalamic Dysfunction

Kallmann Syndrome This syndrome results from defective hypothalamic gonadotropin-releasing hormone (GnRH) synthesis and is associated with anosmia or hyposmia due to olfactory bulb agenesis or hypoplasia (Chap. 335). The syndrome may also be associated with color blindness, optic atrophy, nerve deafness, cleft palate, renal abnormalities, cryptorchidism, and neurologic abnormalities such as mirror movements. Defects in the *KAL* gene, which maps to chromosome Xp22.3, prevent embryonic migration of GnRH neurons from the hypothalamic olfactory placode to the hypothalamus. Genetic abnormalities, in addition to *KAL* mutations, can also cause isolated GnRH deficiency, as autosomal recessive and dominant modes of transmission have been described. GnRH deficiency prevents progression through puberty. Males present with delayed puberty and pronounced hypogonadal features, including micropenis, probably the result of low testosterone levels during infancy (Chap. 335). Female patients present with primary amenorrhea and failure of secondary sexual development.

Kallmann syndrome and other causes of congenitalGnRHdeficiency are characterized by lowLH andFSHlevels and low concentrations of sex steroids (testosterone or estradiol). In sporadic cases of isolated gonadotropin deficiency, the diagnosis is often one of exclusion after eliminating other causes of hypothalamic-pituitary dysfunction. Repetitive GnRH administration restores normal pituitary gonadotropin responses, pointing to a hypothalamic defect.

Long-term treatment of males with human chorionic gonadotropin (hCG) or testosterone restores pubertal development and secondary sex characteristics; females can be treated with cyclic estrogen and progestin. Fertility may also be restored by the administration of subcutaneous, pulsatileGnRHusing a portable infusion pump.

Laurence-Moon-Bardet-Biedl Syndrome This rare autosomal recessive disorder is characterized by mental retardation; obesity; and hexadactyly, brachydactyly, or syndactyly. Central diabetes insipidus may or may not be associated.GnRHdeficiency occurs in 75% of males and half of affected females. Retinal degeneration begins in early childhood, and most patients are blind by age 30.

Frohlich Syndrome (Adipose Genital Dystrophy) A broad spectrum of hypothalamic lesions may be associated with hyperphagia, obesity, and central hypogonadism. DecreasedGnRHproduction in these patients results in attenuated pituitaryFSH andLHsynthesis and release.

Prader-Willi Syndrome Chromosome 15q deletions are associated with hypogonadotropic hypogonadism, hyperphagia-obesity, chronic muscle hypotonia, mental retardation, and adult-onset diabetes mellitus (Chap. 66). Multiple somatic defects also involve the skull, eyes, ears, hands, and feet. Diminished hypothalamic oxytocin- and vasopressin-producing nuclei have been reported.

DeficientGnRHsynthesis is suggested by the observation that chronic GnRH treatment restores pituitaryLH andFSHrelease.

ACQUIRED HYPOPITUITARISM

Hypopituitarism may be caused by accidental or neurosurgical trauma; vascular events such as apoplexy; pituitary or hypothalamic neoplasms such as pituitary adenomas, craniopharyngiomas, or metastatic deposits; inflammatory disease such as lymphocytic hypophysitis; infiltrative disorders such as sarcoidosis, hemochromatosis (Chap. 345), and tuberculosis; or irradiation. It is often difficult to localize the site of hormonal dysfunction as many processes, including hemochromatosis and radiation, may affect both hypothalamic and pituitary function.

Hypothalamic Infiltration Disorders These disorders -- including those associated with sarcoidosis, histiocytosis X, amyloidosis, and hemochromatosis -- frequently involve both hypothalamic and pituitary neuronal and neurochemical tracts. Consequently, diabetes insipidus occurs in half of patients with these disorders. Growth retardation is seen if attenuatedGHsecretion occurs before pubertal epiphyseal closure. Hypogonadotropic hypogonadism and hyperprolactinemia are also common.

Inflammatory Lesions Pituitary damage and subsequent dysfunction can be seen with chronic infections such as tuberculosis, opportunistic fungal infections associated with AIDS, and in tertiary syphilis. Other inflammatory processes, such as granulomas or sarcoidosis, may mimic a pituitary adenoma. These lesions may cause extensive hypothalamic and pituitary damage, leading to trophic hormone failure.

Cranial Irradiation Cranial irradiation may result in long-term hypothalamic and pituitary dysfunction, especially in children and adolescents who are more susceptible to damage following whole-brain or head and neck therapeutic irradiation. The development of hormonal abnormalities correlates strongly with irradiation dosage and the time interval after completion of radiotherapy. Up to two-thirds of patients ultimately develop hormone insufficiency afer a median dose of 50 Gy (5000 rad) directed at the skull base. The development of hypopituitarism occurs over 5 to 15 years and usually reflects hypothalamic damage rather than absolute destruction of pituitary cells. Though the pattern of hormone loss is variable,GHdeficiency is most commonly followed by gonadotropin andACTHdeficiency. When deficiency of one or more hormones is documented, the possibility of diminished reserve of other hormones is likely. Accordingly, anterior pituitary function should be evaluated over the long term in previously irradiated patients, and replacement therapy instituted when appropriate (see below).

Lymphocytic Hypophysitis This occurs mainly in pregnant or post-partum women; it usually presents with hyperprolactinemia andMRIevidence of a prominent pituitary mass resembling an adenoma, with mildly elevatedPRLlevels. Pituitary failure caused by diffuse lymphocytic infiltration may be transient or permanent but requires immediate evaluation and treatment. Rarely, isolated pituitary hormone deficiencies have been described, suggesting a selective autoimmune process targeted to specific cell types. Most patients manifest symptoms of progressive mass effects with headache and visual disturbance. The erythrocyte sedimentation rate is often elevated. As the MRI image may be indistinguishable from that of a pituitary adenoma, hypophysitis should be considered in a post-partum woman with a newly diagnosed pituitary mass before embarking on unnecessary surgical intervention. The inflammatory process often resolves after several months of glucocorticoid treatment, and pituitary function may be restored, depending on the extent of damage.

Pituitary Apoplexy Acute intrapituitary hemorrhagic vascular events can cause substantial damage to the pituitary and surrounding sellar structures. Pituitary apoplexy may occur spontaneously in a preexisting adenoma (usually nonfunctioning); postpartum (Sheehan's syndrome); or in association with diabetes, hypertension, sickle cell anemia, or acute shock. The hyperplastic enlargement of the pituitary during pregnancy increases the risk for hemorrhage and infarction. Apoplexy is an endocrine emergency that may result in severe hypoglycemia, hypotension, central nervous system (CNS) hemorrhage, and death. Acute symptoms include severe headache with signs of meningeal irritation, bilateral visual changes, ophthalmoplegia that varies, and, in severe cases, cardiovascular collapse and loss of consciousness. Pituitary computed tomography (CT) orMRI may reveal signs of intratumoral or sellar hemorrhage, with deviation of the pituitary stalk and compression of pituitary tissue.

Patients with no evident visual loss or impaired consciousness can be observed and managed conservatively with high-dose glucocorticoids. Those with significant or progressive visual loss or loss of consciousness require urgent surgical decompression. Visual recovery after surgery is inversely correlated with the length of time after the acute event. Therefore, severe ophthalmoplegia or visual deficits are indications for early surgery. Hypopituitarism is very common after apoplexy.

Empty Sella A partial or apparently totally empty sella is usually an incidentalMRIfinding. These patients usually exhibit normal pituitary function, implying that the surrounding rim of pituitary tissue is fully functional. Hypopituitarism, however, may develop insidiously. Pituitary masses may undergo clinically silent infarction with development of a partial or totally empty sella by cerebrospinal fluid (CSF) filling the dural herniation. Rarely, functional pituitary adenomas may arise within the rim of pituitary tissue, and these are not always visible on MRI.

PRESENTATION AND DIAGNOSIS

The clinical manifestations of hypopituitarism depend on which hormones are lost and the extent of the hormone deficiency.GHdeficiency causes growth disorders in children and leads to abnormal body composition in adults (see below). Gonadotropin deficiency causes menstrual disorders and infertility in women and decreased sexual function, infertility, and loss of secondary sexual characteristics in men.TSH andACTHdeficiency usually develop later in the course of pituitary failure. TSH deficiency leads to growth retardation in children and features of hypothyroidism in children and in adults. The secondary form of adrenal insufficiency caused by ACTH deficiency leads to hypocortisolism with relative preservation of mineralocorticoid production.PRLdeficiency causes failure of lactation. When lesions involve the posterior pituitary tracts, polyuria and polydipsia reflect loss of vasopressin secretion. Epidemiologic studies have documented an increased mortality rate in patients with longstanding pituitary damage, primarily from increased cardiovascular and cerebrovascular disease.

LABORATORY INVESTIGATION

Biochemical diagnosis of pituitary insufficiency is made by demonstrating low levels of trophic hormones in the setting of low target hormone levels. For example, low free thyroxine in the setting of a low or inappropriately normalTSHlevel suggests secondary hypothyroidism. Similarly, a low testosterone level without elevation of gonadotropins suggests hypogonadotropic hypogonadism. Provocative tests may be required to assess pituitary reserve (Table 328-3).GHresponses to insulin-induced hypoglycemia, arginine, L-dopa, growth hormone-releasing hormone (GHRH), or growth hormone-releasing peptides (GHRPs) can be used to assess GH reserve.PRL and TSH responses to thyrotropin-releasing hormone (TRH) reflect lactotrope and thyrotrope function. Corticotropin-releasing hormone (CRH) administration inducesACTHrelease, and administration of synthetic ACTH (cortrosyn) evokes adrenal cortisol release as an indirect indicator of pituitary ACTH reserve (Chap. 331). ACTH reserve is most reliably assessed during insulin-induced hypoglycemia. However, this test should be performed cautiously in patients with suspected adrenal insufficiency because of increased risk of hypoglycemia and hypotension.

TREATMENT

Hormone replacement therapy, including glucocorticoids, thyroid hormone, sex steroids, growth hormone, and vasopressin, is usually free of complications. Treatment regimens that mimic physiologic hormone production allow for maintenance of satisfactory clinical homeostasis. Effective dosage schedules are outlined inTable 328-4. Patients in need of glucocorticoid replacement require careful dose adjustments during stressful events such as acute illness, dental procedures, trauma, and acute hospitalization (Chap. 331).

HYPOTHALAMIC, PITUITARY, AND OTHER SELLAR MASSES

PITUITARY TUMORS

Pituitary adenomas are the most common cause of pituitary hormone hypersecretion and hyposecretion syndromes in adults. They account for ~10% of all intracranial neoplasms. At autopsy, up to a quarter of all pituitary glands harbor an unsuspected microadenoma (<10 mm diameter). Similarly, pituitary imaging detects small pituitary lesions in at least 10% of normal individuals.

Pathogenesis Pituitary adenomas are benign neoplasms that arise from one of the five anterior pituitary cell types. The clinical and biochemical phenotype of pituitary adenomas depend on the cell type from which they are derived and are described in detail below. Thus, tumors arising from lactotrope (PRL), somatotrope (GH), corticotrope (ACTH), thyrotrope (TSH), or gonadotrope (LH,FSH) cells hypersecrete their respective hormones (Table 328-5). Plurihormonal tumors that express combinations of GH, PRL, TSH, ACTH, and the glycoprotein hormonea subunit may be diagnosed by careful immunocytochemistry or may, in fact, present with mixed clinical features of these hormonal hypersecretory syndromes. Morphologically, these tumors may arise from a single polysecreting cell type or consist of cells with mixed function within the same tumor.

Hormonally active tumors are characterized by autonomous hormone secretion with diminished responsiveness to the normal physiologic pathways of inhibition. Hormone production does not always correlate with tumor size. Small hormone-secreting adenomas may cause significant clinical perturbations, whereas larger adenomas that produce less hormone may be clinically silent and remain undiagnosed (if no central compressive effects occur). About one-third of all adenomas are clinically nonfunctioning and produce no distinct clinical hypersecretory syndrome. Most arise from gonadotrope cells and may secrete a- andb-glycoprotein hormone subunits or, very rarely, intact circulating gonadotropins. True pituitary carcinomas with documented extracranial metastases are exceedingly rare.

Almost all pituitary adenomas are monoclonal in origin, implying the acquisition of one or more somatic mutations that confer a selective growth advantage. In addition to direct studies of oncogene mutations, this idea is supported by X-chromosomal inactivation analyses of tumors in female patients heterozygous for X-linked genes. Consistent with their clonal origin, complete surgical resection of small pituitary adenomas usually cures hormone hypersecretion. Nevertheless, hypothalamic hormones, such asGHRH orCRH, also enhance the mitotic activity of their respective pituitary target cells, in addition to their role in pituitary hormone regulation. Thus, patients harboring rare abdominal or chest tumors elaborating ectopic GHRH or CRH may present with somatotrope or corticotrope hyperplasia.

Several etiologic genetic events have been implicated in the development of pituitary tumors. The pathogenesis of sporadic forms of acromegaly has been particularly informative as a model of tumorigenesis.GHRH, after binding to its G protein-coupled somatotrope receptor, utilizes cyclic AMP as a second messenger to stimulateGHsecretion and somatotrope proliferation. A subset (~35%) of GH-secreting pituitary tumors contain mutations in Gsa (Arg 201® Cys or His; Gln 227® Arg). These mutations inhibit intrinsic GTPase activity, resulting in constitutive elevation of cyclic AMP, Pit-1 induction, and activation of cyclic AMP response element binding protein (CREB), thereby promoting somatotrope cell proliferation.

Characteristic loss of heterozygosity (LOH) in various chromosomes has been documented in large or invasive macroadenomas, suggesting the presence of putative tumor suppressor genes at these loci. LOH of chromosome region on 11q13, 13, and 9 is present in up to 20% of sporadic pituitary tumors includingGH-,PRL-, andACTH-producing adenomas and in some nonfunctioning tumors.

Compelling evidence also favors growth factor promotion of pituitary tumor proliferation. Basic fibroblast growth factor (bFGF) is abundant in the pituitary and has been shown to stimulate pituitary cell mitogenesis. Other factors involved in initiation and promotion of pituitary tumors include loss of negative-feedback inhibition (as seen with primary hypothyroidism or hypogonadism) and estrogen-mediated or paracrine angiogenesis. Growth characteristics and neoplastic behavior may also be influenced by several activated oncogenes, including *RAS* and pituitary tumor transforming gene (*PTTG*).

Genetic Syndromes Associated with Pituitary Tumors Several familial syndromes are associated with pituitary tumors, and the genetic mechanisms for some of these have been unraveled (Table 328-6).

Multiple endocrine neoplasia (MEN) 1 is an autosomal dominant syndrome characterized primarily by a genetic predisposition to parathyroid, pancreatic islet, and pituitary adenomas (Chap. 339). MEN-1 is caused by inactivating germline mutations in *MENIN*, a constitutively expressed tumor-suppressor gene located on chromosome 11q13. Loss of heterozygosity, or a somatic mutation of the remaining normal *MENIN* allele, leads to tumorigenesis. About half of affected patients develop prolactinomas; acromegaly and Cushing's syndrome are less commonly encountered.

Carney syndrome is characterized by spotty skin pigmentation, myxomas, and endocrine tumors including testicular, adrenal, and pituitary adenomas. Acromegaly occurs in about 20% of patients. This autosomal dominant syndrome is associated with microsatellite alterations on chromosome 2p16.

McCune-Albright syndrome consists of polyostotic fibrous dysplasia, pigmented skin patches, and a variety of endocrine disorders, includingGH-secreting pituitary tumors, adrenal adenomas, and autonomous ovarian function (Chap. 343). Hormonal hypersecretion is due to constitutive cyclic AMP production caused by inactivation of the GTPase activity of Gsa. The Gsa mutations occur postzygotically, leading to a mosaic pattern of mutant expression.

Familial acromegaly is a rare disorder in which family members may manifest either acromegaly or gigantism. The disorder is associated withLOH at a chromosome 11q13 locus distinct from that of *MENIN*.

OTHER SELLAR MASSES

Craniopharyngiomas are derived from Rathke's pouch. They arise near the pituitary stalk and commonly extend into the suprasellar cistern. These tumors are often large, cystic, and locally invasive. Many are partially calcified, providing a characteristic appearance on skull x-ray andCTimages. More than half of all patients present before age 20, usually with signs of increased intracranial pressure, including headache, vomiting, papilledema, and hydrocephalus. Associated symptoms include visual field abnormalities, personality changes and cognitive deterioration, cranial nerve damage, sleep difficulties, and weight gain. Anterior pituitary dysfunction and diabetes insipidus are common. About half of affected children present with growth retardation.

Treatment usually involves transcranial or transsphenoidal surgical resection followed by postoperative radiation of residual tumor. This approach can result in long-term survival and ultimate cure, but most patients require lifelong pituitary hormone replacement. If the pituitary stalk is uninvolved and can be preserved at the time of surgery, the incidence of subsequent anterior pituitary dysfunction is significantly diminished.

Developmental failure of Rathke's pouch obliteration may lead to *Rathke's cysts*, which are small (<5 mm) cysts entrapped by squamous epithelium; these cysts are found in about 20% of individuals at autopsy. Although Rathke's cleft cysts do not usually grow and are often diagnosed incidentally, about a third present in adulthood with compressive symptoms, diabetes insipidus, and hyperprolactinemia due to stalk compression. Rarely, internal hydrocephalus develops. The diagnosis is suggested preoperatively by visualizing the cyst wall onMRI, which distinguishes these lesions from craniopharyngiomas. Cyst contents range fromCSF-like fluid to mucoid material. *Arachnoid cysts* are rare and generate an MRI image isointense with cerebrospinal fluid.

Sella chordomas usually present with bony clival erosion, local invasiveness, and, on occasion, calcification. Normal pituitary tissue may be visible onMRI, distinguishing chordomas from aggressive pituitary adenomas. Mucinous material may be obtained by fine-needle aspiration.

Meningiomas arising in the sellar region may be difficult to distinguish from nonfunctioning pituitary adenomas. OnMRI they may be asymmetric, and onCT they may show evidence of bony erosion. Meningiomas may cause compressive symptoms.

Histiocytosis X comprises a variety of syndromes associated with foci of eosinophilic granulomas. Diabetes insipidus, exophthalmos, and punched-out lytic bone lesions (*Hand-Schuller-Christian disease*) are associated with granulomatous lesions visible onMRI, as well as a characteristic axillary skin rash. Rarely, the pituitary stalk may be involved.

Pituitary metastases occur in ~3% of cancer patients. Blood-borne metastatic deposits are found almost exclusively in the posterior pituitary. Accordingly, diabetes insipidus can be a presenting feature of lung, gastrointestinal, breast, and other pituitary metastases. About half of pituitary metastases originate from breast cancer; about 25% of patients with breast cancer have such deposits. Rarely, pituitary stalk involvement results in anterior pituitary insufficiency. TheMRIdiagnosis of a metastatic lesion may be difficult to distinguish from an aggressive pituitary adenoma; the diagnosis may require histologic examination of excised tumor tissue. Primary or metastatic lymphoma, leukemias, and plasmacytomas also occur within the sella.

Hypothalamic hamartomas and *gangliocytomas* may arise from astrocytes, oligodendrocytes, and neurons with varying degrees of differentiation. These tumors may overexpress hypothalamic neuropeptides includingGnRH,GHRH, orCRH. In GnRH-producing tumors, children present with precocious puberty, psychomotor delay, and laughing-associated seizures. Medical treatment of GnRH-producing hamartomas with long-acting GnRH analogues effectively suppresses gonadotropin secretion and controls pubertal development. Rarely, hamartomas are also associated with craniofacial abnormalities; imperforate anus; cardiac, renal, and lung disorders; and pituitary failure (*Pallister-Hall syndrome*). Hypothalamic hamartomas are often contiguous with the pituitary, and preoperativeMRIdiagnosis may not be possible. Histologic evidence of hypothalamic neurons in tissue resected at transsphenoidal surgery may be the first indication of a primary hypothalamic lesion.

Hypothalamic gliomas and *optic gliomas* occur mainly in childhood and usually present with visual loss. Adults have more aggressive tumors; about a third are associated with neurofibromatosis.

Brain germ-cell tumors may arise within the sellar region. These include *dysgerminomas*, which are associated with diabetes insipidus and visual loss and rarely metastasize. *Germinomas*, *embryonal carcinomas*, *teratomas*, and *choriocarcinomas* may arise in the parasellar region and producehCG. These germ-cell tumors present with precocious puberty, diabetes insipidus, visual field defects, and thirst disorders. Many patients areGH-deficient with short stature.

METABOLIC EFFECTS OF HYPOTHALAMIC LESIONS

The hypothalamus is subject to injury from mass lesions, granulomatous disorders, infections, and hemorrhage. Lesions involving the anterior and preoptic hypothalamic regions cause paradoxical vasoconstriction, tachycardia, and hyperthermia. Acute hyperthermia is usually due to a hemorrhagic insult, but poikilothermia may also occur. Central disorders of thermoregulation result from posterior hypothalamic damage. The *periodic hypothermia syndrome* comprises episodic attacks of rectal temperatures<30°C, sweating, vasodilation, vomiting, and bradycardia (Chap. 20). Damage to the ventromedial nuclei by craniopharyngiomas, hypothalamic trauma, or inflammatory disorders may be associated with *hyperphagia* and *obesity*. This region appears to contain an energy-satiety center where melanocortin receptors are influenced by leptin, insulin,POMCproducts, and gastrointestinal peptides (Chap. 77). Median eminence involvement results in diabetes insipidus in about 50% of patients. Hypothalamic gliomas in early childhood may be associated with a diencephalic syndrome characterized by progressive severe emaciation and growth failure. Polydipsia or hypodipsia are associated with damage to central osmo-receptors located in preoptic nuclei (Chap. 329). Slow-growing hypothalamic lesions can cause increased somnolence and disturbed sleep cycles as well as obesity, hypothermia, and emotional outbursts. Lesions of the central hypothalamus may stimulate sympathetic neurons, leading to elevated serum catecholamine and cortisol levels. These patients are predisposed to cardiac arrhythmias, hypertension, and gastric erosions.

EVALUATION

Local Mass Effects Clinical manifestations of sellar lesions vary, depending on the anatomic location of the mass and direction of its extension (Table 328-7). The dorsal roof of the sella presents the least resistance to soft tissue expansion from within the confines of the sella; consequently, pituitary adenomas frequently extend in a suprasellar direction. Bony invasion may ultimately occur as well.

Headaches are common features of small intrasellar tumors, even with no demonstrable suprasellar extension. Because of the confined nature of the pituitary, small changes in intrasellar pressure stretch the dural plate; however, the severity of the headache correlates poorly with adenoma size or extension.

Suprasellar extension can lead to visual loss by several mechanisms, the most common being compression of the optic chiasm, but direct invasion of the optic nerves or obstruction ofCSF flow leading to secondary visual disturbances also occur. Pituitary stalk compression by a hormonally active or inactive intrasellar mass may compress the portal vessels, disrupting pituitary access to the hypothalamic hormones and dopamine; this results in hyperprolactinemia and concurrent loss of other pituitary hormones. This "stalk section" phenomenon may also be caused by trauma, whiplash injury with posterior clinoid stalk compression, or skull base fractures. Lateral mass invasion may impinge on the cavernous sinus and compress its neural contents, leading to cranial nerve III, IV, and VI palsies as well as effects on the ophthalmic and maxillary branches of the fifth cranial nerve (Chap. 367). Patients may present with diplopia, ptosis, ophthalmoplegia, and decreased facial sensation, depending on the extent of neural damage. Extension into the sphenoid sinus indicates that the pituitary mass has eroded through the sellar floor. Aggressive tumors may also invade the palate roof and cause nasopharyngeal obstruction, infection, and, rarely, CSF leakage. Both temporal and frontal lobes may be invaded, leading to uncinate seizures, personality disorders, and anosmia. Direct hypothalamic encroachment by an invasive pituitary mass may cause important metabolic sequelae, precocious puberty or hypogonadism, diabetes insipidus, sleep disturbances, dysthermia, and appetite disorders.

MRISagittal and coronal T1-weighted spin-echo MRI imaging, before and after administration of gadolinium, allow precise visualization of the pituitary gland with clear delineation of the hypothalamus, pituitary stalk, pituitary tissue and surrounding suprasellar cisterns, cavernous sinuses, sphenoid sinus, and optic chiasm. Pituitary gland height ranges from 6 mm in children to 8 mm in adults; during pregnancy and puberty, the height may reach 10 to 12 mm. The upper aspect of the adult pituitary is flat or slightly concave, but in adolescent and pregnant individuals, this surface may be convex, reflecting physiologic pituitary enlargement. The stalk should be vertical.CT scan is indicated to define the extent of bony erosion or the presence of calcification.

The soft tissue consistency of the pituitary gland is slightly heterogeneous onMRI. Anterior pituitary signal intensity resembles that of brain matter on T1-imaging (Fig. 328-5). Adenoma density is usually lower than that of surrounding normal tissue on T1-weighted imaging, and the signal intensity increases with T2-weighted images. The high phospholipid content of the posterior pituitary results in a bright enhancing signal.

Sellar masses are commonly encountered as incidental findings onMRI, and most of these are pituitary adenomas (incidentalomas). This finding is consistent with the observation that clinically silent pituitary microadenomas can be identified in up to 25% of pituitaries in autopsy series. In the absence of hormone hypersecretion, these small lesions can be safely monitored by MRI, which is performed annually and then less often if there is no evidence of growth. Resection should be considered for incidentally discovered macroadenomas, as about one-third become invasive or cause local pressure effects. If hormone hypersecretion is evident, specific therapies are indicated. When larger masses (>1 cm) are encountered, they should also be distinguished from nonadenomatous lesions. Meningiomas are often associated with bony hyperostosis; craniopharyngiomas may be calcified and are usually hypodense, whereas gliomas are hyperdense on T2-weighted images.

Ophthalmologic Evaluation Because optic tracts may be contiguous to an expanding pituitary mass, reproducible visual field assessment that uses perimetry techniques should be performed on all patients with sellar mass lesions that abut the optic chiasm. Loss of red perception is an early sign of optic tract pressure. Bitemporal hemianopia or superior bitemporal defects are classically observed, reflecting the location of these tracts within the inferior and posterior part of the chiasm. Early diagnosis reduces the risk of blindness, scotomas, or other visual disturbances.

Laboratory Investigation The presenting clinical features of functional pituitary adenomas (e.g., acromegaly, prolactinomas, or Cushing's disease) should guide the laboratory studies (see below). However, for a sellar mass with no obvious clinical features of hormone excess, laboratory studies are geared towards determining the nature of the tumor and assessing the possible presence of hypopituitarism. When a pituitary adenoma is suspected based onMRI, initial hormonal evaluation usually includes: (1) basalPRL; (2) insulin-like growth factor (IGF) I; (3) 24-h urinary free cortisol (UFC) and/or overnight oral dexamethasone (1 mg) suppression test; (4)a-subunit,FSH, andLHlevels; and (5) thyroid function tests. Additional hormonal evaluation may be indicated based on the results of these tests. Pending more detailed assessment of hypopituitarism, a menstrual history, testosterone level, 8 A.M. cortisol, and thyroid function tests usually identify patients with pituitary hormone deficiencies that require hormone replacement before further testing or surgery.

Histologic Evaluation Immunohistochemical staining of pituitary tumor specimens obtained at transsphenoidal surgery confirm clinical and laboratory studies and provide a histologic diagnosis when hormone studies are equivocal and in cases of clinically nonfunctioning tumors. Occasionally, ultrastructural assessment by electron microscopy is required for diagnosis.

TREATMENT

Overview Successful management of sellar masses requires accurate diagnosis as well as selection of optimal therapeutic modalities. Most pituitary tumors are benign and slow-growing. Clinical features result from local mass effects and hormonal hypo- or hypersecretion syndromes caused directly by the adenoma or as a consequence of treatment. Thus, lifelong management and follow-up are necessary for these patients.

ImprovedMRItechnology with gadolinium enhancement for pituitary visualization, new advances in transsphenoidal surgery and in stereotactic radiotherapy (including gamma-knife radiotherapy), and novel therapeutic agents have improved pituitary tumor management. The goals of pituitary tumor treatment include normalization of excess pituitary secretion, amelioration of symptoms and signs of hormonal hypersecretion syndromes, and shrinkage or ablation of large tumor masses with relief of adjacent structure compression. Residual anterior pituitary function should be preserved and can sometimes be restored by removing tumor mass. Ideally, adenoma recurrence should be prevented.

Transsphenoidal Surgery Transsphenoidal rather than transfrontal resection is the desired surgical approach for pituitary tumors, except for the rare invasive suprasellar mass surrounding the frontal or middle fossa, the optic nerves, or invading posteriorly behind the clivus. Intraoperative microscopy facilitates visual distinction between adenomatous and normal pituitary tissue, as well as microdissection of small tumors that may not be visible byMRI(Fig. 328-6). Transsphenoidal surgery also avoids the cranial invasion and manipulation of brain tissue required by subfrontal surgical approaches. Endoscopic techniques with three-dimensional intraoperative localization have improved visualization and access to tumor tissue. The endoscopic approach is also less traumatic, as the technique is endonasal and does not require a transsphenoidal retractor.

In addition to correction of hormonal hypersecretion, pituitary surgery is indicated for mass lesions that impinge on surrounding structures. Surgical decompression and resection are required for an expanding pituitary mass accompanied by pesistent headache, progressive visual field defects, cranial nerve palsies, internal hydrocephalus, and, occasionally, intrapituitary hemorrhage and apoplexy. Repeat surgery may be required for persistent postoperativeCSFleakage. Transsphenoidal surgery is sometimes used for pituitary tissue biopsy and histologic diagnosis.

Whenever possible, the pituitary mass lesion should be selectively excised; normal tissue should be manipulated or resected only when critical for effective dissection. Nonselective hemihypophysectomy or total hypophysectomy may be indicated if no mass lesion is clearly discernible, multifocal lesions are present, or the remaining nontumorous pituitary tissue is obviously necrotic. This strategy increases the likelihood of hypopituitarism and the need for lifelong hormonal replacement.

Preoperative local compression signs, including visual field defects or compromised pituitary function, may be reversed by surgery, particularly when these deficits are not long-standing. For large and invasive tumors, it is necessary to determine the optimal balance between maximal tumor resection and preservation of anterior pituitary function, especially for preserving growth and reproductive function in younger patients. Similarly, tumor invasion outside of the sella is rarely amenable to surgical cure; the surgeon must judge the risk:benefit ratio of extensive tumor resection.

Side Effects Tumor size and the degree of invasiveness largely determine the incidence of surgical complications. Operative mortality is about 1%. Transient diabetes insipidus and hypopituitarism occur in up to 20% of patients. Permanent diabetes insipidus, cranial nerve damage, nasal septal perforation, or visual disturbances may be encountered in up to 10% of patients.CSFleaks occur in 4% of patients. Less common complications include carotid artery injury, loss of vision, hypothalamic damage, and meningitis. Permanent side effects are rarely encountered after surgery for microadenomas.

Radiation Radiation is used either as a primary therapy for pituitary or parasellar masses or, more commonly, as an adjunct to surgery or medical therapy. Focused megavoltage irradiation is achieved by preciseMRIlocalization, using a high-voltage linear accelerator and accurate isocentric rotational arcing. A major determinant of accurate irradiation is to reproduce the patient's head position during multiple visits and to maintain absolute head immobility. A total of <50 Gy (5000 rad) is given as 180-cGy (180 rad) fractions split over about 6 weeks. Stereotactic radiosurgery delivers a large single high-energy dose from a cobalt 60 source (gamma knife), linear accelerator, or cyclotron. Long-term effects of gamma-knife surgery are as yet unknown.

The role of radiation therapy in pituitary tumor management depends on multiple factors including the nature of the tumor, age of the patient, and the availability of surgical and radiation expertise. Because of its relatively slow onset of action, radiation therapy is usually reserved for postsurgical management. As an adjuvant to surgery, radiation is used to treat residual tumor and in an attempt to prevent regrowth. Irradiation offers the