INTRODUCTION

To a medical student who requires hours to collect a patient’s history, perform a physical examination, and organize that information into a coherent presentation, an experienced clinician’s ability to decide on a diagnosis and management plan in minutes may seem extraordinary. What separates the master clinician from the novice is an elusive quality called “expertise.” The first part of this chapter provides an overview of our current understanding of expertise in clinical reasoning, what it is, and how it can be developed.

The proper use of diagnostic tests and the integration of the results into the patient’s clinical assessment may also be equally bewildering to students. Hoping to hit the unknown diagnostic target, novice medical practitioners typically apply a “shotgun” approach to testing. The expert, in contrast, usually focuses her testing strategy to specific diagnostic hypotheses. The second part of the chapter reviews basic statistical concepts useful for interpreting diagnostic tests and quantitative tools useful for clinical decision-making.

Evidence-based medicine (EBM) constitutes the integration of the best available research evidence with clinical judgment as applied to the care of individual patients. The third part of the chapter provides an overview of the tools of EBM.

BRIEF INTRODUCTION TO CLINICAL REASONING

Clinical Expertise

Defining “clinical expertise” remains surprisingly difficult. Chess has an objective ranking system based on skill and performance criteria. Athletics, similarly, have ranking systems to distinguish novices from Olympians. But in medicine, after physicians complete training and pass the boards, no further tests or benchmarks identify those who have attained the highest levels of clinical performance. Of course, physicians often consult a few “elite” clinicians for their “special problem-solving prowess” when particularly difficult or obscure cases have baffled everyone else. Yet despite their skill, even master clinicians typically cannot explain their exact processes and methods, thereby limiting the acquisition and dissemination of the expertise used to achieve their impressive results. Furthermore, clinical virtuosity appears not to be generalizable, e.g., an expert on hypertrophic cardiomyopathy may be no better (and possibly worse) than a first-year medical resident at diagnosing and managing a patient with neutropenia, fever, and hypotension.
Broadly construed, clinical expertise includes not only cognitive dimensions and the integration of verbal and visual cues or information but also complex fine-motor skills necessary for invasive and noninvasive procedures and tests. In addition, “the complete package” of expertise in medicine includes the ability to communicate effectively with patients and work well with members of the medical team. Research on medical expertise remains relatively sparse overall, with most of the work focused on diagnostic reasoning, and much less work focused on treatment decisions or the technical skills involved in the performance of procedures. Thus, in this chapter, we focus primarily on the cognitive elements of clinical reasoning.

Because clinical reasoning takes place in the heads of doctors, it is therefore not readily observable, making it obviously difficult to study. One method of research on reasoning asks doctors to “think out loud” as they receive increments of clinical information in a manner meant to simulate a clinical encounter. Another research approach has focused on how doctors should reason diagnostically rather than on how they actually do reason. Much of what is known about clinical reasoning comes from empirical studies of nonmedical problem-solving behavior. Because of the diverse perspectives contributing to this area, with important contributions from cognitive psychology, sociology, medical education, economics, informatics, and decision sciences, no single integrated model of clinical reasoning exists, and not infrequently, different terms and models describe similar phenomena.

**Intuitive versus Analytic Reasoning**

A contemporary model of reasoning, dual-process theory distinguishes two general systems of cognitive processes. *Intuition* (System 1) provides rapid effortless judgments from memorized associations using pattern recognition and other simplifying “rules of thumb” (i.e., heuristics). For example, a very simple pattern that could be useful in certain situations is “African-American women plus hilar adenopathy equals sarcoid.” Because no effort is involved in recalling the pattern, typically, the clinician is unable to say how those judgments were formulated. In contrast, *analysis* (System 2), the other form of reasoning in the dual-process model, is slow, methodical, deliberative, and effortful. These are, of course, idealized extremes of the cognitive continuum. How these systems interact in different decision problems, how experts use them differently from novices, and when their use can lead to errors in judgment remain the subject of considerable study and debate.

*Pattern recognition* is a complex cognitive process that appears largely effortless. One can recognize people’s faces, the breed of a dog, or an automobile model without necessarily being able to say what specific features prompted the recognition. Analogously, experienced clinicians often recognize familiar diagnosis patterns quickly. In the absence of an extensive stored repertoire of diagnostic patterns, students (as well as more experienced clinicians operating outside their area of expertise) often use the more laborious System 2 analytic approach along with more intensive and comprehensive data collection to reach the diagnosis.

The following three brief scenarios of a patient with hemoptysis demonstrate three distinct patterns:

A 46-year-old man presents to his internist with a chief complaint of hemoptysis. An otherwise healthy nonsmoker, he is recovering from an apparent viral bronchitis. This presentation pattern suggests that the
small amount of blood-streaked sputum is due to acute bronchitis, so that a chest x-ray provides sufficient reassurance that a more serious disorder is absent.

In the second scenario, a 46-year-old patient who has the same chief complaint but with a 100-pack-year smoking history, a productive morning cough, and episodes of blood-streaked sputum fits the pattern of carcinoma of the lung. Consequently, along with the chest x-ray, the physician obtains a sputum cytology examination and refers this patient for a chest computed tomography (CT) scan.

In the third scenario, a 46-year-old patient with hemoptysis who immigrated from a developing country has an echocardiogram as well, because the physician hears a soft diastolic rumbling murmur at the apex on cardiacl auscultation, suggesting rheumatic mitral stenosis and possibly pulmonary hypertension.

Although rapid, pattern recognition used without sufficient reflection can result in premature closure: mistakenly concluding that one already knows the correct diagnosis and therefore failing to complete the data collection that would demonstrate the lack of fit of the initial pattern selected. For example, a 45-year-old man presents with a 3-week history of a “flu-like” upper respiratory infection (URI) including symptoms of dyspnea and a productive cough. On the basis of the presenting complaints, the clinician uses a “URI assessment form” to improve the quality and efficiency of care by standardizing the information gathered. After quickly acquiring the requisite structured examination components and noting in particular the absence of fever and a clear chest examination, the physician prescribes medication for acute bronchitis and sends the patient home with the reassurance that his illness was not serious. Following a sleepless night with significant dyspnea, the patient develops nausea and vomiting and collapses. He presents to the emergency department in cardiac arrest and is unable to be resuscitated. His autopsy shows a posterior wall myocardial infarction and a fresh thrombus in an atherosclerotic right coronary artery. What went wrong? The clinician had decided, based on the patient’s appearance, even before starting the history, that the patient’s complaints were not serious. Therefore, he felt confident that he could perform an abbreviated and focused examination by using the URI assessment protocol rather than considering the broader range of possibilities and performing appropriate tests to confirm or refute his initial hypotheses. In particular, by concentrating on the URI, the clinician failed to elicit the full dyspnea history, which would have suggested a far more serious disorder, and he neglected to search for other symptoms that could have directed him to the correct diagnosis.

Heuristics, also referred to as cognitive shortcuts or rules of thumb, are simplifying decision strategies that ignore part of the data available so as to provide an efficient path to the desired judgment. They are generally part of the intuitive system tools. Two major research programs have come to different conclusions about the value of heuristics in clinical judgment. The “heuristics and biases” program focused on understanding how heuristics in problem solving could be biased by testing the numerical intuition of psychology undergraduates against the rules of statistics. In contrast, the “fast and frugal heuristics” research program explored how and when decision makers’ reliance on simple heuristics can produce good decisions. Although many heuristics have relevance to clinical reasoning, only four will be mentioned here.
When assessing a particular patient, clinicians often weigh the similarity of that patient’s symptoms, signs, and risk factors against those of their mental representations of the diagnostic hypotheses being considered. In other words, among the diagnostic possibilities, clinicians identify the diagnosis for which the patient appears to be a representative example. Analogous to pattern recognition, this cognitive shortcut is called the *representativeness heuristic*. However, physicians using the representativeness heuristic can reach erroneous conclusions if they fail to consider the underlying prevalence (i.e., the prior, or pretest, probabilities) of the two competing diagnoses that could explain the patient’s symptoms. Consider a patient with hypertension and headache, palpitations, and diaphoresis. Inexperienced clinicians might judge pheochromocytoma to be quite likely based on the representativeness heuristic with this classic symptom triad suggesting pheochromocytoma. Doing so would be incorrect given that other causes of hypertension are much more common than pheochromocytoma, and this triad of symptoms can occur in patients who do not have pheochromocytoma. Less experience with a particular diagnosis and with the breadth of presentations (e.g., diseases that affect multiple organ systems such as sarcoid) may also lead to errors.

A second commonly used cognitive shortcut, the *availability heuristic*, involves judgments based on how easily prior similar cases or outcomes can be brought to mind. For example, an experienced clinician may recall 20 elderly patients seen over the last few years who presented with painless dyspnea of acute onset and were found to have acute myocardial infarction (MI). A novice clinician may spend valuable time seeking a pulmonary cause for the symptoms before considering and then confirming the cardiac diagnosis. In this situation, the patient’s clinical pattern does not fit the most common pattern of acute MI, but experience with this atypical presentation, along with the ability to recall it, directs the physician to the diagnosis.

Errors with the availability heuristic arise from several sources of recall bias. Rare catastrophes are likely to be remembered with a clarity and force disproportionate to their likelihood for future diagnosis—for example, a patient with a sore throat eventually found to have leukemia or a young athlete with leg pain eventually found to have a sarcoma—and those publicized in the media or that are recent experiences are, of course, easier to recall and therefore more influential on clinical judgments.

The third commonly used cognitive shortcut, the *anchoring heuristic* (also called conservatism or stickiness), involves estimating a probability of disease (the anchor) and then insufficiently adjusting that probability up or down (compared with Bayes’ rule) when interpreting new data about the patient, i.e., sticking to their initial diagnosis. For example, a clinician may still judge the probability of coronary artery disease (CAD) to be high after a negative exercise thallium test and proceed to cardiac catheterization (see “Measures of Disease Probability and Bayes’ Rule,” below).

The fourth heuristic states that clinicians should use the simplest explanation possible that will account adequately for the patient’s symptoms or findings (Occam’s razor or, alternatively, the *simplicity heuristic*). Although this is an attractive and often used principle, it is important to remember that no biologic basis for it exists. Errors from the simplicity heuristic include premature closure leading to the neglect of unexplained significant symptoms or findings.
Even experienced physicians use analytic reasoning processes (System 2) when the problem they face is recognized to be complex or to involve important unfamiliar elements or features. In such situations, clinicians proceed much more methodically in what has been referred to as the hypothetico-deductive model of reasoning. From the outset, expert clinicians working analytically generate, refine, and discard diagnostic hypotheses. The hypotheses drive questions asked during history taking and may change based on the working hypotheses of the moment. Even the physical examination is focused by the working hypotheses. Is the spleen enlarged? How big is the liver? Is it tender? Are there any palpable masses or nodules? Each question must be answered (with the exclusion of all other inputs) before the examiner can move on to the next specific question. Each diagnostic hypothesis provides testable predictions and sets a context for the next question or step to follow. For example, if the enlarged and quite tender liver felt on physical examination is due to acute hepatitis (the hypothesis), certain specific liver function tests should be markedly elevated (the prediction). If the tests come back normal, the hypothesis may have to be discarded or substantially modified.

Negative findings often are neglected but are as important as positive ones because they often reduce the likelihood of the diagnostic hypotheses under consideration. Chest discomfort that is not provoked or worsened by exertion in an active patient reduces the likelihood that chronic ischemic heart disease is the underlying cause. The absence of a resting tachycardia and thyroid gland enlargement reduces the likelihood of hyperthyroidism in a patient with paroxysmal atrial fibrillation.

The acuity of a patient’s illness may override considerations of prevalence and the other issues described above. “Diagnostic imperatives” recognize the significance of relatively rare but potentially catastrophic diagnoses if undiagnosed and untreated. For example, clinicians are taught to consider aortic dissection routinely as a possible cause of acute severe chest discomfort. Even though the typical history of dissection differs from that of MI, dissection is far less prevalent, so diagnosing dissection remains challenging unless it is explicitly and routinely considered as a diagnostic imperative (Chap. 301). If the clinician fails to elicit any of the characteristic features of dissection by history and finds equivalent blood pressures in both arms and no pulse deficits, he may feel comfortable discarding the aortic dissection hypothesis. If, however, the chest x-ray shows a possible widened mediastinum, the hypothesis may be reinstated and an appropriate imaging test ordered (e.g., thoracic CT scan, transesophageal echocardiogram) to evaluate more fully. In nonacute situations, the prevalence of potential alternative diagnoses should play a much more prominent role in diagnostic hypothesis generation.

Cognitive scientists studying the thought processes of expert clinicians have observed that clinicians group data into packets, or “chunks,” that are stored in short-term or “working memory” and manipulated to generate diagnostic hypotheses. Because short-term memory can typically retain only 5–9 items at a time, the number of packets that can be actively integrated into hypothesis-generating activities is similarly limited. For this reason, the cognitive shortcuts discussed above play a key role in the generation of diagnostic hypotheses, many of which are discarded as rapidly as they are formed (thereby demonstrating that the distinction between analytic and intuitive reasoning is an arbitrary and simplistic, but nonetheless useful, representation of cognition).
Research into the hypothetico-deductive model of reasoning has had surprising difficulty identifying the elements of the reasoning process that distinguish experts from novices. This has led to a shift from examining the problem-solving process of experts to analyzing the organization of their knowledge. For example, diagnosis may be based on the resemblance of a new case to prior individual instances (exemplars). Experts have a much larger store of memorized cases, for example, visual long-term memory in radiology. However, clinicians do not simply rely on literal recall of specific cases but have constructed elaborate conceptual networks of memorized information or models of disease to aid in arriving at their conclusions. That is, expertise involves an increased ability to connect symptoms, signs, and risk factors to one another in meaningful ways; relate those findings to possible diagnoses; and identify the additional information necessary to confirm the diagnosis.

No single theory accounts for all the key features of expertise in medical diagnosis. Experts have more knowledge about more things and a larger repertoire of cognitive tools to employ in problem solving than do novices. One definition of expertise highlights the ability to make powerful distinctions. In this sense, expertise involves a working knowledge of the diagnostic possibilities and what features distinguish one disease from another. Memorization alone is insufficient. Memorizing a medical textbook would not make one an expert. But having access to detailed and specific relevant information is critically important. Clinicians of the past primarily accessed their own remembered experience. Clinicians of the future will be able to access the experience of large numbers of clinicians using electronic tools, but, as with the memorized textbook, the data alone will not create an instant expert. The expert adds these data to an extensive internalized database of knowledge and experience not available to the novice (and nonexpert).

Despite all the work that has been done to understand expertise, in medicine and other disciplines, it remains uncertain whether there is any didactic program that can accelerate the progression from novice to expert or from experienced clinician to master clinician. Deliberate effortful practice (over an extended period of time, sometimes said to be 10 years or 10,000 practice hours) and personal coaching are two strategies that are often used outside medicine (e.g., music, athletics, chess) to promote expertise. Their use in developing medical expertise and maintaining or enhancing it has not yet been adequately explored.

**DIAGNOSTIC VERSUS THERAPEUTIC DECISION MAKING**

The modern ideal of medical therapeutic decision making is to “personalize” the recommendation. In the abstract, personalizing treatment involves combining the best available evidence about what works with an individual patient’s unique features (e.g., risk factors) and his or her preferences and health goals to craft an optimal treatment recommendation with the patient. Operationally, there are two different and complementary levels of personalization possible: individualizing the evidence for the specific patient based on relevant clinical and other characteristics, and personalizing the patient interaction by incorporating their values, often referred to as shared decision-making, which is critically important, but falls outside the scope of this chapter.

Individualizing the evidence about therapy does not mean relying on physician impressions of what works based on personal experience. Because of small sample sizes and rare events, the chance of drawing
erroneous causal inferences from one’s own clinical experience is very high. For most chronic diseases, therapeutic effectiveness is only demonstrable statistically in patient populations. It would be incorrect to infer with any certainty, for example, that treating a hypertensive patient with angiotensin-converting enzyme (ACE) inhibitors necessarily prevented a stroke from occurring during treatment, or that an untreated patient would definitely have avoided a stroke had he or she been treated. For many chronic diseases, a majority of patients will remain event free regardless of treatment choices; some will have events regardless of which treatment is selected; and those who avoided having an event through treatment cannot be individually identified. Blood pressure lowering, a readily observable surrogate endpoint, does not have a tightly coupled relationship with strokes prevented. Consequently, demonstrating therapeutic effectiveness cannot rely simply on observing the outcome of an individual patient but should instead be based on large groups of patients carefully studied and properly analyzed.

Therapeutic decision making, therefore, should be based on the best available evidence from clinical trials and well-done outcome studies. Authoritative, well-done clinical practice guidelines that synthesize such evidence offer readily available, reliable, and trustworthy information relevant to many treatment decisions clinicians face. However, all guidelines recognize that their “one size fits all” recommendations may not apply to individual patients. Increased attention is now being paid to understand how best to adjust group-level clinical evidence of treatment harms and benefits to account for the absolute level of risks faced by subgroups and even individual patients, using, for example, validated clinical risk scores.

**NONCLINICAL INFLUENCES ON CLINICAL DECISION-MAKING**

More than a decade of research on variations in clinician practice patterns has shed much light on the forces that shape clinical decisions. These factors can be grouped conceptually into three overlapping categories: (1) factors related to physicians’ personal characteristics and practice style, (2) factors related to the practice setting, and (3) factors related to economic incentives.

**Factors Related to Practice Style**

To ensure that necessary care is provided at a high level of quality, physicians fulfill a key role in medical care by serving as the patient’s agent. Factors that influence performance in this role include the physician’s knowledge, training, and experience. Clearly, physicians cannot practice EBM (described later in the chapter) if they are unfamiliar with the evidence. As would be expected, specialists generally know the evidence in their field better than do generalists. Beyond published evidence and practice guidelines, a major set of influences on physician practice can be subsumed under the general concept of “practice style.” The practice style serves to define norms of clinical behavior. Beliefs about effectiveness of different therapies and preferred patterns of diagnostic test use are examples of different facets of a practice style. The physician beliefs that drive these different practice styles may be based on personal experience, recollection, and interpretation of the available medical evidence. For example, heart failure specialists are much more likely than generalists to achieve target doses of ACE inhibitor therapy in their heart failure patients because they are more familiar with what the targets are (as defined by large clinical trials), have more familiarity with the
specific drugs (including adverse effects), and are less likely to overreact to foreseeable problems in therapy such as a rise in creatinine levels or asymptomatic hypotension.

Beyond the patient’s welfare, physician perceptions about the risk of a malpractice suit resulting from either an erroneous decision or a bad outcome may drive clinical decisions and create a practice referred to as defensive medicine. This practice involves using tests and therapies with very small marginal benefit, ostensibly to preclude future criticism should an adverse outcome occur. Without any conscious awareness of a connection to the risk of litigation, however, over time such patterns of care may become accepted as part of the practice norm, thereby perpetuating their overuse, e.g., annual cardiac exercise testing in asymptomatic patients.

**Practice Setting Factors**

Factors in this category relate to the physical resources available to the physician’s practice and the practice environment. Physician-induced demand is a term that refers to the repeated observation that once medical facilities and technologies are made available to physicians, they will use them. Other environmental factors that can influence decision-making include the local availability of specialists for consultations and procedures; “high-tech” advanced imaging or procedure facilities such as MRI machines and proton beam therapy centers; and fragmentation of care.

**Economic Incentives**

Economic incentives are closely related to the other two categories of practice-modifying factors. Financial issues can exert both stimulatory and inhibitory influences on clinical practice. In general, physicians are paid on a fee-for-service, capitation, or salary basis. In fee-for-service, physicians who do more get paid more, thereby encouraging overuse, consciously or unconsciously. When fees are reduced (discounted reimbursement), doctors tend to increase the number of services provided to maintain revenue. Capitation, in contrast, provides a fixed payment per patient per year to encourage physicians to consider a global population budget in managing individual patients and ideally reducing the use of interventions with small marginal benefit. In contrast to inexpensive preventive services, however, this type of incentive is more likely to affect expensive interventions. To discourage volume-based excessive utilization, fixed salary compensation plans pay physicians the same regardless of the clinical effort expended, but may provide an incentive to see fewer patients.

**INTERPRETATION OF DIAGNOSTIC TESTS IN THE CONTEXT OF DECISION-MAKING**

Despite the great technological advances in medicine over the last century, uncertainty remains a key challenge in all aspects of medical decision-making. Compounding this challenge is the massive information overload that characterizes modern medicine. Today’s clinician needs access to close to 2 million pieces of information to practice medicine. According to one estimate, doctors subscribe to an average of seven
journals, representing over 2500 new articles each year. Of course, to be useful, this information must be sifted for applicability to and then integrated with patient-specific data. Although computers appear to offer an obvious solution both for information management and for quantification of medical care uncertainties, many practical problems must be solved before computerized decision support can be routinely incorporated into the clinical reasoning process in a way that demonstrably improves the quality of care. For the present, understanding the nature of diagnostic test information can help clinicians become more efficient users of such data. The next section reviews important concepts related to diagnostic testing.

**DIAGNOSTIC TESTING: MEASURES OF TEST ACCURACY**

The purpose of performing a test on a patient is to reduce uncertainty about the patient’s diagnosis or prognosis in order to facilitate optimal management. Although diagnostic tests commonly are thought of as laboratory tests (e.g., blood count) or procedures (e.g., colonoscopy or bronchoscopy), any technology that changes a physician’s understanding of the patient’s problem qualifies as a diagnostic test. Thus, even the history and physical examination can be considered a form of diagnostic test. In clinical medicine, it is common to reduce the results of a test to a dichotomous outcome, such as positive or negative, normal or abnormal. Although this simplification ignores useful information (such as the degree of abnormality), such simplification does make it easier to demonstrate the fundamental principles of test interpretation discussed below.

The accuracy of diagnostic tests is defined in relation to an accepted “gold standard,” which defines the presumably true state of the patient (Table 3-1). Characterizing the diagnostic performance of a new test requires identifying an appropriate population (ideally, patients in whom the new test would be used) and applying both the new and the gold standard tests to all subjects. Biased estimates of test performance may occur from using an inappropriate population or from incompletely applying the gold standard test. By comparing the two tests, the characteristics of the new test are determined. The sensitivity or true-positive rate of the new test is the proportion of patients with disease (defined by the gold standard) who have a positive (new) test. This measure reflects how well the new test identifies patients with disease. The proportion of patients with disease who have a negative test is the false-negative rate and is calculated as 1 – sensitivity. Among patients without disease, the proportion who have a negative test is the specificity, or true-negative rate. This measure reflects how well the new test correctly identifies patients without disease. Among patients without disease, the proportion who have a positive test is the false-positive rate, calculated as 1 – specificity. A perfect test would have a sensitivity of 100% and a specificity of 100% and would completely distinguish patients with disease from those without it.
### TABLE 3-1

**Measures of Diagnostic Test Accuracy**

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Disease Status</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False negative (FN)</td>
<td>True negative (TN)</td>
<td></td>
</tr>
</tbody>
</table>

**Test Characteristics in Patients with Disease**

- True-positive rate (sensitivity) = \( \frac{TP}{TP + FN} \)
- False-negative rate = \( \frac{FN}{TP + FN} \)
- True-positive rate = 1 – false-negative rate

**Test Characteristics in Patients without Disease**

- True-negative rate (specificity) = \( \frac{TN}{TN + FP} \)
- False-positive rate = \( \frac{FP}{TN + FP} \)
- True-negative rate = 1 – false-positive rate

Calculating sensitivity and specificity requires selection of a threshold value or cut point above which the test is considered “positive.” Making the cut point “stricter” (e.g., raising it) lowers sensitivity but improves specificity, whereas making it “laxer” (e.g., lowering it) raises sensitivity but lowers specificity. This dynamic trade-off between more accurate identification of subjects with disease versus those without disease is often displayed graphically as a receiver operating characteristic (ROC) curve ([Fig. 3-1](#)) by plotting sensitivity \((y\text{ axis})\) versus \(1 - \text{specificity (x axis)}\). Each point on the curve represents a potential cut point with an associated sensitivity and specificity value. The area under the ROC curve often is used as a quantitative measure of the information content of a test. Values range from 0.5 (no diagnostic information from testing at all; the test is equivalent to flipping a coin) to 1.0 (perfect test). The choice of cut point should depend on the relative harms and benefits of treatment for those without versus those with disease. For example, if treatment was safe with substantial benefit, then choosing a high-sensitivity cut point (upper right of the ROC curve) for a low-risk test may be appropriate (e.g., phenylketonuria in newborns), but if treatment had substantial risk for harm, then choosing a high-specificity cut point (lower left of the ROC curve) may be appropriate (e.g.,
amniocentesis that may lead to therapeutic abortion of a normal fetus). The choice of cut point may also depend on the likelihood of disease, with low likelihoods placing a greater emphasis on the harms of treating false-positive tests and higher likelihoods placing a greater emphasis on missed benefit by not treating false-negative tests.

**FIGURE 3-1**

Each receiver operating characteristic (ROC) curve illustrates a trade-off that occurs between improved test sensitivity (accurate detection of patients with disease) and improved test specificity (accurate detection of patients without disease), because the test value defining when the test turns from “negative” to “positive” is varied. A 45° line would indicate a test with no predictive value (sensitivity = specificity at every test value). The area under each ROC curve is a measure of the information content of the test. Thus, a larger ROC area signifies increased diagnostic accuracy.

![ROC Curve Diagram](https://www.accessmedicine.com)

**MEASURES OF DISEASE PROBABILITY AND BAYES’ RULE**

Unfortunately, there are no perfect tests. After every test is completed, the true disease state of the patient remains uncertain. Quantifying this residual uncertainty can be done with Bayes’ rule, which provides a simple way to calculate the likelihood of disease after a test result or posttest probability from three parameters: the pretest probability of disease, the test sensitivity, and the test specificity. The pretest probability is a quantitative estimate of the likelihood of the diagnosis before the test is performed and is usually the prevalence of the disease in the underlying population although occasionally it can be the disease incidence. For some common conditions, such as CAD, nomograms and statistical models generate estimates of pretest probability that account for history, physical examination, and test findings. The posttest
probability (also called the predictive value of the test) is a revised statement of the likelihood of the diagnosis, accounting for both pretest probability and test results. For the likelihood of disease following a positive test (i.e., positive predictive value), Bayes' rule is calculated as:

\[
\text{Posttest probability} = \frac{\text{Pretest probability} \times \text{Sensitivity}}{\text{Pretest probability} \times \text{Sensitivity} + (1 - \text{Pretest probability}) \times \text{False-positive Rate}}
\]

For example, with a pretest probability of 0.50 and a “positive” diagnostic test result (test sensitivity = 0.90 and specificity = 0.90):

\[
\text{Posttest probability} = \frac{0.50 \times 0.90}{0.50 \times 0.90 + (1 - 0.50) \times 0.10} = 0.90
\]

The term *predictive value* often is used as a synonym for the posttest probability. Unfortunately, clinicians commonly misinterpret reported predictive values as intrinsic measures of test accuracy. Studies of diagnostic tests compound the confusion by calculating predictive values on the same sample used to measure sensitivity and specificity. Since all posttest probabilities are a function of the prevalence of disease in the tested population, such calculations may be misleading unless the test is applied subsequently to populations with the same disease prevalence. For these reasons, the term *predictive value* is best avoided in favor of the more informative *posttest probability* following a positive or a negative test result.

The nomogram version of Bayes’ rule (Fig. 3-2) helps us to conceptually understand how it estimates the posttest probability of disease. In this nomogram, the impact of the diagnostic test result is summarized by the *likelihood ratio*, which is defined as the ratio of the probability of a given test result (e.g., “positive” or “negative”) in a patient with disease to the probability of that result in a patient without disease, thereby providing a measure of how well the test distinguishes those with from those without disease.

**FIGURE 3-2**

Nomogram version of Bayes’ rule used to predict the posttest probability of disease (right-hand scale) using the pretest probability of disease (left-hand scale) and the likelihood ratio for a positive test (middle scale). See text for information on calculation of likelihood ratios. To use, place a straight edge connecting the pretest probability and the likelihood ratio and read off the posttest probability. The right-hand part of the figure illustrates the value of a positive exercise treadmill test (likelihood ratio 4, green line) and a positive exercise thallium single-photon emission computed tomography perfusion study (likelihood ratio 9, broken yellow line) in a patient with a pretest probability of coronary artery disease of 50%. (*Adapted from Centre for Evidence-Based Medicine: Likelihood ratios. Available at http://www.cebm.net/.*
For a positive test, the likelihood ratio positive is calculated as the ratio of the true-positive rate to the false-positive rate (or sensitivity/[1 – specificity]). For example, a test with a sensitivity of 0.90 and a specificity of 0.90 has a likelihood ratio of 0.90/(1 – 0.90), or 9. Thus, for this hypothetical test, a “positive” result is nine times more likely in a patient with the disease than in a patient without it. Most tests in medicine have likelihood ratios for a positive result between 1.5 and 20. Higher values are associated with tests that more substantially increase the posttest likelihood of disease. A very high likelihood ratio positive (exceeding 10) usually implies high specificity, so a positive high-specificity test helps “rule in” disease. If sensitivity is excellent but specificity is less so, the likelihood ratio will be reduced substantially (e.g., with a 90% sensitivity but a 55% specificity, the likelihood ratio is 2.0).

For a negative test, the corresponding likelihood ratio negative is the ratio of the false-negative rate to the true-negative rate (or [1 – sensitivity]/specificity). Lower likelihood ratio values more substantially lower the posttest likelihood of disease. A very low likelihood ratio negative (falling below 0.10) usually implies high sensitivity, so a negative high-sensitivity test helps “rule out” disease. The hypothetical test considered
above with a sensitivity of 0.9 and a specificity of 0.9 would have a likelihood ratio for a negative test result of 
\[(1 - 0.9)/0.9,\] or 0.11, meaning that a negative result is about one-tenth as likely in patients with disease than in those without disease (or 10 times more likely in those without disease than in those with disease).

APPLICATIONS TO DIAGNOSTIC TESTING IN CAD

Consider two tests commonly used in the diagnosis of CAD: an exercise treadmill test and an exercise single-photon emission CT (SPECT) myocardial perfusion imaging test (Chap. 270e). Meta-analysis has shown that a positive treadmill ST-segment response has an average sensitivity of 66% and an average specificity of 84%, yielding a likelihood ratio of 4.1 \((0.66/[1 – 0.84])\) (consistent with small discriminatory ability because it falls between 2 and 5). For a patient with a 10% pretest probability of CAD, the posttest probability of disease after a positive result rises to only about 30%. If a patient with a pretest probability of CAD of 80% has a positive test result, the posttest probability of disease is about 95%.

In contrast, exercise SPECT myocardial perfusion test is more accurate for CAD. For simplicity, assume that the finding of a reversible exercise-induced perfusion defect has both a sensitivity and a specificity of 90%, yielding a likelihood ratio for a positive test of 9.0 \((0.90/[1 – 0.90])\) (consistent with moderate discriminatory ability because it falls between 5 and 10). For the same 10% pretest probability patient, a positive test raises the probability of CAD to 50% (Fig. 3-2). However, despite the differences in posttest probabilities between these two tests (30% versus 50%), the more accurate test may not improve diagnostic likelihood enough to change patient management (e.g., decision to refer to cardiac catheterization) because the more accurate test has only moved the physician from being fairly certain that the patient did not have CAD to a 50:50 chance of disease. In a patient with a pretest probability of 80%, exercise SPECT test raises the posttest probability to 97% (compared with 95% for the exercise treadmill). Again, the more accurate test does not provide enough improvement in posttest confidence to alter management, and neither test has improved much on what was known from clinical data alone.

In general, positive results with an accurate test (e.g., likelihood ratio positive 10) when the pretest probability is low (e.g., 20%) do not move the posttest probability to a range high enough to rule in disease (e.g., 80%). In screening situations, pretest probabilities are often particularly low because patients are asymptomatic. In such cases, specificity becomes particularly important. For example, in screening first-time female blood donors without risk factors for HIV, a positive test raised the likelihood of HIV to only 67% despite a specificity of 99.995% because the prevalence was 0.01%. Conversely, with a high pretest probability, a negative test may not rule out disease adequately if it is not sufficiently sensitive. Thus, the largest change in diagnostic likelihood following a test result occurs when the clinician is most uncertain (i.e., pretest probability between 30% and 70%). For example, if a patient has a pretest probability for CAD of 50%, a positive exercise treadmill test will move the posttest probability to 80% and a positive exercise SPECT perfusion test will move it to 90% (Fig. 3-2).

As presented above, Bayes’ rule employs a number of important simplifications that should be considered. First, few tests have only positive or negative results, and many tests provide multiple outcomes (e.g., ST-segment depression and exercise duration with exercise testing). Although Bayes’ rule can be adapted to this
more detailed test result format, it is computationally more complex to do so. Similarly, when multiple tests are performed, the posttest probability may be used as the pretest probability to interpret the second test. However, this simplification assumes conditional independence—that is, that the results of the first test do not affect the likelihood of the second test result—and this is often not true.

Finally, it has long been asserted that sensitivity and specificity are prevalence-independent parameters of test accuracy, and many texts still make this statement. This statistically useful assumption, however, is clinically simplistic. A treadmill exercise test, for example, has a sensitivity in a population of patients with one-vessel CAD of around 30%, whereas its sensitivity in patients with severe three-vessel CAD approaches 80%. Thus, the best estimate of sensitivity to use in a particular decision may vary, depending on the severity of disease in the local population. A hospitalized, symptomatic, or referral population typically has a higher prevalence of disease and, in particular, a higher prevalence of more advanced disease than does an outpatient population. Consequently, test sensitivity will likely be higher in hospitalized patients, and test specificity higher in outpatients.

**STATISTICAL PREDICTION MODELS**

Bayes’ rule, while illustrative as presented above, provides an unrealistically simple solution to most problems a clinician faces. Predictions based on multivariable statistical models, however, can more accurately address these more complex problems by accounting for specific patient characteristics. In particular, these models explicitly account for multiple possibly overlapping pieces of patient-specific information and assign a relative weight to each on the basis of its unique contribution to the prediction in question. For example, a logistic regression model to predict the probability of CAD considers all the relevant independent factors from the clinical examination and diagnostic testing and their significance instead of the limited data that clinicians can manage in their heads or with Bayes’ rule. However, despite this strength, prediction models are usually too complex computationally to use without a calculator or computer (although this limitation may be overcome once medicine is practiced from a fully computerized platform).

To date, only a handful of prediction models have been validated properly (for example, Wells criteria for pulmonary embolism) (Table 3-2). The importance of independent validation in a population separate from the one used to develop the model cannot be overstated. An unvalidated prediction model should be viewed with the skepticism appropriate for any new drug or medical device that has not had rigorous clinical trial testing.
TABLE 3-2

Wells Clinical Prediction Rule for Pulmonary Embolism

<table>
<thead>
<tr>
<th>Clinical Feature</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical signs of deep vein thrombosis</td>
<td>3</td>
</tr>
<tr>
<td>Alternative diagnosis is less likely than pulmonary embolism</td>
<td>3</td>
</tr>
<tr>
<td>Heart rate &gt;100 beats/min</td>
<td>1.5</td>
</tr>
<tr>
<td>Immobilization ≥3 days or surgery in previous 4 weeks</td>
<td>1.5</td>
</tr>
<tr>
<td>History of deep vein thrombosis or pulmonary embolism</td>
<td>1.5</td>
</tr>
<tr>
<td>Hemoptysis</td>
<td>1</td>
</tr>
<tr>
<td>Malignancy (with treatment within 6 months) or palliative</td>
<td>1</td>
</tr>
</tbody>
</table>

**Interpretation**

- Score >6.0                                                               | High   |
- Score 2.0–6.0                                                           | Intermediate |
- Score <2.0                                                               | Low    |

When statistical models have been compared directly with expert clinicians, they have been found to be more consistent, as would be expected, but not significantly more accurate. Their biggest promise, then, may be in helping less-experienced clinicians identify critical discriminating patient characteristics and become more accurate in their predictions.

**FORMAL DECISION SUPPORT TOOLS**

**DECISION SUPPORT SYSTEMS**

Over the last 40 years, many attempts have been made to develop computer systems to aid clinical decision-making and patient management. Conceptually attractive because computers offer ready access to the vast information available to today’s physicians, they may also support management decisions by making accurate predictions of outcome, simulating the whole decision process, or providing algorithmic guidance. Computer-based predictions using Bayesian or statistical regression models inform a clinical decision but do
not actually reach a “conclusion” or “recommendation.” Artificial intelligence systems attempt to simulate or replace human reasoning with a computer-based analogue. To date, such approaches have achieved only limited success. Reminder or protocol-directed systems do not make predictions but use existing algorithms, such as guidelines, to guide clinical practice. In general, however, decision support systems have had little impact on practice. Reminder systems, although not yet in widespread use, have shown the most promise, particularly in correcting drug dosing and promoting adherence to guidelines. Checklists, as used by pilots for example, have garnered recent support as an approach to avoid or reduce errors.

DECISION ANALYSIS

Compared with the decision support methods discussed above, decision analysis represents a prescriptive approach to decision making in the face of uncertainty. Its principal application is in complex decisions that involve substantial risk, abundant uncertainty, trade-offs in the outcomes emphasizing a role for preferences, or absence of evidence due to an idiosyncratic feature. For a public health example, **Fig. 3-3** displays a decision tree to evaluate strategies for screening for HIV infection. Infected individuals who are unaware of their illness cause up to 20,000 new cases of HIV infection annually in the United States, and about 40% of HIV-positive patients progress to AIDS within a year of the initial diagnosis because of delayed diagnosis. Early identification offers the opportunity to prevent progression to AIDS through CD4 count and viral load monitoring and combination antiretroviral therapy and to reduce spread by reducing risky injection or sexual behaviors.

**FIGURE 3-3**

*Basic structure of decision model used to evaluate strategies for screening for HIV in the general population.*

HAART, highly active antiretroviral therapy. *(Provided courtesy of G. Sanders, with permission.)*

In 2003, the Centers for Disease Control and Prevention (CDC) proposed that routine universal HIV testing should be incorporated into standard adult medical care and, in part, cited a decision analysis model comparing HIV screening with usual care. Assuming a 1% prevalence of unidentified HIV infection in the population, routine screening of a cohort of 43-year-old men and women increased life expectancy by 5.5 days and lifetime costs by $194 per person screened, yielding an incremental cost-effectiveness ratio for screening versus usual care of $15,078 per quality-adjusted life-year (the additional cost to society to increase population health by 1 year of perfect health). Factors that influenced the results included
assumptions about the effectiveness of behavior modification on subsequent sexual behavior, the benefits of early therapy for HIV infection, and the prevalence and incidence of HIV infection in the population targeted. This model, which required over 75 separate data points, provided novel insights into a public health problem in the absence of a randomized clinical trial and helped weigh the pros and cons of such a health policy recommendation. Although such models have been developed for selected clinical problems, their benefit and application to individual real-time clinical management have yet to be demonstrated.

DIAGNOSIS AS AN ELEMENT OF QUALITY OF CARE

High-quality medical care begins with accurate diagnosis. Recently, diagnostic errors have been re-envisioned: the old view was that they were caused by a lack of sufficient skill of an individual clinician; the new view is that they represent a quality of care patient-safety problem traceable to breakdowns in the health care system. Whether this conceptual shift will lead to new ways to improve diagnosis is uncertain. An annual rate of diagnostic errors of 10–15%, possibly leading to 40,000 deaths in the United States, is commonly cited, but these figures are imprecise.

Solutions to the “diagnostic errors as a system of care problem” have focused on system-level approaches, such as decision support and other tools integrated into electronic medical records. The use of checklists has been proposed as a means of reducing some of the cognitive errors discussed earlier in the chapter, such as premature closure. Although checklists have been shown to be useful in certain medical contexts, such as operating rooms and intensive care units, their value in preventing diagnostic errors that lead to patient adverse events remains to be shown.

EVIDENCE-BASED MEDICINE

Clinical medicine is defined traditionally as a practice combining medical knowledge (including scientific evidence), intuition, and judgment in the care of patients (Chap. 1). EBM updates this construct by placing much greater emphasis on the processes by which clinicians gain knowledge of the most up-to-date and relevant clinical research to determine for themselves whether medical interventions alter the disease course and improve the length or quality of life. The meaning of practicing EBM becomes clearer through an examination of its four key steps:

1. Formulating the management question to be answered

2. Searching the literature and online databases for applicable research data

3. Appraising the evidence gathered with regard to its validity and relevance

4. Integrating this appraisal with knowledge about the unique aspects of the patient (including the patient’s preferences about the possible outcomes)
The process of searching the world’s research literature and appraising the quality and relevance of studies thus identified can be quite time-consuming and requires skills and training that most clinicians do not possess. Thus, identifying recent systematic overviews of the problem in question (Table 3-3) may offer the best starting point for most EBM searches.

Generally, the EBM tools listed in Table 3-3 provide access to research information in one of two forms. The first, primary research reports, is the original peer-reviewed research work that is published in medical journals and accessible through MEDLINE in abstract form. However, without training in using MEDLINE, quickly and efficiently locating reports that are on point in a huge sea of irrelevant or unhelpful citations may be difficult, and important studies could also be missed. The second form, systematic reviews, is the highest level of evidence in the hierarchy because it comprehensively summarizes the available evidence on a particular topic up to a certain date. To avoid the potential biases in review articles, predefined explicit search strategies and inclusion and exclusion criteria are used to find all of the relevant scientific research and grade its quality. The prototype for this kind of resource is the Cochrane Database of Systematic Reviews. When appropriate, a meta-analysis quantitatively summarizes the systematic review findings. The next two sections explicate the major types of clinical research reports available in the literature and the process of aggregating those data into meta-analyses.

**SOURCES OF EVIDENCE: CLINICAL TRIALS AND REGISTRIES**

The notion of learning from observation of patients is as old as medicine itself. Over the last 50 years, physicians’ understanding of how best to turn raw observation into useful evidence has evolved considerably. Case reports, personal anecdotal experience, and small single-center case series are now recognized as having severe limitations in validity and generalizability, and although they may generate hypotheses or be the first reports of adverse events, they have no role in formulating modern standards of practice. The major tools used to develop reliable evidence consist of the randomized clinical trial and the large observational registry. A registry or database typically is focused on a disease or syndrome (e.g., cancer, CAD, heart failure), a clinical procedure (e.g., bone marrow transplantation, coronary revascularization), or an administrative process (e.g., claims data used for billing and reimbursement).
**TABLE 3-3**

Selected Tools for Finding the Evidence in Evidence-Based Medicine (EBM)

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Web Address</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviews</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cochrane Library</td>
<td>Collection of EBM databases, including The Cochrane Database of Systematic Reviews—full text articles reviewing specific health care topics.</td>
<td><a href="http://www.cochrane.org">www.cochrane.org</a></td>
<td>Subscription required. Abstracts of systematic reviews available free online. Some countries have funding to provide free access to all residents.</td>
</tr>
<tr>
<td>ACP Journal Club</td>
<td>Collection of summaries of original studies and systematic reviews. Published bimonthly. All data since 1991 available on website, updated yearly.</td>
<td><a href="http://www.acpjc.org">www.acpjc.org</a></td>
<td>Subscription required.</td>
</tr>
<tr>
<td>Clinical Evidence</td>
<td>Monthly updated directory of concise overviews of common clinical interventions.</td>
<td><a href="http://www.clinicalevidence.com">www.clinicalevidence.com</a></td>
<td>Subscription required. Free access for United Kingdom and developing countries.</td>
</tr>
</tbody>
</table>
By definition, in observational data, the investigator does not control patient care. Carefully collected prospective observational data, however, can achieve a level of evidence quality approaching that of major clinical trial data. At the other end of the spectrum, data collected retrospectively (e.g., chart review) are limited in form and content to what previous observers recorded, which may not include the specific research data being sought, e.g., claims data. Advantages of observational data include the inclusion of a broader population as encountered in practice than is typically represented in clinical trials because of their restrictive inclusion and exclusion criteria. In addition, observational data provide primary evidence for research questions when a randomized trial cannot be performed. For example, it would be difficult to randomize patients to test diagnostic or therapeutic strategies that are unproven but widely accepted in practice, and it would be unethical to randomize based on sex, racial/ethnic group, socioeconomic status, or country of residence or to randomize patients to a potentially harmful intervention, such as smoking or deliberately overeating to develop obesity.

A well-done prospective observational study of a particular management strategy differs from a well-done randomized clinical trial most importantly by its lack of protection from treatment selection bias. The use of observational data to compare diagnostic or therapeutic strategies assumes that sufficient uncertainty exists in clinical practice to ensure that similar patients will be managed differently by different physicians. In short, the analysis assumes that a sufficient element of randomness (in the sense of disorder rather than in the formal statistical sense) exists in clinical management. In such cases, statistical models attempt to adjust for important imbalances to “level the playing field” so that a fair comparison among treatment options can be made. When management is clearly not random (e.g., all eligible left main CAD patients are referred for coronary bypass surgery), the problem may be too confounded (biased) for statistical correction, and observational data may not provide reliable evidence.

In general, the use of concurrent controls is vastly preferable to that of historical controls. For example, comparison of current surgical management of left main CAD with left main CAD patients treated medically during the 1970s (the last time these patients were routinely treated with medicine alone) would be extremely misleading because “medical therapy” has substantially improved in the interim.

Randomized controlled clinical trials include the careful prospective design features of the best observational data studies but also include the use of random allocation of treatment. This design provides the best protection against measured and unmeasured confounding due to treatment selection bias (a major aspect of internal validity). However, the randomized trial may not have good external validity (generalizability) if the process of recruitment into the trial resulted in the exclusion of many patients seen in clinical practice.

Consumers of medical evidence need to be aware that randomized trials vary widely in their quality and applicability to practice. The process of designing such a trial often involves many compromises. For example, trials designed to gain U.S. Food and Drug Administration (FDA) approval for an investigational drug or device must fulfill regulatory requirements that may result in a trial population and design that differs substantially from what practicing clinicians would find most useful.
META-ANALYSIS

The Greek prefix *meta* signifies something at a later or higher stage of development. Meta-analysis is research that combines and summarizes the available evidence quantitatively. Although occasionally used to examine nonrandomized studies, meta-analysis is used most typically to summarize all randomized trials examining a particular therapy. Ideally, unpublished trials should be identified and included to avoid publication bias (i.e., missing “negative” trials that may not be published). Furthermore, the best meta-analyses obtain and analyze individual patient-level data from all trials rather than working only the summary data in published reports of each trial. Nonetheless, not all published meta-analyses yield reliable evidence for a particular problem, so their methodology should be scrutinized carefully to ensure proper study design and analysis. The results of a well-done meta-analysis are likely to be most persuasive if they include at least several large-scale, properly performed randomized trials. Meta-analysis can especially help detect benefits when individual trials are inadequately powered (e.g., the benefits of streptokinase thrombolytic therapy in acute MI demonstrated by ISIS-2 in 1988 were evident by the early 1970s through meta-analysis). However, in cases in which the available trials are small or poorly done, meta-analysis should not be viewed as a remedy for the deficiency in primary trial data.

Meta-analyses typically focus on summary measures of relative treatment benefit, such as odds ratios or relative risks. Clinicians also should examine what absolute risk reduction (ARR) can be expected from the therapy. A useful summary metric of absolute treatment benefit is the number needed to treat (NNT) to prevent one adverse outcome event (e.g., death, stroke). NNT is simply 1/ARR. For example, if a hypothetical therapy reduced mortality rates over a 5-year follow-up by 33% (the relative treatment benefit) from 12% (control arm) to 8% (treatment arm), the ARR would be 12% – 8% = 4%, and the NNT would be 1/0.04, or 25. Thus, it would be necessary to treat 25 patients for 5 years to prevent 1 death. If the hypothetical treatment was applied to a lower-risk population, say, with a 6% 5-year mortality, the 33% relative treatment benefit would reduce absolute mortality by 2% (from 6% to 4%), and the NNT for the same therapy in this lower-risk group of patients would be 50. Although not always made explicit, comparisons of NNT estimates from different studies should account for the duration of follow-up used to create each estimate.

CLINICAL PRACTICE GUIDELINES

According to the 1990 Institute of Medicine definition, clinical practice guidelines are “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances.” This definition emphasizes several crucial features of modern guideline development. First, guidelines are created by using the tools of EBM. In particular, the core of the development process is a systematic literature search followed by a review of the relevant peer-reviewed literature. Second, guidelines usually are focused on a clinical disorder (e.g., adult diabetes, stable angina pectoris) or a health care intervention (e.g., cancer screening). Third, the primary objective of guidelines is to improve the quality of medical care by identifying care practices that should be routinely implemented, based on high-quality evidence and high benefit-to-harm ratios for the interventions. Guidelines are intended to “assist” decision-making, not to define explicitly what decisions should be made in a particular
situation, in part because evidence alone is never sufficient for clinical decision-making (e.g., deciding whether to intubate and administer antibiotics for pneumonia in a terminally ill individual, in an individual with dementia, or in an otherwise healthy 30-year-old mother).

Guidelines are narrative documents constructed by expert panels whose composition often is determined by interested professional organizations. These panels vary in the degree to which they represent all relevant stakeholders. The guideline documents consist of a series of specific management recommendations, a summary indication of the quantity and quality of evidence supporting each recommendation, an assessment of the benefit-to-harm ratio for the recommendation, and a narrative discussion of the recommendations. Many recommendations simply reflect the expert consensus of the guideline panel because literature-based evidence is absent. The final step in guideline construction is peer review, followed by a final revision in response to the critiques provided. To improve the reliability and trustworthiness of guidelines, the Institute of Medicine has made methodologic recommendations for guideline development.

Guidelines are closely tied to the process of quality improvement in medicine through their identification of evidence-based best practices. Such practices can be used as quality indicators. Examples include the proportion of acute MI patients who receive aspirin upon admission to a hospital and the proportion of heart failure patients with a depressed ejection fraction treated with an ACE inhibitor.

**CONCLUSIONS**

In this era of EBM, it is tempting to think that all the difficult decisions practitioners face have been or soon will be solved and digested into practice guidelines and computerized reminders. However, EBM provides practitioners with an ideal rather than a finished set of tools with which to manage patients. Moreover, even with such evidence, it is always worth remembering that the response to therapy of the “average” patient represented by the summary clinical trial outcomes may not be what can be expected for the specific patient sitting in front of a physician in the clinic or hospital. In addition, meta-analyses cannot generate evidence when there are no adequate randomized trials, and most of what clinicians confront in practice will never be thoroughly tested in a randomized trial. For the foreseeable future, excellent clinical reasoning skills, experience supplemented by well-designed quantitative tools, and a keen appreciation for the role of individual patient preferences in their health care will continue to be of paramount importance in the practice of clinical medicine.