

This chapter appears in Harrison's Principles of Internal Medicine

Chapter 3, 20th edition

Authors

Daniel B. Mark, MD, MPH
Professor of Medicine
Duke University Medical Center
Director, Outcomes Research

Email: Daniel.mark@duke.edu

Web: <http://www.danielmarkmd.com>

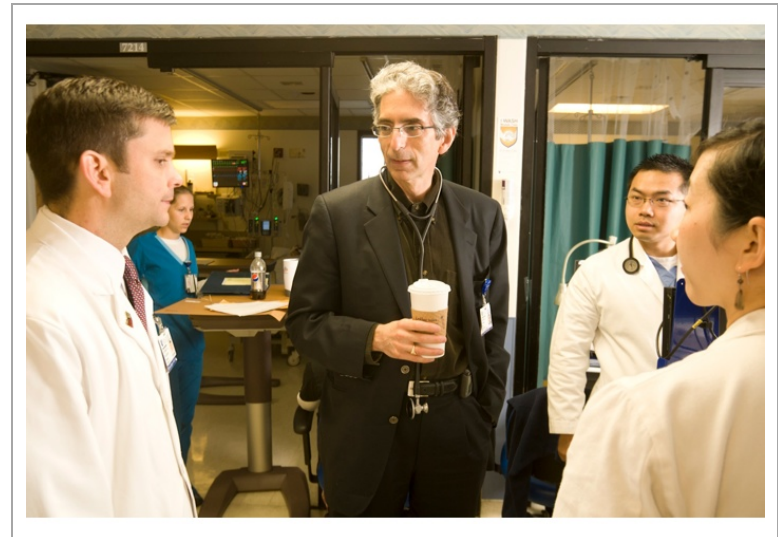
John B. Wong, MD
Chief, Division of Clinical Decision Making
Department of Medicine
Professor of Medicine
Tufts University School of Medicine
Tufts New England Medical Center

Email: jwong@tuftsmedicalcenter.org

We have adapted our pre-print version for your personal use under
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0
International license. CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

Decision- Making in Clinical Medicine



DANIEL B. MARK
JOHN B. WONG

Table of Contents

Introduction	1
Brief Introduction to Clinical Reasoning	2
Clinical expertise.....	2
Intuitive versus analytic reasoning.....	2
Diagnostic Versus Therapeutic Decision Making	7
Non-Clinical Influences on Clinical Decision-Making.....	7
Factors related to practice style	7
Practice setting factors	8
Payment systems	8
Interpretation of Diagnostic Tests	9
Diagnostic Testing: Measures of Test Accuracy	9
Measures of Disease Probability and Bayes' Rule	11
Applications to Diagnostic Testing in CAD	12
Statistical Prediction Models	13
Decision Support Tools	15
Diagnosis as an Element of Quality of Care	16
Evidence-Based Medicine	16
Clinical Practice Guidelines	19

Introduction

Sir William Osler’s familiar quote “Medicine is a science of uncertainty and an art of probability” captures well the complex nature of clinical medicine. Although the science of medicine is often taught as if the mechanisms of the human body operate with Newtonian predictability, every aspect of medical practice is infused with an element of irreducible uncertainty that the clinician ignores at her peril. Clinical medicine has deep roots in science, but it is an imprecise science. More than 100 years after the practice of medicine took its modern form, it remains at its core a craft, to which individual doctors bring varying levels of skill and understanding. With the exponential growth in medical literature and other technical information and an ever-increasing number of testing and treatment options, 21st century physicians who seek excellence in their craft must master a more diverse and complex set of skills than any of the generations that preceded them. This chapter provides an introduction to three of the pillars upon which the craft of modern medicine rests: 1) expertise in clinical reasoning (what it is and how it can be developed); 2) rational diagnostic tests, use and interpretation; and 3) integration of the best available research evidence with clinical judgment in the care of individual patients (*evidence-based medicine* or *EBM* and 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 (or get recertified), no tests or benchmarks are used to identify those who have attained the highest levels of clinical performance. 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 such 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 involving the integration of disease knowledge with verbal and visual cues and test interpretation but also potentially the complex fine-motor skills necessary for invasive procedures and tests. In addition, “the complete package” of expertise in medicine requires effective communication and care coordination with patients and members of the medical team. Research on medical expertise remains sparse overall and mostly centered on diagnostic reasoning, so in this chapter, we focus primarily on the cognitive elements of clinical reasoning.

Because clinical reasoning occurs in the heads of clinicians, objective study of the process is difficult. One research method used for this area asks clinicians to “think out loud” as they receive increments of clinical information in a manner meant to simulate a clinical encounter. Another research approach focuses on how doctors should reason diagnostically to identify remediable “errors” 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, medical education, behavioral economics, sociology, informatics, and decision sciences, no single integrated model of clinical reasoning exists, and not infrequently, different terms and reasoning models describe similar phenomena.

Intuitive Versus Analytic Reasoning

A useful 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. A student might read about lymph nodes in the lung and from that list, identify diseases more common in African-American women or examine the patient for skin or eye findings that may occur with sarcoid. These dual processes, of course, represent two exemplars taken from the cognitive continuum. They provide helpful descriptive insights but very little guidance in how to develop expertise in clinical reasoning. How these idealized 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 study and considerable debate.

Pattern recognition, an important part of System 1 reasoning, is a complex cognitive process that appears largely effortless. One can recognize people's faces, the breed of a dog, an automobile model, or a piece of music from just a few notes within milliseconds without necessarily being able to articulate the specific features that prompted the recognition. Analogously, experienced clinicians often recognize familiar diagnosis patterns very quickly. The key here is having a large library of stored patterns that can be rapidly accessed. In the absence of an extensive stored repertoire of diagnostic patterns, students (as well as more experienced clinicians operating outside their area of expertise and familiarity) often must 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 illustrate three distinct patterns that experienced clinicians recognize without effort:

- 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, with blood-streaked sputum, and weight loss fits the pattern of carcinoma of the lung. Consequently, along with the chest x-ray, the clinician obtains a sputum cytology examination and refers this patient for a chest CT scan.

- In the third scenario, the clinician hears a soft diastolic rumbling murmur at the apex on cardiac auscultation in a 46-year-old patient with hemoptysis who immigrated from a developing country and orders an echocardiogram as well, because of possible pulmonary hypertension from suspected rheumatic mitral stenosis.

Pattern recognition by itself is not, however, sufficient for secure diagnosis. Without deliberative systematic reflection, pattern recognition can result in premature closure: mistakenly jumping to the conclusion that one has correct diagnosis before all the relevant data are in. A critical second step, even when the diagnosis seems obvious, is *diagnostic verification*: considering whether the diagnosis adequately accounts for the presenting symptoms and signs and can explain all the ancillary findings. An example of premature closure is contained in the following case, modified from a real clinical encounter. A 45-year-old man presents with a 3-week history of a "flulike" upper respiratory infection (URI) including dyspnea and a productive cough. The Emergency Department clinician pulled out a "URI assessment form" which defines and standardizes 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 prescribed a cough suppressant for acute bronchitis and reassured the patient that his illness was not serious. Following a sleepless night at home with significant dyspnea, the patient developed nausea and vomiting and collapsed. He was brought back to the emergency department in cardiac arrest and was unable to be resuscitated. His autopsy showed a posterior wall myocardial infarction and a fresh thrombus in an atherosclerotic right coronary artery. What went wrong? Presumably, the ED clinician felt that the patient was basically healthy (one can be misled by the way the patient appears on exam – a patient that does not "appear sick" may be incorrectly assumed to have an innocuous illness). So in this case, the physician, upon hearing the overview of the patient from the triage nurse, elected to use the URI assessment protocol even before starting the history, closing consideration of the broader range of possibilities and associated tests

required to confirm or refute these possibilities. In particular, by concentrating on the abbreviated and focused URI protocol, the clinician failed to elicit the full dyspnea history, which was precipitated by exertion and accompanied by chest heaviness and relieved by rest, suggesting a far more serious disorder

Heuristics or rules of thumb are a part of the intuitive system. These cognitive shortcuts provide a quick and easy path to reaching conclusions and making choices, but when used improperly they can lead to errors. Two major research programs have studied heuristics in a mostly non-medical context and have reached very different conclusions about the value of these cognitive tools. The “heuristics and biases” program focuses on how relying on heuristics can lead to cognitive biases and incorrect judgments. Over 100 different cognitive biases have been described. So far, however, there is little evidence that educating physicians and other decision makers to watch for these cognitive biases has any effect on the rate of diagnostic errors. In contrast, the “fast and frugal heuristics” research program explores how and when relying on simple heuristics can produce good decisions. Although many heuristics have relevance to clinical reasoning, only four will be mentioned here.

When diagnosing patients, clinicians usually develop diagnostic hypotheses based on the similarity of that patient’s symptoms, signs and other data to their mental representations (memorized patterns) of the disease possibilities. In other words, clinicians pattern match to identify the diagnoses which share the most similar findings to the patient at hand. This cognitive shortcut is called the representativeness heuristic. Consider a patient with hypertension and headache, palpitations and diaphoresis. Based on the representativeness heuristic, clinicians might judge pheochromocytoma to be quite likely given this classic presenting symptom triad suggesting pheochromocytoma. Doing so however, 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 it. Thus, clinicians using the representativeness heuristic may overestimate the likelihood of a particular disease based on its representativeness by failing to recognize

the low underlying prevalence (i.e., the prior, or pretest, probabilities). Conversely, atypical presentations of common diseases may lead to underestimating the likelihood of a particular disease. Thus, inexperience with a specific disease and with the breadth of its presentations may also lead to diagnostic delays or errors, e.g., diseases that affect multiple organ systems, such as sarcoid or tuberculosis, may be particularly challenging to diagnose because of the many different patterns they may manifest.

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, a clinician may recall a case from a morbidity and mortality conference in which an elderly patient presented with painless dyspnea of acute onset and was evaluated for a pulmonary cause but eventually found to have acute MI with the diagnostic delay likely contributing to the development of ischemic cardiomyopathy. If the case was associated with a malpractice accusation, such examples may be even more memorable. 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 subsequently found to have a sarcoma—and those publicized in the media or recent experience 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 insufficiently adjusting the initial probability of disease up (or down) following a positive (or negative test) when compared with Bayes’ theorem, i.e., sticking to the initial diagnosis. For example, a clinician may still judge the probability of coronary artery disease (CAD) to be high despite a negative exercise perfusion test and go on to cardiac catheterization (see “Measures of Disease Probability and Bayes’ Theorem,” below).

The fourth heuristic states that clinicians should use the simplest explanation possible that will adequately account for the patient's symptoms and 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.

For complex or unfamiliar diagnostic problems, clinicians typically resort to analytic reasoning processes (System 2) and proceed methodically using the *hypothetico-deductive model of reasoning*. Based on the stated reasons for seeking medical attention, clinicians develop an initial list of diagnostic possibilities in *hypothesis generation*. During the history of the present illness, the initial hypotheses evolve in *diagnostic refinement* as emerging information is tested against the mental models of the diseases being considered with diagnoses increasing and decreasing in likelihood or even being dropped from consideration as the working hypotheses of the moment. These mental models often generate additional questions that distinguish the diagnostic possibilities from one another. The focused physical examination contributes further distinguishing the working hypotheses. Is the spleen enlarged? How big is the liver? Is it tender? Are there any palpable masses or nodules? *Diagnostic verification* involves testing the adequacy (whether the diagnosis accounts for all symptoms and signs) and coherency (whether the signs and symptoms are consistent with the underlying pathophysiological causal mechanism) of the diagnosis. For example, if the enlarged and quite tender liver felt on physical examination is due to acute hepatitis (the hypothesis), then certain specific liver function tests will be markedly elevated (the prediction). Should the tests come back normal, the hypothesis may have to be discarded or substantially modified.

Although often neglected, negative findings are as important as positive ones because they reduce the likelihood of the diagnostic hypotheses

under consideration. Chest discomfort that is not provoked or worsened by exertion and not relieved by rest 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 should consider aortic dissection routinely as a possible cause of acute severe chest discomfort. Although the typical presenting symptoms of dissection differ from that of MI, dissection may mimic MI, and because it is far less prevalent and potentially fatal if mistreated, diagnosing dissection remains a challenging diagnostic imperative. Clinicians taking care of acute, severe chest pain patients should explicitly and routinely inquire about symptoms suggestive of dissection, measure blood pressures in both arms for discrepancies and examine for pulse deficits. When these are all negative, clinicians may feel sufficiently reassured to discard the aortic dissection hypothesis. If, however, the chest x-ray shows a possible widened mediastinum, the hypothesis should be reinstated and an appropriate imaging test ordered [e.g., thoracic computed tomography (CT) scan or transesophageal echocardiogram]. In non-acute 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 is limited (classically humans can accurately repeat a list of 7 ± 2 numbers read to them), the number of diagnoses that can be actively considered in

hypothesis-generating activities is similarly limited. For this reason, 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 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 pattern matching as exemplars, prototypes and illness scripts. For example, diagnosis may be based on the resemblance of a new case to patients seen previously (exemplars). As abstract mental models of disease, prototypes incorporate the likelihood of various disease features. Illness scripts include risk factors, pathophysiology, and symptoms and signs. Experts have a much larger store of exemplar and prototype cases, an example of which is the visual long-term memory of experienced radiologists. 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 (illness scripts). That is, expertise involves an enhanced 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 presenting symptoms of diseases and a larger repertoire of cognitive tools to employ in problem solving than non-experts. One definition of expertise highlights the ability to make powerful distinctions. In this sense, expertise involves a working knowledge of the diagnostic possibilities and those features that distinguish one disease from another.

Memorization alone is insufficient, e.g., photographic memory of a medical textbook would not make one an expert. But having access to detailed case-specific relevant information is critically important. In the past, clinicians primarily acquired clinical knowledge through their patient experiences, but now clinicians have access to a plethora of information sources (see evidence-based medicine (EBM) below). Clinicians of the future will be able to leverage the experiences of large numbers of other clinicians using electronic tools, but, as with the memorized textbook, the data alone will be insufficient for becoming an expert. Nonetheless, availability of these data removes one barrier for acquiring experience with connecting symptoms, signs, and risk factors to the possible diagnoses and identifying the additional distinguishing information necessary to confirm the diagnosis, thereby potentially facilitating the development of the working knowledge necessary for becoming an expert.

Despite all of the research seeking to understand expertise in medicine and other disciplines, it remains uncertain whether any didactic program can actually 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. Some studies in medicine suggest that didactic education exposing students to both the signs and symptoms of specific diseases and, in addition, the diseases that may present with specific symptoms and signs may be beneficial. Developing a personal learning system (e.g., metacognition) through for example EBM processes below and follow-up to identify diagnoses and treatments for patients that you have cared for provide active learning opportunities.

Diagnostic Versus Therapeutic Decision Making

The modern ideal of medical therapeutic decision making is to “personalize” treatment recommendations. 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, genomics and co-morbidities) and his or her preferences and health goals to craft an optimal treatment recommendation with the patient. Operationally, two different and complementary levels of personalization are possible: individualizing the risk of harm and benefit for the options being considered based on the specific patient characteristics (precision medicine), and personalizing the therapeutic decision process by incorporating the patient’s preferences and values for the possible health outcomes. This latter process is sometimes referred to as shared decision-making, and typically involves clinicians sharing their knowledge about the options and the associated consequences and tradeoffs, and patients sharing their health goals, e.g., avoiding a short-term risk of dying from coronary artery bypass grafting to see their grandchild get married in a few months.

Individualizing the evidence about therapy does not mean relying on physician impressions of benefit and harm from their 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 large 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 their stroke had they 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, in most situations 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. Trustworthy clinical practice guidelines that synthesize such evidence offer normative guidance for many testing and treatment decisions. However, all guidelines recognize that “one size fits all” recommendations may not apply to individual patients. Increased research into the heterogeneity of treatment effects seeks 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 by individual patients, using, for example, validated clinical risk scores.

Non-Clinical Influences on Clinical Decision-Making

More than three decades of research on variations in clinician practice patterns has identified important non-clinical forces that shape clinical decisions. These factors can be grouped conceptually into three overlapping categories: (1) factors related to individual physicians practice, (2) factors related to practice setting, and (3) factors related to payment systems.

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

advocate. Factors that influence performance in this role include the physician's knowledge, training, and experience. Clearly, physicians cannot practice evidence-based medicine (described below) 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 training, personal experience, and medical evidence. For example, in heart failure patients, heart failure specialists have more familiarity than general internists with the target doses of angiotensin-converting enzyme (ACE) inhibitor therapy as defined by large clinical trials and 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. Not surprisingly, the specialists are much more likely than generalists to achieve target doses of ACE inhibitor therapy. By contrast, perhaps due to specialization, cardiologists may overestimate the benefit and underestimate the harm of coronary revascularization relative to general internists.

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 benefits, ostensibly to preclude future criticism should an adverse outcome occur. With conscious or unconscious awareness of a connection to the risk of litigation or to payment, 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 work systems including tasks and work flow (interruptions, inefficiencies, workload), technology (poor design or implementation, errors in use, failure, misuse), organizational characteristics (e.g., culture, leadership, staffing, scheduling), and the physical environment (e.g., noise, lighting, layout). Physician-induced demand is a term that refers to the repeated observation that once medical facilities and technologies become 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.

Payment systems

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. Historically, 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), clinicians 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. 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. In

recognition of the non-sustainability of continued growth in medical expenditures and the opportunity costs associated with that (funds that might be more beneficially applied to education, energy, social welfare or defense), current efforts seek to transition to a value-based payment system to reduce overuse and to reflect benefit. Work to define how to actually tie payment to value has mostly focused so far on “pay for performance” models. High quality clinical trial evidence for the effectiveness of these models is still mostly lacking.

Interpretation of Diagnostic Tests

Despite impressive technological advances in medicine over the last century, uncertainty still abounds and challenges all aspects of medical decision-making. Compounding this challenge, massive information overload characterizes modern medicine. Clinicians on average subscribe to seven journals, presenting them with over 2500 new articles each year, and need access to 2 million pieces of information to practice medicine. Of course, to be useful, this information must be sifted for quality and examined for applicability for integration into patient-specific care. 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 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 appropriate management. Although diagnostic tests commonly refer to laboratory (e.g., blood count) or imaging tests or procedures (e.g., colonoscopy or bronchoscopy), any information that changes a

provider’s understanding of the patient’s problem qualifies as a diagnostic test. Thus, even the history and physical examination should be considered as diagnostic tests. 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), it facilitates illustrating some important principles of test interpretation which are described below.

The accuracy of any diagnostic test is assessed relative to a “gold standard,” where a positive gold standard test defines the patients who have disease and a negative test rules out disease (Table 3-1).

TABLE 3-1 Measures of Diagnostic Test Accuracy

Test Result	Disease Status	
	Present	Absent
Positive	True positives (TP)	False positives (FP)
Negative	False negatives (FN)	True Negatives (TN)

Test Characteristics in Patients with Disease

$$\text{True-positive rate (sensitivity)} = \frac{TP}{TP + FN}$$

$$\text{False-negative rate} = \frac{FN}{TP + FN} = 1 - \text{true positive rate}$$

Test Characteristics in Patients without Disease

$$\text{True-negative rate (specificity)} = \frac{TN}{TN + FP}$$

$$\text{False-positive rate} = \frac{FP}{TN + FP} = 1 - \text{true-negative rate}$$

Characterizing the diagnostic performance of a new test requires identifying an appropriate population (ideally, patients representative of those 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 occur when diagnostic accuracy is defined using an inappropriate population or one in which gold standard determination of disease status is incomplete. The accuracy of the new test in distinguishing disease from health is determined relative to the gold standard results and summarized in 4 estimates. The sensitivity or true-positive rate of the new test reflects how well the new test identifies patients with disease. It is the proportion of patients with disease (defined by the gold standard) who have a positive test. The proportion of patients with disease who have a negative test is the false-negative rate, calculated as $1 - \text{sensitivity}$. The specificity, or true-negative rate reflects how well the new test correctly identifies patients without disease. It is the proportion of patients without disease (defined by the gold standard) who have a negative test. The proportion of patients without disease who have positive test is the false-positive rate, calculated as $1 - \text{specificity}$. In theory, a perfect test would be one with a sensitivity of 100% and a specificity of 100% and would completely distinguish patients with disease from those without it. A useful mnemonic is the following: a *negative* high sensitivity (*Sn*) test helps rule *out* disease (Negative SnOut), and a *positive* high specificity (*Sp*) test helps rule *in* disease (Positive SpIn).

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, while 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 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 in theory reflect 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

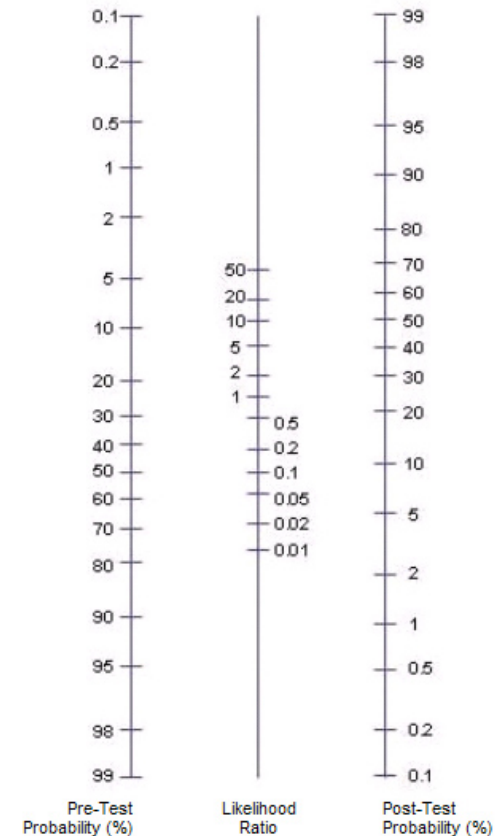


Figure 3-1 Each ROC curve illustrates a trade-off that occurs between improved test sensitivity (accurate detection of patients with disease) and improved test specificity (accurate detection of patients without disease), as the test value defining when the test turns from “negative” to “positive” is 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.

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., chemotherapy for cancer). The choice of cut point may also depend on the likelihood of disease with low likelihoods placing a greater emphasis on the harms of false positive tests (e.g., HIV testing in marriage applicants) or the harms of false negative tests (e.g., HIV testing in blood donors).

Measures of Disease Probability and Bayes' Rule

In the absence of perfect tests, the true disease state of the patient remains uncertain after every test. Bayes' rule provides a way to quantify the revised uncertainty using simple probability mathematics (and thereby avoid anchoring bias). It calculates the *posttest probability* or likelihood of disease after a test result, 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 estimated from the prevalence of the disease in the underlying population (if known) or clinical context (e.g., age, sex and type of chest pain). For some common conditions, such as CAD, existing 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, see below) is a recalibrated 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:

$$\text{Posttest probability} = \frac{\text{Pretest probability} \times \text{test sensitivity}}{\text{Pretest probability} \times \text{test sensitivity} + (1 - \text{Pretest probability}) \times \text{test false-positive rate}}$$

For example, consider a 64-year-old woman with atypical chest pain

who has a pretest probability of 0.50 and "positive" diagnostic test result (assuming test sensitivity = 0.90 and specificity = 0.90).

$$\begin{aligned} \text{Posttest probability} &= \frac{(0.50)(0.90)}{(0.50)(0.90) + (0.50)(0.10)} \\ &= 0.90 \end{aligned}$$

The term predictive value has often been used as a synonym for the posttest probability. Unfortunately, clinicians commonly misinterpret reported predictive values as intrinsic measures of test accuracy rather than calculated probabilities. Studies of diagnostic test performance compound the confusion by calculating predictive values from the same sample used to measure sensitivity and specificity. Such calculations are misleading unless the test is applied subsequently to populations with exactly the same disease prevalence. For these reasons, the term predictive value is best avoided in favor of the more descriptive posttest probability following a positive or a negative test result. The nomogram version of Bayes' rule (Fig. 3-2) helps us to understand at a conceptual level 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.

The *likelihood ratio for a positive test* 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 9 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

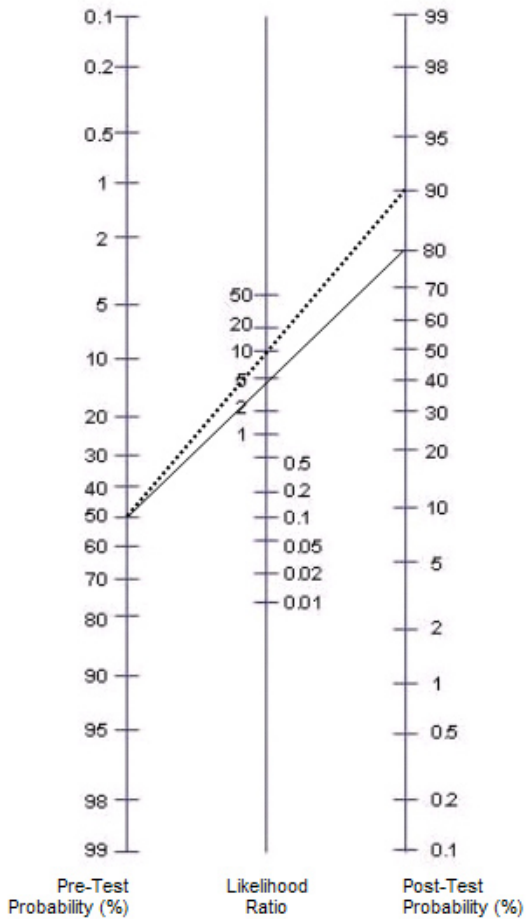


Figure 3-2: Nomogram version of Bayes' theorem used to predict the posttest probability of disease (right-hand scale) using the pretest probability of disease (left-hand scale) and the likelihood ratio for a positive test (middle scale). See text for information on calculation of likelihood ratios. To use, place a straightedge 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 CT 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/likelihood-ratios/>).

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 positive is 2.0).

The corresponding *likelihood ratio for a negative test* is the ratio of the false-negative rate to the true-negative rate [or $(1 - \text{sensitivity})/\text{specificity}$]. Lower likelihood ratio negative 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 about ten 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 and an exercise single-photon emission CT (SPECT) myocardial perfusion imaging test (Chap. 236). Meta-analysis has shown that a positive treadmill ST-segment response has an average sensitivity of 60% and an average specificity of 75%, yielding a likelihood ratio positive of 2.4 $[0.60/(1 - 0.75)]$ (consistent with modest discriminatory ability because it falls between 2-5). For a 41-year old man with nonanginal pain and a 10% pretest probability of CAD, the posttest probability of disease after a positive result rises to only about 30%. For a 60-year old woman with typical angina and a pretest probability of CAD of 80%, a positive test result raises the posttest probability of disease to about 95%.

In contrast, exercise SPECT myocardial perfusion test is more accurate for diagnosis of CAD. For simplicity, assume that the finding of a reversible exercise-induced perfusion defect has both a sensitivity and a specificity of 90% (a bit higher than reported), yielding a likelihood ratio for a positive test of 9.0 [$0.90/(1 - 0.90)$] (consistent with intermediate discriminatory ability because it falls between 5-10). 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 provide only "positive" or "negative" results. Many tests have multi-dimensional outcomes (e.g., extent of ST-segment depression, exercise duration and exercise-induced symptoms with exercise testing). Although Bayes' theorem can be adapted to this more detailed test result format, it is computationally more complex to do so. Similarly, when multiple sequential 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, many texts assert that sensitivity and specificity are prevalence-independent parameters of test accuracy. This statistically useful assumption, however, is clinically simplistic. A treadmill exercise test, for example, has a sensitivity of around 30% in a population of patients with 1-vessel CAD, whereas its sensitivity in patients with severe 3-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, when used as presented above, is useful in studying diagnostic testing concepts but may prove too simplistic for use in actual patient care decisions. Predictions based on multivariable statistical models can more accurately address these more complex problems by simultaneously accounting for additional relevant patient

characteristics. In particular, these models explicitly account for multiple, even possibly overlapping, pieces of patient-specific information and assign a relative weight to each on the basis of its unique independent contribution to the prediction in question. For example, a logistic regression model to predict the probability of CAD ideally considers all the relevant independent factors from the clinical examination and diagnostic testing and their relative importance 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. Guideline driven treatment recommendations based on statistical prediction models available online, e.g., the ACC/AHA risk calculator for primary prevention with statins and the CHA₂DS₂-VASC calculator for anticoagulation for atrial fibrillation have generated more widespread usage. Whether the adoption of electronic health records will promote more use of predictive models in clinical practice and increase their impact on clinical encounters and outcomes remains unclear.

One reason for limited clinical use is that, to date, only a handful of prediction models have been validated properly (for example, Wells' criteria for pulmonary embolism, see 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.

When statistical survival models in cancer and heart disease have been compared directly with clinicians' predictions, the survival models have been found to be more consistent, as would be expected but not always more accurate. On the other hand, comparison of clinicians with

websites and apps that generate lists of possible diagnoses to help patients with self-diagnosis found that physicians outperformed the currently available programs. For students and less-experienced clinicians, the biggest value of diagnostic decision support may be in extending diagnostic possibilities and triggering "rational override" but their impact on knowledge, information-seeking, and problem-solving needs additional research.

TABLE 3-2 Wells Clinical Prediction Rule for Pulmonary Embolism

Clinical Feature	Points
Clinical signs of deep vein thrombosis	3
Alternative diagnosis is less likely than PE	3
Heart rate > 100 beats per minute	1.5
Immobilization ≥3d or surgery in previous four weeks	1.5
History of deep vein thrombosis or pulmonary embolism	1.5
Hemoptysis	1
Malignancy (with treatment within 6 months) or palliative	1
Interpretation	
Score >6.0	High
Score 2.0 to 6.0	Intermediate
Score <2.0	Low

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, computers offer several levels of potentially useful support for clinicians. At the most basic level, they provide ready access to vast reservoirs of information, which may, however, be quite difficult to sort through to find what is needed. At higher levels, computers can support care management decisions by making accurate predictions of outcome, or can simulate the whole decision process, and provide algorithmic guidance. Computer-based predictions using Bayesian or statistical regression models inform a clinical decision but do not actually reach a “conclusion” or “recommendation.” Machine learning methods are being applied to pattern recognition tasks such as the examination of skin lesions and the interpretation of x-rays. 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 or appropriate utilization criteria, to direct clinical practice. In general, however, decision support systems have had little impact on practice. Reminder systems built into electronic health records have shown the most promise, particularly in correcting drug dosing and promoting adherence to guidelines. Checklists may also help avoid or reduce errors.

Decision Analysis

Compared with the decision support methods above, decision analysis

represents a normative prescriptive approach to decision making in the face of uncertainty. Its principal application is in complex decisions. For example, public health policy decisions often involve *trade-offs* in length versus quality of life, benefits versus resource use, population versus individual health, and *uncertainty* regarding efficacy, effectiveness, and adverse events as well as *values* or preferences regarding mortality and morbidity outcomes.

One recent analysis using this approach involved the optimal screening strategy for breast cancer, which has remained controversial, in part because a randomized controlled trial to determine when to begin screening and how often to repeat screening mammography is impractical. In 2016, the National Cancer Institute sponsored Cancer Intervention and Surveillance Network (CISNET) examined 8 strategies differing by whether to initiate mammography screening at age 40, 45, or 50 years and whether to screen annually, biennially, or annually for women in their 40s and biennially thereafter (hybrid). The 6 simulation models found biennial strategies to be the most efficient for average-risk women. Biennial screening for 1000 women from age 50 to 74 years versus no screening avoided 7 breast cancer deaths. Screening annually from age 40 to 74 years avoided 3 additional deaths but required 20,000 additional mammograms and yielded 1988 more false-positive results. Factors that influenced the results included patients with a 2-4 fold higher risk for developing breast cancer in whom annual screening from 40 to 74 yielded similar benefits as biennial screening from age 50 to 74. For average-risk patients with moderate or severe co-morbidities, screening could be stopped earlier at ages 66 to 68 years.

This analysis involved 6 models that reproduced epidemiologic trends and a screening trial result, accounted for digital technology and treatments advances, and considered quality of life, risk factors, breast density, and comorbidity. It 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 has yet to be demonstrated.

Diagnosis as An Element of Care

High quality medical care begins with accurate diagnosis. The incidence of diagnostic errors has been estimated by a variety of methods including postmortem exams, medical record reviews, and medical malpractice claims, with each yielding complementary but different estimates of this quality of care patient-safety problem. In the past, diagnostic errors tended to be viewed as a failure of individual clinicians. The modern view is that they are mostly system of care deficiencies. Current estimates suggest that nearly everyone will experience at least one diagnostic error in their lifetime, leading to mortality, morbidity, unnecessary tests and procedures, costs and anxiety.

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. While checklists have been shown useful in certain medical contexts, such as the ORs and ICUs, 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). Evidence-based medicine (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 can be 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. However, the medical literature is now being flooded with systematic reviews of varying quality and clinical utility. Therefore, systematic reviews should be used in conjunction with selective reading of some of the best empirical studies.

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, locating reports quickly and efficiently that are on point in a sea of irrelevant or unhelpful citations remains difficult, and important studies are easily missed. Systematic reviews, the second form, are regarded by some as the highest level of evidence in the hierarchy because they are intended to comprehensively summarize the available evidence on a particular topic. To avoid the potential biases found in review articles, predefined reproducible

explicit search strategies and inclusion and exclusion criteria seek 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 is used to quantitatively summarize 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.

TABLE 3-3 Selected Tools for Finding the Evidence in Evidence-Based Medicine

Name	Description	Web Address	Availability
Evidence-Based Medicine Reviews	Comprehensive electronic database that combines and integrates: 1. The Cochrane Database of Systematic Reviews 2. ACP Journal Club 3. The Database of Abstracts of Reviews of Effectiveness	www.ovid.com	Subscription required. Available through medical center libraries and other institutions.
Cochrane Library	Collection of EBM databases, including The Cochrane Database of Systematic Reviews—full text articles reviewing specific health care topics.	www.cochrane.org	Subscription required. Abstracts of systematic reviews available free online. Some countries have funding to provide free access to all residents.
ACP Journal Club	Collection of summaries of original studies and systematic reviews. Published bimonthly. All data since 1991 available on Web site, updated yearly.	www.acpjc.org	Subscription required.
Clinical Evidence	Monthly updated directory of concise overviews of common clinical interventions.	www.clinicalevidence.com	Subscription required. Free access for United Kingdom and developing countries.
MEDLINE	National Library of Medicine database with citations back to 1966.	www.nlm.nih.gov	Free via Internet.

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 or therapeutic benefit, 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., different types of cancer, acute or chronic CAD, pacemaker capture or chronic 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).

By definition, in observational data, the investigator does not control patient care. Carefully collected prospective observational data, however, can at times 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 and 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 and heterogeneity exists in clinical practice to ensure that similar patients will be managed differently by diverse 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 coronary artery disease 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 medically treated patients with left main CAD 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 potentially eligible subjects or if the nominal eligibility for the trial describe a very heterogeneous population.

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 (such as the use of a placebo control) 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 it is used to examine nonrandomized studies, meta-analysis is most useful for summarizing all randomized trials examining a particular therapy. Ideally, unpublished trials should be identified and included to avoid publication bias (i.e., missing “negative” trials which may not be published). Furthermore, the best meta-analyses obtain and analyze individual patient-level data from all trials rather than using only the summary data from published reports. 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 deficiencies in primary trial data or trial design.

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 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 absolute risk reduction would be $12\% - 8\% = 4\%$ and the NNT would be $1/.04$, 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. In addition, the NNT concept assumes a homogeneity in response to treatment that may not be accurate. The NNT is simply another way of summarizing the absolute treatment difference and does not provide any unique information.

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., diabetes mellitus, 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 which 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 guideline level 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 expertise and 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 insufficient or 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 National Academy of Medicine (formerly Institute of Medicine) has made methodological 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 provider 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 and 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.

FURTHER READING

- CROSKERRY P: A universal model of diagnostic reasoning. *Acad Med* 84:1022, 2009.
- DHALIWAL G, DETSKY AS: The evolution of the master diagnostician. *JAMA* 310:579, 2013.
- HUNINK MGM et al: *Decision Making in Health and Medicine: Integrating Evidence and Values*. 2nd Ed. Cambridge, Cambridge University Press, 2014.
- KASSIRER JP et al: *Learning Clinical Reasoning*. 2nd Ed., Baltimore, Lippincott Williams & Wilkins, 2009.
- KAHNEMAN D: *Thinking Fast and Slow*. New York, Farrar, Straus and Giroux, 2013.
- MANDELBLATT JS et al: Collaborative modeling of the benefits and harms of associated with different U.S. breast cancer screening strategies. *Ann Intern Med* 164:215, 2016.
- MCDONALD KM et al: Patient safety strategies targeted at diagnostic errors. A systematic review. *Ann Intern Med* 158:381, 2013.
- NORMAN GR et al: The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad Med* 92:23, 2017.
- NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. *Improving Diagnosis in Health Care*. Washington, DC: The National Academies Press, 2015.
- SAPOSNIK G et al: Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak* 16:138, 2016.