

Challenges for evidence-based diagnosis

Introduction

We wrestled for a long time with the question of whether to include the term “evidence-based” in the title of this book. Although both of us are firm believers in the principles and goals of evidence-based medicine (EBM), we also knew that the term “evidence-based” would be viewed negatively by some potential readers (Grahame-Smith 1995; Lancet 1995; Healy 2006). We decided to keep “evidence-based” in the title and use this chapter to directly address some of the criticisms of EBM, many of which we believe have merit. We also recognize that, as elegant and satisfying as evidence-based diagnosis is, there are some very real barriers to applying it in a clinical setting. These barriers are the second topic of this chapter. Finally, we end the book with some thoughts on the future of evidence-based diagnosis and why it will be increasingly important.

Criticisms of evidence-based medicine

1. EBM overvalues randomized blinded trials and denigrates other forms of evidence, including clinical experience

EBM is frequently misrepresented as requiring randomized blinded trials (or better yet, a systematic review of such trials) to prove that a treatment is useful. This has been humorously illustrated in a “systematic review” of “Parachute Use to Prevent Death and Major Trauma Related to Gravitational Challenge” (Smith and Pell 2003). The authors found no controlled trials of parachute use for the “gravitationally challenged” (people jumping out of airplanes) and concluded that “everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute.”

We admit this criticism finds some resonance with us, which was one of the reasons for including Chapter 10 (“Alternatives to Randomized Trials”) in this book.

However, the solution is not to dismiss EBM; rather, it is to help its users to understand better the strengths and limitations of different types of evidence. Although we favor healthy skepticism about results of observational studies, particularly studies of treatments, EBM should not and does not require randomized blinded trials to prove the effectiveness of every treatment. We do not require randomized trials to prove that people with myopia see better with glasses, that electric shock works for ventricular fibrillation, or that antibiotics help patients with bacterial pneumonia (Gorman 2007). However, most treatments are not as obviously effective as wearing glasses. Often, the effectiveness of a treatment is doubtful enough to justify a randomized controlled trial. In this case, the results of a properly done trial will generally trump the observational results.

2. Evidence-based treatment recommendations tend toward the nihilistic

Related to the criticism that EBM insists too much on randomized trials is the concern that EBM either recommends against or fails to recommend treatments, when many people believe they work. Although we sympathize with patients and clinicians who find uncertainty uncomfortable and who appreciate being told what to do, we are distressed by a sense of paternalism and intellectual dishonesty that accompanies recommendations for tests and treatments that go far beyond available data, often making assumptions about patients' values that may be unwarranted. This is particularly problematic when those making the recommendations have a conflict of interest (Hayward 2008), as described in Chapter 6 ("Screening Tests").

A particularly contentious area for EBM is cancer screening. Results of randomized controlled trials suggest that mammographic screening for asymptomatic breast cancer in women aged 40 to 49 has only a small effect on breast cancer mortality (RRR ~15%; estimated NNT to prevent 1 death over 14 years is 1792; Humphrey et al. 2002), with a significant risk (20% to 56% for 10 mammograms) of false-positive results (Armstrong et al. 2007). The low prevalence of breast cancer in this age group, combined with the inaccuracy of the test, means that false positives and the consequent costly and uncomfortable biopsies will occur much more often than true positives. There is also the problem of pseudodisease; we know that some biopsy-proven breast "cancers" never progress to overt disease, but will nonetheless be treated as cancer. Reviewing the data in 1997, an expert panel convened by the National Institutes of Health recommended neither for nor against screening mammography in women aged 40 to 49. Instead, the panel recommended that a woman in this age group be counseled about potential benefits and harms before making her own decision about mammography (NIH 1997). Specialty groups, such as the American College of Radiology and the American College of Obstetricians, and disease-specific advocacy groups, such as the American Cancer Society, disagreed and recommended screening mammography in women aged 40 to 49. In 2006, former director of the National Institutes of Health Bernadine Healy summarized this controversy as part of a *U.S. News and World Report* critique of EBM (Healy 2006):

Remember the mammogram wars over whether women should get them during their 40s? The protagonists were the EBM-ers who said no and the radiologists and oncologists who said yes.

For the naysayers, randomized clinical trials were inadequate to show that the test saved lives, even though it did detect cancers sooner. Such a mammogram program would be costly, and unnecessary biopsies for false positive readings even costlier. But based on their interpretation of clinical evidence, cancer experts maintained that the test saved lives. What's more, they factored in the nature of the disease: more aggressive in younger women and best cured if picked up early. But in 1997 the Department of Health and Human Services gave a thumbs down to recommending that women start having mammograms in their 40s. Women promptly exercised their political clout, which led to an HHS reversal. (In fact, the trend has been for more screening in this age group, not less.)

It is instructive to note that Dr. Healy characterizes as a “thumbs down” the 1997 panel's recommendation that the screening decision be individualized. To some, particularly those concerned about reimbursement by third-party payers (see Criticism #3 below), recommending that a treatment decision be individualized appears to be the same as recommending against it. Dr. Healy failed to make this distinction when discussing prostate cancer screening as well (Box 12.1).

EBM provides an approach to validating research studies and quantitative methods for summarizing their results. Using this approach and these methods, different groups can arrive at different answers about the utility of a treatment or screening program, depending on their prior probabilities and values. (There is no right answer to the question of how many additional false-positive mammograms it is worth to get one more true-positive.) When a group that identifies itself as using the methods of EBM comes to a conclusion with which we disagree, we should review the evidence and how EBM was applied, not blame EBM for the conclusion we do not like.

3. EBM has been or might be used by payers as an excuse to deny payment and limit clinician autonomy

A 2007 *Time Magazine* article on EBM says that “insurance companies have been very aggressive in using evidence-based arguments to deny payment for untested treatments” (Gorman 2007). The plaintiff's lawyer in the Merenstein case (Box 12.1) defined EBM as purely a cost-saving method.

We share the concern that the language and methods of EBM may be misappropriated by organizations for which maximizing profit, rather than health, is the goal. A problem arises when the standards of evidence devised to determine whether to recommend population-wide preventive health interventions (which, for reasons described in Chapter 6, should be conservative) are applied to decisions about whether third-party payers will pay for particular tests and treatments (Woolf and George 2000). In this case, judgments and recommendations aimed at preserving physician and patient autonomy can end up preserving neither, if reimbursement for the desired care is denied. On the other hand, the perceived need to force third-party payers to provide coverage may lead to guidelines that are overly aggressive, leading to excess treatment, loss of patient and clinician autonomy, and creation of liability where none should exist (Newman and Maisels 2000).

It will always be necessary to set priorities for the allocation of limited health care resources. Efforts to control health care costs pre-dated EBM and would continue regardless of whether EBM existed. If payers did not use (or claim to use) the

Box 12.1: EBM as malpractice

We recommend the *JAMA* essay “Winners and Losers” by Dr. Daniel Merenstein (2004), in which he describes his experience of being sued for not obtaining a prostate-specific antigen (PSA) test on a 53-year-old man in 1999. The plaintiff he had not screened was diagnosed in 2002 with incurable prostate cancer. The balance of benefits and risks for PSA testing for prostate cancer in 1999 was at least as questionable as for mammography in women aged 40 to 49 (US Preventive Services Task Force 2002). False positives lead to unnecessary biopsies, and treatments for indolent cancers (pseudodisease) carry the risk of death, incontinence, and impotence. If the patient is unfortunate enough, as in this case, to have an aggressive cancer, it is unclear whether early diagnosis prolongs life, although for the reasons described in Chapter 6, it will appear to do so. As with the NIH panel’s recommendation about mammography, the evidence-based recommendation for PSA screening was, in Merenstein’s words, “discussing with the patient the risks and benefits, providing thorough informed consent, and coming to a shared decision.” Merenstein had documented this discussion and the shared decision not to obtain the PSA test. The plaintiff’s lawyer showed that most doctors in the state would have obtained the PSA test without discussing the risks and benefits with the patient. In his closing arguments, the plaintiff’s lawyer also put EBM on trial:

He threw EBM around like a dirty word and named the residency and me as believers in EBM, and our experts as founders of EBM. . . . He urged the jury to return a verdict to teach residencies not to send any more residents on the street believing in EBM (Merenstein 2004).

The jury found that Merenstein was not liable, but the residency program that trained him in evidence-based practice was – for \$1 million – despite the lack of evidence that an earlier diagnosis would have made any difference to the patient.

In her *U.S. News and World Report* essay critical of EBM, here is how Bernadine Healy summarized the case:

EBM also questions the prostate-specific antigen test, or PSA, for prostate cancer. The evidence-based method concludes that the test brings more harm than benefit, as it leads to unneeded biopsies and surgeries on often slow-growing cancers. This is at odds with the American Cancer Society, which says that men should have annual PSAs starting at age 50, and African-Americans, who have a higher prostate cancer rate, at age 45. This does not help that young primary-care doctor who published a mournful essay in the *Journal of the American Medical Association* in 2004. He did not get a PSA on his 53-year-old patient, based on his dutiful practice of evidence-based medicine. When found to have advanced prostate cancer, the patient sued and won. The jury put its faith in the medical experts who testified that PSAs are the best way to pick up tumors when they are most treatable.

We disagree with Dr. Healy’s framing of the issue. The question is not whether the PSA test is the best way to identify prostate tumors when they are most treatable. The questions are whether the expected benefit of the PSA test outweighs the risks of testing and overtreatment and whether patients should have any say in the decision to assume these risks.

methods of EBM to justify denying payment, they would rely on expert panels, common practice, and even more arbitrary justifications. The solution is not to attack EBM, but rather to attack third-party payers who use it inappropriately to limit reasonable care.

Barriers to the idealized process of evidence-based diagnosis

In Chapters 3 and 4, we learned how to use Bayes's Rule to calculate the post-test probability of disease based on pre-test probability and the LR associated with the test's result. This is an oversimplification of the diagnostic process. How should we estimate the pre-test probability? If we are considering classic diagnostic tests, such as x-rays or laboratories, then the pre-test probability is the probability of disease based on the population prevalence, the patient's history, and the physical exam. But if the test is a physical exam finding, then the pre-test probability is based on whatever information is available prior to examining for that finding. The post-test probability after one test can become the pre-test probability for the next test, but as we discussed in Chapter 8, unless the two tests are independent, the LRs of the results on each sequential test depend on the results of previous tests. Also, clinicians do not do all tests in series, they do many tests in parallel – that is, simultaneously. Finally, as we have pointed out before, the question is not, “Does this patient have disease X?” but “What disease is causing this patient's illness?” Tests help clinicians choose between multiple possible diagnoses.

We will discuss three barriers to the Bayesian process of diagnosis: 1) clinicians do not estimate or even understand probabilities very well; 2) oversimplification of diagnostic problems to fit the evidence-based model can lead to questionable conclusions; and 3) the process is impractical to apply on a patient-by-patient basis.

Clinicians, probability, and cognitive errors

A barrier for evidence-based diagnosis is that clinicians, like most people, do not estimate or even understand probabilities very well. They show wide variability, inconsistency, and irrationality in their estimates of probabilities. Even when given the pre-test probability, they do not properly use the test result and its LR to calculate post-test probabilities. Interestingly, however, asking clinicians, not for a probability, but for a *clinical decision*, reduces variability, inconsistency, and irrationality.

Errors in pre-test probability estimates

Several surveys have shown that different physicians given the same clinical vignette will provide widely different estimates for the probability of disease (Phelps and Levitt 2004; Dolan et al. 1986; Cahan et al. 2003, 2005). In one such survey, Cahan et al. (2003) gave clinicians the history, physical exam, and ECG description of a 58-year-old woman with chest pain. They asked for the probability of multiple different possible diagnoses, including active coronary artery disease, thoracic aortic dissection, esophageal reflux, and biliary colic. The probability estimates for any given diagnosis in the differential varied widely between clinicians. The estimated

Box 12.2: Bias

We have used the term “bias” many times in this book. The dictionary definition of “bias” is: “deviation of the expected value of an estimate from the quantity it estimates.” In discussing a bias, we should be clear about what quantity is being systematically under- or overestimated. We have previously discussed multiple biases in clinical research that distort estimates of test accuracy or treatment efficacy. Now, we are discussing biases in clinicians’ subjective estimates of disease probability – biases that arise from the use of heuristics. In the literature, the bias is named according to the heuristic from which it results: Representativeness Bias, Availability Bias, and Anchoring Bias.

probability of active coronary artery disease ranged from 1% to 99% with a median of 65% and an interquartile range of 30%. Moreover, the probabilities assigned by an individual physician to each diagnosis in the differential usually summed to much greater than 100%, even though the diagnoses were supposed to be mutually exclusive.

In a classic paper, Tversky and Kahneman (1974) pointed out that we all have difficulty dealing with probabilities and simplify the complex task of assessing probabilities by using heuristics that can lead to severe and systematic errors (i.e., bias; see Box 12.2). A “heuristic” is a rule of thumb used to simplify a problem, such as estimation of a quantity or probability, sometimes at the expense of precision and accuracy. Tversky and Kahneman’s example of a heuristic is the subjective estimate of an object’s distance from the viewer based on its visual clarity. This leads to overestimates of distance on foggy days and underestimates on clear days. Tversky and Kahneman described three heuristics commonly used to estimate probabilities: representativeness, availability, and adjustment from an anchor. Use of these heuristics can result in biased estimates of disease probability.

Representativeness. One of the heuristics used in estimating probability is representativeness, in which likelihood is confused with similarity. In medicine, if a clinical presentation is similar to the typical presentation of a rare disease, the clinician will assign that rare disease a high probability, without taking into account its very low prior probability. For example, among patients who present with chest pain, acute cardiac ischemia is thought to be between 50 and 500 times more likely than thoracic aortic dissection (Burt 1999; Kohn et al. 2005). Because of this, even if the chest pain has a characteristic typical of aortic dissection, such as radiation to the back, the probability of cardiac ischemia may still be at least as high as the probability of aortic dissection. However, many physicians will assign a much higher likelihood to dissection than to ischemia.

Availability. Availability is another heuristic used to estimate probabilities. Availability refers to the ease with which instances or occurrences of an event can be brought to mind. Of course, representativeness may be one contributor to availability: the presence of classic symptoms of a rare disease may make it available in memory.

However, other factors affect availability as well. For example, recent events are likely to be more available than earlier events. The Tversky and Kahneman (1974) article points out that “the subjective probability of traffic accidents rises temporarily when one sees a car overturned by the side of the road.” An emergency physician is more likely to assign a high probability to aortic dissection if a case was discussed at the last department conference.

One’s own experience is obviously more available than the experience of others. For example, surgeons at a hospital were asked to estimate overall (hospital-wide) surgical mortality. The estimates of surgeons from high-mortality specialties (e.g., neurosurgeons) were at least double the estimates of surgeons from low-mortality specialties (e.g., plastic surgeons). Thus, the mortality rate from personally performed operations exerted a disproportionate influence on judgment about the whole hospital’s surgical mortality rate (Detmer et al. 1978).

Clinicians often overestimate the probability of a diagnosis with severe consequences because of the anticipated regret if the diagnosis were missed (Bornstein and Emler 2001). This could also be classified as “regret bias.” Kahneman and Tversky did not use the term “regret bias,” but it is related to use of the availability heuristic, because diagnoses with severe consequences are often more easily brought to mind. We mentioned the Cahan study in which clinicians were surveyed about likely diagnoses in a 58-year-old woman with 2 days of “episodic pressing/burning chest pain.” The clinicians assigned aortic dissection a mean probability of 16%, whereas more common (and more likely) problems such as reflux and anxiety were assigned lower probabilities. Perhaps this was because failing to diagnose reflux or anxiety has minor consequences compared with failing to diagnose aortic dissection. When asked for the probability of a particular diagnosis, clinicians usually respond with their level of concern – not the actual likelihood.

Similarly, in Chapter 2 on reliability, we suggested that the same radiologist interpreting the same set of x-rays might be systematically more likely to rate them as abnormal after being sued for missing an abnormality. This is because the lawsuit makes the abnormality more *available* to the radiologist either by increasing its subjective probability or because the level of concern has increased.

Adjustment from an anchor. A third heuristic discussed by Tversky and Kahneman is to estimate a probability by starting from an initial value, called the “anchor,” and adjusting to reach a final answer. As we shall see, even when the initial value is meaningful, adjustment can be inadequate. But this heuristic is especially problematic when the initial anchor is irrelevant.

For example, Brewer et al. (2007) presented to family physicians (via a mailed survey) a clinical vignette about a 32-year-old woman with cough, pleuritic chest pain, and low-grade fever. First, they established an irrelevant anchor. Half the participants were asked whether the chance of pulmonary embolism was greater or less than 1%; the other half were asked whether the chance was greater or less than 90%. Then, all the participants were asked to give a point estimate of the probability

of pulmonary embolism. Physicians in the low anchor group estimated the likelihood of pulmonary embolism at 23% on average, whereas physicians in the high anchor group estimated the likelihood at 53%.

Even when the anchor has some relevance, we can underadjust our probability estimates. In any emergency department, there are critical care rooms for the sickest patients and areas (often multipatient rooms or wards) for the lowest acuity patients. A triage nurse places the patient in an area prior to the emergency physician's initial evaluation. In general, as the emergency physician approaches a patient in the low-acuity area, the chance of a serious condition requiring hospitalization is very low. But, one of the more common errors in emergency medicine is to send a patient home from the multipatient ward who would have been admitted from a critical care room. This is probably anchoring bias; the patient in the ward starts out with a low probability of serious illness that the emergency physician insufficiently adjusts upward. This phenomenon has also been called "triage cueing bias" (Croskerry and Wears 2003).

Errors in post-test probability estimates

The discussion of adjusting from an anchor and its possible effect on pre-test probability estimates naturally leads to a discussion of cognitive bias in test interpretation. As mentioned in the introduction to this section, the pre-test probability for one test can be the post-test probability from a prior test, and many tests are done in parallel rather than in series. Because of this, the distinction between cognitive bias in test interpretation and cognitive bias in estimating pre-test probabilities is somewhat arbitrary. Attempts have been made to name the cognitive biases that contribute to our misinterpretation of test results (Dawson and Arkes 1987). For example, "confirmation bias" consists of cognitive "cherry-picking"; unconsciously, we both pay more attention to test results that support our initial impression and misinterpret nonspecific findings as confirmatory. "Premature closure" is choosing (and often labeling a patient with) a specific diagnosis that is not sufficiently supported by the test results. This can occur because of the patient's or our own discomfort with uncertainty. Confirmation bias and premature closure can be especially problematic if we stake ego on our initial impression by mentioning it to the patient or a member of the house staff, or if we are fatigued or under time pressure. Unlike representativeness and availability bias, cognitive errors in test interpretation, such as confirmation bias and premature closure, do not arise from commonly used heuristics for estimating probabilities.

Intuition versus math. Anchoring bias occurs when we are influenced by an irrelevant anchor or underuse new information to adjust from a relevant anchor. On the other hand, we often tend to overadjust probabilities of disease based on positive test results. Recall the example in Chapter 3 of a positive screening mammogram in a 45-year-old woman. The prevalence of breast cancer was 2.8/1,000. Before we teach probability updating in our class, we ask our students to estimate the probability of cancer given the prevalence, test characteristics, and the positive mammogram. The answers tend

to exceed 50%. We saw in Chapter 3 that, assuming a sensitivity of 75% and a specificity of 93%, the actual answer is about 3%. This systematic bias is obviously not due to underadjustment from the anchor of 2.8/1,000. Rather, it represents failure to consider the very low pre-test probability. Using a mammography example similar to this one, Eddy (1982) concluded that physicians grossly overestimate probability of disease in patients with positive screening tests for rare diseases.

We spent Chapters 3 and 4 teaching you the mathematics of adjusting an initial pre-test probability to get a post-test probability, but as shown in Box 12.3, if you rely on your intuition instead of the math, you may get it wrong.

Box 12.3

Without doing any mental arithmetic, try answering the following problem adapted from Raiffa (1968):

There are two large, outwardly identical bags filled with red and white marbles. In Bag R3W1, 3/4 of the marbles are red and 1/4 are white. In Bag W3R1, 3/4 are white and 1/4 are red. You have been given one of the 2 bags, and are trying to figure out which one you got. You draw 12 marbles and find that 4 are white and 8 are red. (To simplify the math, we'll assume you drew them in that order, although it doesn't actually matter.) The bags hold thousands of marbles, so you don't have to worry about sampling without replacement versus with replacement. Keep in mind that the sample held predominantly *red* marbles. This obviously *lowers* likelihood that you started out with Bag W3R1, which has 3/4 white marbles. What is the probability that you got Bag W3R1, with its 3/4 white marbles and only 1/4 red marbles?

- a) >50%
- b) 35–50%
- c) 20–34%
- d) 5–19%
- e) <5%

Circle your answer before reading on.

The pre-test odds = 1:1.

The probability of the sample¹ given Bag W3R1, with its 3/4 white balls is

$$P(4 \text{ white, } 8 \text{ red} \mid W3R1) = \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \\ \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}.$$

$$P(4 \text{ white, } 8 \text{ red} \mid R3W1) = \left(\frac{3}{4}\right)^4 \times \left(\frac{1}{4}\right)^8$$

The probability of the sample given Bag R3W1, which contains only 1/4 white balls is

$$P(4 \text{ white, } 8 \text{ red} \mid W3R1) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \\ \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4}.$$

$$P(4 \text{ white, } 8 \text{ red} \mid R3W1) = \left(\frac{1}{4}\right)^4 \times \left(\frac{3}{4}\right)^8$$

¹ This is a sample with one specific sequence of 4 white and 8 red marbles. If we wanted the probability of *any* combination of 8 red and 4 white marbles, we would have to multiply by a factor of $(12 \times 11 \times 10 \times 9) / (4 \times 3 \times 2 \times 1)$, but this would be in both the numerator and denominator of the LR, and hence would cancel out anyway.

The LR is

$$\begin{aligned} \text{LR} &= \frac{P(4 \text{ white, 8 red} \mid W3R1)}{P(4 \text{ white, 8 red} \mid R3W1)} \\ \text{LR} &= \frac{(3/4)^4 \times (1/4)^8}{(3/4)^8 \times (1/4)^4} = \frac{(1/4)^4}{(3/4)^4} \\ &= \frac{1}{3^4} \\ &= 1/81 \end{aligned}$$

Post-test odds is pre-test odds times the LR

Post-test odds = $1 \times 1/81 = 0.012 \approx$ post-test probability.

So, the correct answer is (e).

How did you do? Most people exhibit insufficient adjustment from the pre-test probability of 50% and answer (b) or (c).

This is illustrated by a study by Sox et al. (2006), who asked pediatricians for the post-test probability of pertussis given a pre-test probability of 30% and a negative pertussis direct fluorescent antibody (DFA) test. One-third of the physicians were given the sensitivity (50%) and specificity (95%) of the DFA; one-third were given the test characteristics explained in nontechnical terms; and one-third received no information about test characteristics.

The correct post-test probability is 18%.² Two-thirds of the respondents estimated a post-test probability *higher than* the pre-test probability of 30%, despite the negative DFA result. This was *worse* in the two groups that were given the test's characteristics. We hope readers of this book would do better!

Probability estimates versus decision making

When clinicians estimate disease probabilities, we use heuristics that can result in significant biases. Also, despite the medical school and CME courses on clinical epidemiology and EBM, and despite the nomograms, slide-rules, and on-line calculators designed to make the process easier, many clinicians still cannot properly update pre-test probabilities based on the results of a diagnostic test. On the other hand, clinicians are probably more consistent and rational in their clinical decision making than they are in their probability estimates. In the literature on cognitive biases, this is the distinction between judgment (probability estimates) and choice (decision making) (Kahneman et al. 1982; Brewer et al. 2007).

Although experienced clinicians do not estimate or update probabilities well, they often make good decisions that take into account varying presentations and consequences of error. In the study of Brewer et al. (2007), physician estimates for the probability of pulmonary embolism were influenced by the irrelevant anchors of 1% and 90%. However, the authors went on to ask the physicians for a decision

² You can just about do this in your head. Convert 30% probability to pre-test odds of 3/7. Calculate $\text{LR}(-) = 50\%/95\% \approx 1/2$. Post-test odds = $3/7 \times 1/2 = 3/14$. Convert to post-test probability = $3/17 \approx 0.18$.

about next steps.³ Although the initial anchor affected probability estimates, it did not appear to affect the treatment decisions. In fact, the physicians in the low anchor group were slightly more aggressive about testing and treating for pulmonary embolism. Similarly, although doctors may not be very good at estimating the probability of serious illness, they may do at least as well as decision rules at deciding whom to admit and treat (Pantell et al. 2004; Tierney et al. 1985; Davison et al. 1990).

This is not to say that cognitive errors do not affect patient outcomes. In *How Doctors Think*, Groopman (2007) gives multiple examples of cognitive errors in diagnosis resulting from the biases mentioned above and contributed to by time pressure, fatigue, and cultural barriers. Groopman's choice for a book title is itself an indication of the importance of cognitive error in medicine. Of course, we all focus on the cognitive errors leading to "misses," failures to identify a serious diagnosis, which lead to the most dramatic stories. But more commonly, flawed thinking leads to overtesting, which is much more mundane. For example, unnecessary tests, like an obligatory throat culture after a negative rapid strep test (Problem 3.5), are often recommended because clinicians misunderstand and miscalculate the implications of imperfect sensitivity (a small but non-zero false-negative rate).

Oversimplification of the diagnostic problem

Another problem with evidence-based diagnosis is that its application sometimes entails oversimplification that leads to highly questionable conclusions.

Cardall et al (2004) recommend against obtaining a WBC count to determine whether a patient with abdominal pain has appendicitis, because it is "not clinically useful" for distinguishing between patients with and without appendicitis. But their study showed that a $WBC \geq 15,000/\mu L$ has an LR of 3.2 for appendicitis. Moreover, the study failed to adequately consider that the WBC count is a continuous test; a WBC count of 28,000/ μL or 500/ μL would appropriately affect a clinician's management decisions. Also, when confronted by a patient with abdominal pain, the question is not whether the patient has appendicitis; the questions are what the patient does have and whether additional testing (e.g., a CT scan) can help identify the problem. A markedly elevated WBC count is associated with other conditions, such as diverticulitis and small bowel obstruction, which are identifiable on CT. Finally, the study did not consider something that clinicians do consider – the WBC count is always part of a complete blood count, which provides a hematocrit and a platelet count as well, both of which may help with diagnosis and treatment decisions.

Experienced clinicians are also justifiably concerned about evidence-based diagnosis when a book on the subject pronounces the electrocardiogram as "useless testing" in identifying which patients with acute chest pain to admit to the hospital (Knottnerus and Van Weel 2002).

³ The choices were: normal care; lung scan; pulmonary angiogram; hospitalize; and treat with anticoagulant.

Making a multilevel test dichotomous or failure to adequately consider the full range of possible test results are oversimplifications addressed in Chapter 4. The multiplicity of possible diagnoses to explain a patient's illness is more difficult to accommodate.

Is evidence-based diagnosis practical?

The main problem with the step-by-step Bayesian process of evidence-based diagnosis is that it is impractical for clinicians to apply on a patient-by-patient basis. According to Croskerry (2002), in their "flesh-and-blood decision making," emergency physicians are not and cannot be formal Bayesians. Instead, they have developed several decision-making strategies that reduce decision complexity and build economy and redundancy into the process. The primary strategy is to make a treatment/disposition decision fairly soon after the presentation of a patient at the emergency department, or to commit to a formal work-up involving an array of tests, imaging techniques, and consultations.

MAK is an emergency physician in a nonteaching, community hospital who sees, without the interposition of medical students or residents, approximately 1,500 patients per year. Many of these patients, such as those with chest or abdominal pain, fever, shortness of breath, or altered mental status, present significant diagnostic problems. TBN is a pediatrician at a teaching hospital who attends in an urgent care clinic and a newborn nursery, where the prevalence of serious illness is lower, but diagnostic barriers still exist. If anyone were going to apply the step-by-step Bayesian approach to diagnosis, the co-authors of a book on evidence-based diagnosis would. But we almost never, on a patient-by-patient basis, estimate pre-test probabilities and then update them using the results of the tests that we order. We do use the basic logic of evidence-based diagnosis with many of the patients we see. For example, material covered in this text has helped us:

- decide not to order tests (e.g., a head CT on a child with a minor head injury) when the disease is so unlikely that the pain, risk, and cost (e.g., radiation exposure) of testing are not worth the negligible chance of a positive result.
- avoid ordering nonspecific tests (e.g., myeloperoxidase and C-reactive protein).
- accept some negative initial tests (e.g., rapid strep test or urinalysis) without ordering confirmatory tests (e.g., throat or urine cultures).
- interpret tests (e.g., BNP) along a whole range of possible values, rather than dichotomizing them as either positive or negative.
- act on mildly abnormal test results (e.g., a slightly elevated D-dimer or WBC count) when our level of concern is high but wait when we get the same results on patients about whom we are less concerned.
- become more aware of how our own biases and cognitive limitations affect our ability to diagnose and treat disease.

One way to address the practical barriers to the step-by-step Bayesian approach is to minimize the need for front-line clinicians to estimate or deal with probabilities by providing them with clinical decision rules and guidelines, to be discussed in the next section.



Figure 12.1 Example of direct-to-consumer advertising from an imaging center, sent via direct mail to TBN.

The future of evidence-based diagnosis

As we come to the end of this book, we cannot resist the temptation to speculate about the direction in which medical tests are moving, and how the material in this book might help readers keep up.

One direction seems clear: more and more new tests will be offered, and they will need to be critically evaluated. These tests will take advantage of advances in technology, particularly in genetics, molecular biology, and imaging. Increasingly, we fear, they may be promoted directly to consumers (Fig. 12.1), who are ill-equipped to critically evaluate the claims of the promoters.

Clinicians, already drowning in a sea of data, will increasingly rely on decision rules and guidelines, sometimes implemented as computer-based decision aids, to assist with deciding which tests to order and how to interpret the results. This will help to overcome both knowledge gaps about pre-test probabilities and LRs, as well as cognitive errors in probability estimation and updating. The authors of the decision rules and guidelines evaluate treatment effectiveness, determine test characteristics, estimate pretest probabilities, do the Bayesian updating for a range of clinical scenarios, and then provide their recommendations to clinicians. However, clinicians will need to be skeptical consumers of these decision rules and guidelines, just as they are of individual tests. As shown in this book, decisions about which tests to order depend not only on the costs and accuracy of tests, but on the efficacy and risks of different treatment options, and assessment of these may depend on the patient's values. For all of the reasons discussed in Chapter 6, it will be important to discern whose values and whose perspective are reflected in any such decision aids. The material in this text should help us select and interpret diagnostic and screening tests so as to maximize the benefit to our patients' health.

Summary

1. Evidence-based medicine has been criticized for being overly reliant on evidence from randomized controlled trials, overly skeptical about the efficacy of many

treatments, and an excuse for insurance companies to deny coverage for treatments. These valid concerns should give rise to caution about the application of EBM, not to its abandonment.

2. Evidence-based diagnosis as a step-by-step Bayesian process faces the challenge that clinicians often do not deal well with probabilities, either estimating pre-test probabilities or interpreting tests and calculating post-test probabilities. Despite this, experienced clinicians often make good clinical decisions.
3. However, with knowledge of evidence-based diagnosis and understanding of our cognitive biases and limitations, we can do even better.
4. Clinicians, as skeptical consumers, can use the methods of evidence-based diagnosis to evaluate and use the increasing number of individual tests, clinical decision rules, and practice guidelines that appear in the literature and the marketplace.

References

- Armstrong, K., E. Moye, et al. (2007). "Screening mammography in women 40 to 49 years of age: a systematic review for the American College of Physicians." *Ann Intern Med* **146**(7): 516–26.
- Bornstein, B. H., and A. C. Emler (2001). "Rationality in medical decision making: a review of the literature on doctors' decision-making biases." *J Eval Clin Pract* **7**(2): 97–107.
- Brewer, N. T., G. B. Chapman, et al. (2007). "The influence of irrelevant anchors on the judgments and choices of doctors and patients." *Med Decis Making* **27**(2): 203–11.
- Burt, C. W. (1999). "Summary statistics for acute cardiac ischemia and chest pain visits to United States EDs, 1995–1996." *Am J Emerg Med* **17**(6): 552–9.
- Cahan, A., D. Gilon, et al. (2003). "Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities?" *QJM* **96**(10): 763–9.
- Cahan, A., D. Gilon, et al. (2005). "Clinical experience did not reduce the variance in physicians' estimates of pretest probability in a cross-sectional survey." *J Clin Epidemiol* **58**(11): 1211–6.
- Cardall, T., J. Glasser, et al. (2004). "Clinical value of the total white blood cell count and temperature in the evaluation of patients with suspected appendicitis." *Acad Emerg Med* **11**(10): 1021–7.
- Croskerry, P. (2002). "Achieving quality in clinical decision making: cognitive strategies and detection of bias." *Acad Emerg Med* **9**(11): 1184–204.
- Croskerry, P., and R. Wears (2003). Safety errors in emergency medicine. In: *Emergency Medicine Secrets*. Markovchik. VJ and Pons. PT. Philadelphia, PA, Hanley and Belfus.
- Davison, G., A. L. Suchman, et al. (1990). "Reducing unnecessary coronary care unit admissions: a comparison of three decision aids [see comments]." *J Gen Intern Med* **5**(6): 474–9.
- Dawson, N. V., and H. R. Arkes (1987). "Systematic errors in medical decision making: judgment limitations." *J Gen Intern Med* **2**(3): 183–7.
- Detmer, D. E., D. G. Fryback, et al. (1978). "Heuristics and biases in medical decision-making." *J Med Educ* **53**(8): 682–3.
- Dolan, J. G., D. R. Bordley, et al. (1986). "An evaluation of clinicians' subjective prior probability estimates." *Med Decis Making* **6**(4): 216–23.
- Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In: *Judgment Under Uncertainty: Heuristics and Biases*. D. Kahneman, P. Slovic, and A. Tversky, editors. Cambridge, Cambridge University Press, pp. 249–267.
- Gorman, C. (2007). "Are Doctors Just Playing Hunches?" *Time Magazine*. February 15, 2007.

- Groopman, J. (2007). *How Doctors Think*. New York: Houghton Mifflin.
- Grahame-Smith, D. (1995). "Evidence based medicine: Socratic dissent [see comments]." *Br Med J* **310**(6987): 1126–7.
- Hayward, R. (2008). "Access to clinically-detailed patient information." *Med Care* **46**(3): 229.
- Healy, B. (2006). "Who Says What's Best?" *U.S. News and World Report*.
- Humphrey, L. L., M. Helfand, et al. (2002). "Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force." *Ann Intern Med* **137**(5 Part 1): 347–60.
- Kahneman, D., P. Slovic, et al. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, Cambridge University Press.
- Knottnerus, J. A., and C. Van Weel (2002). General Introduction: evaluation of diagnostic procedures. In: *The Evidence Base of Clinical Diagnosis*. J. A. Knottnerus, editor. London, BMJ Books, pp. 1–18.
- Kohn, M. A., E. Kwan, et al. (2005). "Prevalence of acute myocardial infarction and other serious diagnoses in patients presenting to an urban emergency department with chest pain." *J Emerg Med* **29**(4): 383–90.
- Lancet (1995). "Evidence-based medicine, in its place." *Lancet* **346**(8978): 785.
- Merenstein, D. (2004). "A piece of my mind. Winners and losers." *JAMA* **291**(1): 15–6.
- Newman, T. B., and M. J. Maisels (2000). "Less aggressive treatment of neonatal jaundice and reports of kernicterus: lessons about practice guidelines." *Pediatrics* **105**(1 Pt 3): 242–5.
- NIH (1997). "NIH Consensus Statement. Breast cancer screening for women ages 40–49." *NIH Consensus Statement* **15**(1): 1–35.
- Pantell, R. H., T. B. Newman, et al. (2004). "Management and outcomes of care of fever in early infancy." *JAMA* **291**(10): 1203–12.
- Phelps, M. A., and M. A. Levitt (2004). "Pretest probability estimates: a pitfall to the clinical utility of evidence-based medicine?" *Acad Emerg Med* **11**(6): 692–4.
- Raiffa, H. (1968). *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Reading, MA, Addison–Wesley.
- Smith, G. C., and J. P. Pell (2003). "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials." *Br Med J* **327**(7429): 1459–61.
- Sox, C. M., T. D. Koepsell, et al. (2006). "Pediatricians' clinical decision making: results of 2 randomized controlled trials of test performance characteristics." *Arch Pediatr Adolesc Med* **160**(5): 487–92.
- Tierney, W. M., B. J. Roth, et al. (1985). "Predictors of myocardial infarction in emergency room patients." *Crit Care Med* **13**(7): 526–31.
- Tversky, A., and D. Kahneman (1974). "Judgment under uncertainty: heuristics and biases." *Science* **185**: 1124–31.
- US Preventive Services Task Force (2002). "Screening for prostate cancer: recommendation and rationale." *Ann Intern Med* **137**(11): 915–6.
- Woolf, S. H., and J. N. George (2000). "Evidence-based medicine. Interpreting studies and setting policy." *Hematol Oncol Clin North Am* **14**(4): 761–84.

