Biomedical Decision Making: Probabilistic Clinical Reasoning

3

Douglas K. Owens and Harold C. Sox

After reading this chapter, you should know the answers to these questions:

- How is the concept of probability useful for understanding test results and for making medical decisions that involve uncertainty?
- How can we characterize the ability of a test to discriminate between disease and health?
- What information do we need to interpret test results accurately?
- What is expected-value decision making? How can this methodology help us to understand particular medical problems?
- What are utilities, and how can we use them to represent patients' preferences?
- What is a sensitivity analysis? How can we use it to examine the robustness of a decision and to identify the important variables in a decision?
- What are influence diagrams? How do they differ from decision trees?

D.K. Owens, MD, MS (✉)
VA Palo Alto Health Care System,
Palo Alto, CA, USA

Henry J. Kaiser Center for Primary Care
and Outcomes Research/Center for Health Policy,
Stanford University, Stanford, CA, USA
e-mail: owens@stanford.edu

H.C. Sox, MD, MACP
Dartmouth Institute, Geisel School of Medicine at
Dartmouth, Dartmouth College, 31 Faraway Lane,
West Lebanon, NH 03784, USA

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_3, © Springer-Verlag London 2014

3.1 The Nature of Clinical Decisions: Uncertainty and the Process of Diagnosis

Because clinical data are imperfect and outcomes of treatment are uncertain, health professionals often are faced with difficult choices. In this chapter, we introduce probabilistic medical reasoning, an approach that can help health care providers to deal with the uncertainty inherent in many medical decisions. Medical decisions are made by a variety of methods; our approach is neither necessary nor appropriate for all decisions. Throughout the chapter, we provide simple clinical examples that illustrate a broad range of problems for which probabilistic medical reasoning does provide valuable insight.

As discussed in Chap. 2, medical practice is medical decision making. In this chapter, we look at the process of medical decision making. Together, Chaps. 2 and 3 lay the groundwork for the rest of the book. In the remaining chapters, we discuss ways that computers can help clinicians with the decision-making process, and we emphasize the relationship between information needs and system design and implementation.

The material in this chapter is presented in the context of the decisions made by an individual clinician. The concepts, however, are more broadly applicable. Sensitivity and specificity are important parameters of laboratory systems that flag abnormal test results, of patient monitoring systems (Chap. 19), and of information-retrieval systems (Chap. 21). An understanding of what

probability is and of how to adjust probabilities after the acquisition of new information is a foundation for our study of clinical decision-support systems (Chap. 22). The importance of probability in medical decision making was noted as long ago as 1922:

> [G]ood medicine does not consist in the indiscriminate application of laboratory examinations to a patient, but rather in having so clear a comprehension of the probabilities and possibilities of a case as to know what tests may be expected to give information of value (Peabody 1922).

Example 1

You are the director of a blood bank. All potential blood donors are tested to ensure that they are not infected with the human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS). You ask whether use of the polymerase chain reaction (PCR), a gene-amplification technique that can diagnose HIV, would be useful to identify people who have HIV. The PCR test is positive 98 % of the time when antibody is present, and negative 99 % of the time antibody is absent.[1]

[1]The test sensitivity and specificity used in Example 1 are consistent with the reported values of the sensitivity and specificity of the PCR test for diagnosis of HIV early in its development (Owens et al. 1996b); the test now has higher sensitivity and specificity.

If the test is positive, what is the likelihood that a donor actually has HIV? If the test is negative, how sure can you be that the person does not have HIV? On an intuitive level, these questions do not seem particularly difficult to answer. The test appears accurate, and we would expect that, if the test is positive, the donated blood specimen is likely to contain the HIV. Thus, we are surprised to find that, if only one in 1,000 donors actually is infected, the test is more often mistaken than it is correct. In fact, of 100 donors with a positive test, fewer than 10 would be infected. There would be ten wrong answers for each correct result. How are we to understand this result? Before we try to find an answer, let us consider a related example.

Example 2

Mr. James is a 59-year-old man with coronary artery disease (narrowing or blockage of the blood vessels that supply the heart tissue). When the heart muscle does not receive enough oxygen (hypoxia) because blood cannot reach it, the patient often experiences chest pain (angina). Mr. James has twice undergone coronary artery bypass graft (CABG) surgery, a procedure in which new vessels, often taken from the leg, are grafted onto the old ones such that blood is shunted past the blocked region. Unfortunately, he has again begun to have chest pain, which becomes progressively more severe, despite medication. If the heart muscle is deprived of oxygen, the result can be a heart attack (myocardial infarction), in which a section of the muscle dies.

Should Mr. James undergo a third operation? The medications are not working; without surgery, he runs a high risk of suffering a heart attack, which may be fatal. On the other hand, the surgery is hazardous. Not only is the surgical mortality rate for a third operation higher than that for a first or second one but also the chance that surgery will relieve the chest pain is lower than that for a first operation. All choices in Example 2 entail considerable uncertainty. Furthermore, the risks are grave; an incorrect decision may substantially increase the chance that Mr. James will die. The decision will be difficult even for experienced clinicians.

These examples illustrate situations in which intuition is either misleading or inadequate. Although the test results in Example 1 are appropriate for the blood bank, a clinician who uncritically reports these results would erroneously inform many people that they had the AIDS virus—a mistake with profound emotional and social consequences. In Example 2, the decision-making skill of the clinician will affect a patient's quality and length of life. Similar situations are

commonplace in medicine. Our goal in this chapter is to show how the use of probability and decision analysis can help to make clear the best course of action.

Decision making is one of the quintessential activities of the healthcare professional. Some decisions are made on the basis of deductive reasoning or of physiological principles. Many decisions, however, are made on the basis of knowledge that has been gained through collective experience: the clinician often must rely on empirical knowledge of associations between symptoms and disease to evaluate a problem. A decision that is based on these usually imperfect associations will be, to some degree, uncertain. In Sects. 3.1.1, 3.1.2 and 3.1.3, we examine decisions made under uncertainty and present an overview of the diagnostic process. As Smith (1985, p. 3) said: "Medical decisions based on probabilities are necessary but also perilous. Even the most astute physician will occasionally be wrong."

3.1.1 Decision Making Under Uncertainty

Example 3

Mr. Kirk, a 33-year-old man with a history of a previous blood clot (thrombus) in a vein in his left leg, presents with the complaint of pain and swelling in that leg for the past 5 days. On physical examination, the leg is tender and swollen to midcalf—signs that suggest the possibility of deep vein thrombosis.[2] A test (ultrasonography) is performed, and the flow of blood in the veins of Mr. Kirk's leg is evaluated. The blood flow is abnormal, but the radiologist cannot tell whether there is a new blood clot.

[2]In medicine, a sign is an objective physical finding (something observed by the clinician) such as a temperature of 101.2 °F. A symptom is a subjective experience of the patient, such as feeling hot or feverish. The distinction may be blurred if the patient's experience also can be observed by the clinician.

Should Mr. Kirk be treated for blood clots? The main diagnostic concern is the recurrence of a blood clot in his leg. A clot in the veins of the leg can dislodge, flow with the blood, and cause a blockage in the vessels of the lungs, a potentially fatal event called a pulmonary embolus. Of patients with a swollen leg, about one-half actually have a blood clot; there are numerous other causes of a swollen leg. Given a swollen leg, therefore, a clinician cannot be sure that a clot is the cause. Thus, the physical findings leave considerable uncertainty. Furthermore, in Example 3, the results of the available diagnostic test are equivocal. The treatment for a blood clot is to administer anticoagulants (drugs that inhibit blood clot formation), which pose the risk of excessive bleeding to the patient. Therefore, clinicians do not want to treat the patient unless they are confident that a thrombus is present. But how much confidence should be required before starting treatment? We will learn that it is possible to answer this question by calculating the benefits and harms of treatment.

This example illustrates an important concept: Clinical data are imperfect. The degree of imperfection varies, but all clinical data—including the results of diagnostic tests, the history given by the patient, and the findings on physical examination—are uncertain.

3.1.2 Probability: An Alternative Method of Expressing Uncertainty

The language that clinicians use to describe a patient's condition often is ambiguous—a factor that further complicates the problem of uncertainty in medical decision making. Clinicians use words such as "probable" and "highly likely" to describe their beliefs about the likelihood of disease. These words have strikingly different meanings to different individuals. Because of the widespread disagreement about the meaning of common descriptive terms, there is ample opportunity for miscommunication.

The problem of how to express degrees of uncertainty is not unique to medicine. How is it handled in other contexts? Horse racing has its

share of uncertainty. If experienced gamblers are deciding whether to place bets, they will find it unsatisfactory to be told that a given horse has a "high chance" of winning. They will demand to know the odds.

The odds are simply an alternate way to express a probability. The use of probability or odds as an expression of uncertainty avoids the ambiguities inherent in common descriptive terms.

3.1.3 Overview of the Diagnostic Process

In Chap. 2, we described the hypothetico-deductive approach, a diagnostic strategy comprising successive iterations of hypothesis generation, data collection, and interpretation. We discussed how observations may evoke a hypothesis and how new information subsequently may increase or decrease our belief in that hypothesis. Here, we review this process briefly in light of a specific example. For the purpose of our discussion, we separate the diagnostic process into three stages.

The first stage involves making an initial judgment about whether a patient is likely to have a disease. After an interview and physical examination, a clinician intuitively develops a belief about the likelihood of disease. This judgment may be based on previous experience or on knowledge of the medical literature. A clinician's belief about the likelihood of disease usually is implicit; he or she can refine it by making an explicit estimation of the probability of disease. This estimated probability, made before further information is obtained, is the **prior probability** or **pretest probability** of disease.

Example 4
Mr. Smith, a 60-year-old man, complains to his clinician that he has pressure-like chest pain that occurs when he walks quickly. After taking his history and examining him, his clinician believes there is a high enough chance that he has heart disease to warrant ordering an exercise stress test. In the stress test, an electrocardiogram (ECG) is taken while Mr. Smith exercises. Because the heart must pump more blood per stroke and must beat faster (and thus requires more oxygen) during exercise, many heart conditions are evident only when the patient is physically stressed. Mr. Smith's results show abnormal changes in the ECG during exercise—a sign of heart disease.

How would the clinician evaluate this patient? The clinician would first talk to the patient about the quality, duration, and severity of his or her pain. Traditionally, the clinician would then decide what to do next based on his or her intuition about the etiology (cause) of the chest pain. Our approach is to ask the clinician to make his or her initial intuition explicit by estimating the pretest probability of disease. The clinician in this example, based on what he or she knows from talking with the patient, might assess the pretest or prior probability of heart disease as 0.5 (50 % chance or 1:1 odds; see Sect. 3.2). We explore methods used to estimate pretest probability accurately in Sect. 3.2.

After the pretest probability of disease has been estimated, the second stage of the diagnostic process involves gathering more information, often by performing a diagnostic test. The clinician in Example 4 ordered a test to reduce the uncertainty about the diagnosis of heart disease. The positive test result supports the diagnosis of heart disease, and this reduction in uncertainty is shown in Fig. 3.1a. Although the clinician in Example 4 chose the exercise stress test, there are many tests available to diagnose heart disease, and the clinician would like to know which test he or she should order next. Some tests reduce uncertainty more than do others (see Fig. 3.1b), but may cost more. The more a test reduces uncertainty, the more useful it is. In Sect. 3.3, we explore ways to measure how well a test reduces uncertainty, expanding the concepts of test sensitivity and specificity first introduced in Chap. 2.

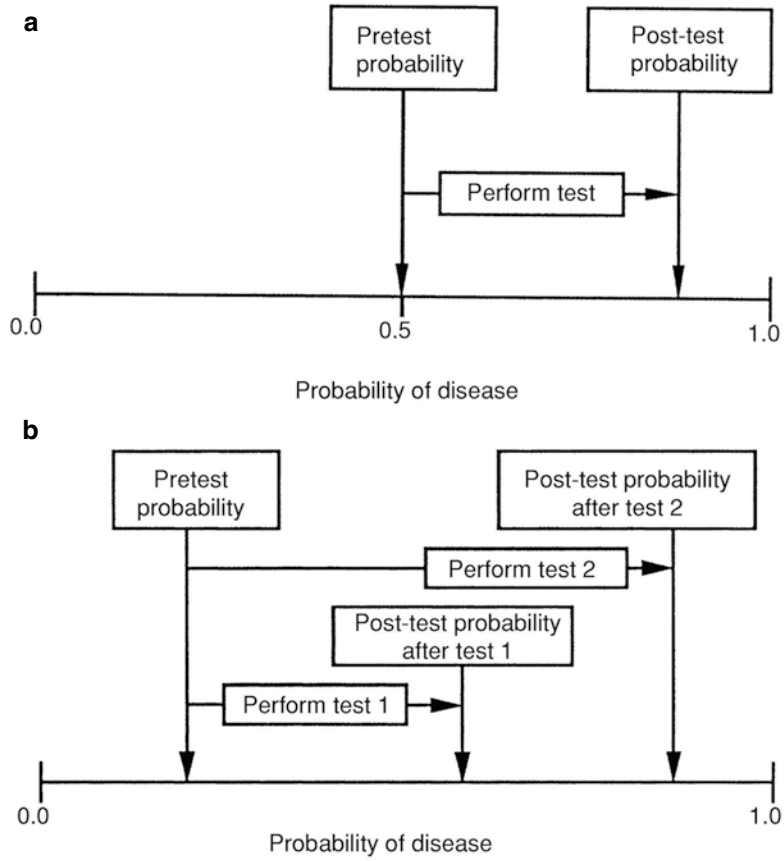

Fig. 3.1 The effect of test results on the probability of disease. (**a**) A positive test result increases the probability of disease. (**b**) Test 2 reduces uncertainty about presence of disease (increases the probability of disease) more than test 1 does

Given new information provided by a test, the third step is to update the initial probability estimate. The clinician in Example 4 must ask: "What is the probability of disease given the abnormal stress test?" The clinician wants to know the **posterior probability**, or **post-test probability**, of disease (see Fig. 3.1a). In Sect. 3.4, we reexamine Bayes' theorem, introduced in Chap. 2, and we discuss its use for calculating the post-test probability of disease. As we noted, to calculate post-test probability, we must know the pretest probability, as well as the sensitivity and specificity, of the test.[3]

[3]Note that pretest and post-test probabilities correspond to the concepts of prevalence and predictive value. The latter terms were used in Chap. 2 because the discussion was about the use of tests for screening populations of patients; in a population, the pretest probability of disease is simply that disease's prevalence in that population.

3.2 Probability Assessment: Methods to Assess Pretest Probability

In this section, we explore the methods that clinicians can use to make judgments about the probability of disease before they order tests. **Probability** is our preferred means of expressing uncertainty. In this framework, probability (p) expresses a clinician's opinion about the likelihood of an event as a number between 0 and 1. An event that is certain to occur has a probability of 1; an event that is certain not to occur has a probability of 0.[4]

The probability of event A is written p[A]. The sum of the probabilities of all possible, collectively exhaustive outcomes of a chance event must be equal to 1. Thus, in a coin flip,

[4]We assume a Bayesian interpretation of probability; there are other statistical interpretations of probability.

$$p[\text{heads}] + p[\text{tails}] = 1.0.$$

The probability of event A and event B occurring together is denoted by p[A&B] or by p[A,B].

Events A and B are considered **independent** if the occurrence of one does not influence the probability of the occurrence of the other. The probability of two independent events A and B both occurring is given by the product of the individual probabilities:

$$p[A,B] = p[A] \times p[B].$$

Thus, the probability of heads on two consecutive coin tosses is $0.5 \times 0.5 = 0.25$. (Regardless of the outcome of the first toss, the probability of heads on the second toss is 0.5.)

The probability that event A will occur given that event B is known to occur is called the **conditional probability** of event A given event B, denoted by p[A|B] and read as "the probability of A given B." Thus a post-test probability is a conditional probability predicated on the test or finding. For example, if 30 % of patients who have a swollen leg have a blood clot, we say the probability of a blood clot given a swollen leg is 0.3, denoted:

$$p[\text{blood clot} \mid \text{swollen leg}] = 0.3.$$

Before the swollen leg is noted, the pretest probability is simply the prevalence of blood clots in the leg in the population from which the patient was selected—a number likely to be much smaller than 0.3.

Now that we have decided to use probability to express uncertainty, how can we estimate probability? We can do so by either subjective or objective methods; each approach has advantages and limitations.

3.2.1 Subjective Probability Assessment

Most assessments that clinicians make about probability are based on personal experience. The clinician may compare the current problem to similar problems encountered previously and then ask: "What was the frequency of disease in similar patients whom I have seen?"

To make these subjective assessments of probability, people rely on several discrete, often unconscious mental processes that have been described and studied by cognitive psychologists (Tversky and Kahneman 1974). These processes are termed **cognitive heuristics**.

More specifically, a cognitive heuristic is a mental process by which we learn, recall, or process information; we can think of heuristics as rules of thumb. Knowledge of heuristics is important because it helps us to understand the underpinnings of our intuitive probability assessment. Both naive and sophisticated decision makers (including clinicians and statisticians) misuse heuristics and therefore make systematic—often serious—errors when estimating probability. So, just as we may underestimate distances on a particularly clear day (Tversky and Kahneman 1974), we may make mistakes in estimating probability in deceptive clinical situations. Three heuristics have been identified as important in estimation of probability:

1. Representativeness. One way that people estimate probability is to ask themselves: What is the probability that object A belongs to class B? For instance, what is the probability that this patient who has a swollen leg belongs to the class of patients who have blood clots? To answer, we often rely on the **representativeness** heuristic in which probabilities are judged by the degree to which A is representative of, or similar to, B. The clinician will judge the probability of the development of a blood clot (thrombosis) by the degree to which the patient with a swollen leg resembles the clinician's mental image of patients with a blood clot. If the patient has all the classic findings (signs and symptoms) associated with a blood clot, the clinician judges that the patient is highly likely to have a blood clot. Difficulties occur with the use of this heuristic when the disease is rare (very low prior probability, or prevalence); when the clinician's previous experience with the disease is atypical, thus giving an incorrect mental representation; when the patient's clinical profile is

atypical; and when the probability of certain findings depends on whether other findings are present.

2. Availability. Our estimate of the probability of an event is influenced by the ease with which we remember similar events. Events more easily remembered are judged more probable; this rule is the **availability** heuristic, and it is often misleading. We remember dramatic, atypical, or emotion-laden events more easily and therefore are likely to overestimate their probability. A clinician who had cared for a patient who had a swollen leg and who then died from a blood clot would vividly remember thrombosis as a cause of a swollen leg. The clinician would remember other causes of swollen legs less easily, and he or she would tend to overestimate the probability of a blood clot in patients with a swollen leg.
3. Anchoring and adjustment. Another common heuristic used to judge probability is **anchoring and adjustment**. A clinician makes an initial probability estimate (the anchor) and then adjusts the estimate based on further information. For instance, the clinician in Example 4 makes an initial estimate of the probability of heart disease as 0.5. If he or she then learns that all the patient's brothers had died of heart disease, the clinician should raise the estimate because the patient's strong family history of heart disease increases the probability that he or she has heart disease, a fact the clinician could ascertain from the literature. The usual mistake is to adjust the initial estimate (the anchor) insufficiently in light of the new information. Instead of raising his or her estimate of prior probability to, say, 0.8, the clinician might adjust it to only 0.6.

Heuristics often introduce error into our judgments about prior probability. Errors in our initial estimates of probabilities will be reflected in the posterior probabilities even if we use quantitative methods to derive those posterior probabilities. An understanding of heuristics is thus important for medical decision making. The clinician can avoid some of these difficulties by using published research results to estimate probabilities.

3.2.2 Objective Probability Estimates

Published research results can serve as a guide for more objective estimates of probabilities. We can use the prevalence of disease in the population or in a subgroup of the population, or clinical prediction rules, to estimate the probability of disease.

As we discussed in Chap. 2, the **prevalence** is the frequency of an event in a population; it is a useful starting point for estimating probability. For example, if you wanted to estimate the probability of prostate cancer in a 50-year-old man, the prevalence of prostate cancer in men of that age (5–14 %) would be a useful anchor point from which you could increase or decrease the probability depending on your findings. Estimates of disease prevalence in a defined population often are available in the medical literature.

Symptoms, such as difficulty with urination, or signs, such as a palpable prostate nodule, can be used to place patients into a **clinical subgroup** in which the probability of disease is known. For patients referred to a urologist for evaluation of a prostate nodule, the prevalence of cancer is about 50 %. This approach may be limited by difficulty in placing a patient in the correct clinically defined subgroup, especially if the criteria for classifying patients are ill-defined. A trend has been to develop guidelines, known as clinical prediction rules, to help clinicians assign patients to well-defined subgroups in which the probability of disease is known.

Clinical prediction rules are developed from systematic study of patients who have a particular diagnostic problem; they define how clinicians can use combinations of clinical findings to estimate probability. The symptoms or signs that make an independent contribution to the probability that a patient has a disease are identified and assigned numerical weights based on statistical analysis of the finding's contribution. The result is a list of symptoms and signs for an individual patient, each with a corresponding numerical contribution to a total score. The total score places a patient in a subgroup with a known probability of disease.

Example 5

Ms. Troy, a 65-year-old woman who had a heart attack 4 months ago, has abnormal heart rhythm (arrhythmia), is in poor medical condition, and is about to undergo elective surgery.

Table 3.1 Diagnostic weights for assessing risk of cardiac complications from noncardiac surgery

Clinical finding	Diagnostic weight
Age greater than 70 years	5
Recent documented heart attack	
>6 months previously	5
<6 months previously	10
Severe angina 20	
Pulmonary edema[a]	
Within 1 week	10
Ever	5
Arrhythmia on most recent ECG 5	
>5 PVCs	5
Critical aortic stenosis	20
Poor medical condition	5
Emergency surgery	10

Source: Modified from Palda et al. (1997)
ECG electrocardiogram, *PVCs* premature ventricular contractions on preoperative electrocardiogram
[a]Fluid in the lungs due to reduced heart function

What is the probability that Ms. Troy will suffer a cardiac complication? Clinical prediction rules have been developed to help clinicians to assess this risk (Palda and Detsky 1997). Table 3.1 lists clinical findings and their corresponding diagnostic weights. We add the diagnostic weights for each of the patient's clinical findings to obtain the total score. The total score places the patient in a group with a defined probability of cardiac complications, as shown in Table 3.2. Ms. Troy receives a score of 20; thus, the clinician can estimate that the patient has a 27 % chance of developing a severe cardiac complication.

Objective estimates of pretest probability are subject to error because of bias in the studies on which the estimates are based. For instance, published prevalence data may not apply directly to a particular patient. A clinical illustration is that early studies indicated that a patient found to have microscopic evidence of blood in the urine (microhematuria) should undergo extensive tests because a significant proportion of the patients would be found to have cancer or other serious diseases. The tests involve some risk, discomfort, and expense to the patient. Nonetheless, the approach of ordering tests for any patient with microhematuria was widely practiced for some years. A later study, however, suggested that the probability of serious disease in asymptomatic patients with only microscopic evidence of blood was only about 2 % (Mohr et al. 1986). In the past, many patients may have undergone unnecessary tests, at considerable financial and personal cost.

Table 3.2 Clinical prediction rule for diagnostic weights in Table 3.1

Total score	Prevalence (%) of cardiac complications[a]
0–15	5
20–30	27
>30	60

Source: Modified from Palda et al. (1997)
[a]Cardiac complications defined as death, heart attack, or congestive heart failure

What explains the discrepancy in the estimates of disease prevalence? The initial studies that showed a high prevalence of disease in patients with microhematuria were performed on patients referred to urologists, who are specialists. The primary care clinician refers patients whom he or she suspects have a disease in the specialist's sphere of expertise. Because of this initial screening by primary care clinicians, the specialists seldom see patients with clinical findings that imply a low probability of disease. Thus, the prevalence of disease in the patient population in a specialist's practice often is much higher than that in a primary care practice; studies performed with the former patients therefore almost always overestimate disease probabilities. This example demonstrates **referral bias**. Referral bias is common because many published studies are performed on patients referred to specialists.

Thus, one may need to adjust published estimates before one uses them to estimate pretest probability in other clinical settings.

We now can use the techniques discussed in this part of the chapter to illustrate how the clinician in Example 4 might estimate the pretest probability of heart disease in his or her patient, Mr. Smith, who has pressure-like chest pain. We begin by using the objective data that are available. The prevalence of heart disease in 60-year-old men could be our starting point. In this case, however, we can obtain a more refined estimate by placing the patient in a clinical subgroup in which the prevalence of disease is known. The prevalence in a clinical subgroup, such as men with symptoms typical of coronary heart disease, will predict the pretest probability more accurately than would the prevalence of heart disease in a group that is heterogeneous with respect to symptoms, such as the population at large. We assume that large studies have shown the prevalence of coronary heart disease in men with typical symptoms of angina pectoris to be about 0.9; this prevalence is useful as an initial estimate that can be adjusted based on information specific to the patient. Although the prevalence of heart disease in men with typical symptoms is high, 10 % of patients with this history do not have heart disease.

The clinician might use subjective methods to adjust his or her estimate further based on other specific information about the patient. For example, the clinician might adjust his or her initial estimate of 0.9 upward to 0.95 or higher based on information about family history of heart disease. The clinician should be careful, however, to avoid the mistakes that can occur when one uses heuristics to make subjective probability estimates. In particular, he or she should be aware of the tendency to stay too close to the initial estimate when adjusting for additional information. By combining subjective and objective methods for assessing pretest probability, the clinician can arrive at a reasonable estimate of the pretest probability of heart disease.

In this section, we summarized subjective and objective methods to determine the pretest probability, and we learned how to adjust the pretest probability after assessing the specific subpopulation of which the patient is representative. The next step in the diagnostic process is to gather further information, usually in the form of formal diagnostic tests (laboratory tests, X-ray studies, etc.). To help you to understand this step more clearly, we discuss in the next two sections how to measure the accuracy of tests and how to use probability to interpret the results of the tests.

3.3 Measurement of the Operating Characteristics of Diagnostic Tests

The first challenge in assessing any test is to determine criteria for deciding whether a result is normal or abnormal. In this section, we present the issues that you need to consider when making such a determination.

3.3.1 Classification of Test Results as Abnormal

Most biological measurements in a population of healthy people are continuous variables that assume different values for different individuals. The distribution of values often is approximated by the normal (gaussian, or bell-shaped) distribution curve (Fig. 3.2). Thus, 95 % of the population will fall within two standard deviations of the mean. About 2.5 % of the population will be more than two standard deviations from the mean at each end of the distribution. The distribution of values for ill individuals may be normally distributed as well. The two distributions usually overlap (see Fig. 3.2).

How is a test result classified as abnormal? Most clinical laboratories report an "upper limit of normal," which usually is defined as two standard deviations above the mean. Thus, a test result greater than two standard deviations above the mean is reported as abnormal (or positive); a test result below that cutoff is reported as normal (or negative). As an example, if the mean cholesterol concentration in the blood is 220 mg/dl, a clinical laboratory might choose as the upper

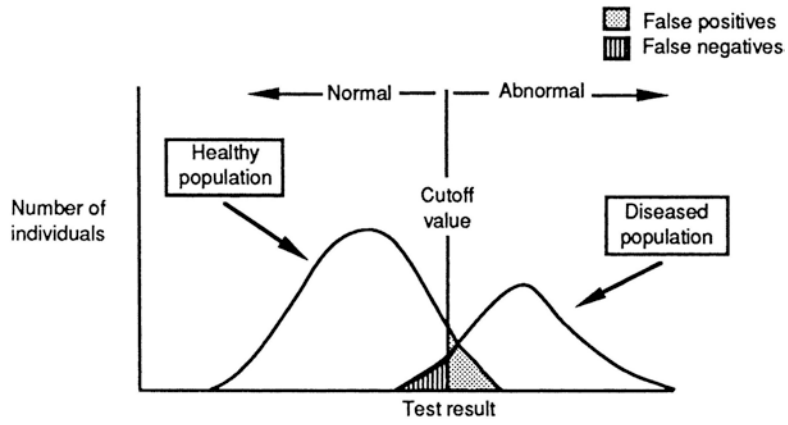

Fig. 3.2 Distribution of test results in healthy and diseased individuals. Varying the cutoff between "normal" and "abnormal" across the continuous range of possible values changes the relative proportions of false positives (FPs) and false negatives (FNs) for the two populations

limit of normal 280 mg/dl because it is two standard deviations above the mean. Note that a cutoff that is based on an arbitrary statistical criterion may not have biological significance.

An ideal test would have no values at which the distribution of diseased and nondiseased people overlap. That is, if the cutoff value were set appropriately, the test would be normal in all healthy individuals and abnormal in all individuals with disease. Few tests meet this standard. If a test result is defined as abnormal by the statistical criterion, 2.5 % of healthy individuals will have an abnormal test. If there is an overlap in the distribution of test results in healthy and diseased individuals, some diseased patients will have a normal test (see Fig. 3.2). You should be familiar with the terms used to denote these groups:

- A **true positive** (TP) is a positive test result obtained for a patient in whom the disease is present (the test result correctly classifies the patient as having the disease).
- A **true negative** (TN) is a negative test result obtained for a patient in whom the disease is absent (the test result correctly classifies the patient as not having the disease).
- A **false positive** (FP) is a positive test result obtained for a patient in whom the disease is absent (the test result incorrectly classifies the patient as having the disease).
- A **false negative** (FN) is a negative test result obtained for a patient in whom the disease is present (the test result incorrectly classifies the patient as not having the disease).

Table 3.3 A 2×2 contingency table for test results

Results of test	Disease present	Disease absent	Total
Positive result	TP	FP	TP+FP
Negative result	FN	TN	FN+TN
	TP+FN	FP+TN	

TP true positive, *TN* true negative, *FP* false positive, *FN* false negative

Figure 3.2 shows that varying the cutoff point (moving the vertical line in the figure) for an abnormal test will change the relative proportions of these groups. As the cutoff is moved further up from the mean of the normal values, the number of FNs increases and the number of FPs decreases. Once we have chosen a cutoff point, we can conveniently summarize test performance—the ability to discriminate disease from nondisease—in a 2×2 **contingency table**, as shown in Table 3.3. The table summarizes the number of patients in each group: TP, FP, TN, and FN. Note that the sum of the first column is the total number of diseased patients, TP+FN. The sum of the second column is the total number of nondiseased patients, FP+TN. The sum of the first row, TP+FP, is the total number of patients with a positive test result. Likewise, FN+TN gives the total number of patients with a negative test result.

A perfect test would have no FN or FP results. Erroneous test results do occur, however, and you can use a 2×2 contingency table to define the measures of test performance that reflect these errors.

3.3.2 Measures of Test Performance

Measures of test performance are of two types: measures of agreement between tests or **measures of concordance**, and measures of disagreement or **measures of discordance**. Two types of **concordant test results** occur in the 2×2 table in Table 3.3: TPs and TNs. The relative frequencies of these results form the basis of the measures of concordance. These measures correspond to the ideas of the sensitivity and specificity of a test, which we introduced in Chap. 2. We define each measure in terms of the 2×2 table and in terms of conditional probabilities.

The **true-positive rate** (TPR), or **sensitivity**, is the likelihood that a diseased patient has a positive test. In conditional-probability notation, sensitivity is expressed as the probability of a positive test given that disease is present:

$$\mathrm{p}[\text{positive test} \mid \text{disease}].$$

Another way to think of the TPR is as a ratio. The likelihood that a diseased patient has a positive test is given by the ratio of diseased patients with a positive test to all diseased patients:

$$\mathrm{TPR} = \left(\frac{\text{number of diseased patients with positive test}}{\text{total number of diseased patients}} \right).$$

We can determine these numbers for our example from the 2×2 table (see Table 3.3). The number of diseased patients with a positive test is TP. The total number of diseased patients is the sum of the first column, TP+FN. So,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

The **true-negative** rate (TNR), or **specificity**, is the likelihood that a nondiseased patient has a negative test result. In terms of conditional probability, specificity is the probability of a negative test given that disease is absent:

$$\mathrm{p}[\text{negative test} \mid \text{no disease}].$$

Viewed as a ratio, the TNR is the number of nondiseased patients with a negative test divided by the total number of nondiseased patients:

$$\mathrm{TNR} = \left(\frac{\text{Number of nondiseased patients with negative test}}{\text{Total number of nondiseased patients}} \right).$$

From the 2×2 table (see Table 3.3),

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

The measures of discordance—the **false-positive rate** (FPR) and the **false-negative rate** (FNR)—are defined similarly. The FNR is the likelihood that a diseased patient has a negative test result. As a ratio,

$$\mathrm{FNR} = \left(\frac{\text{Number of diseased patients with negative test}}{\text{Total number of diseased patients}} \right) = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}.$$

The FPR is the likelihood that a nondiseased patient has a positive test result:

$$\mathrm{FPR} = \left(\frac{\text{Number of nondiseased patients with positive test}}{\text{Total number of nondiseased patients}} \right) = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}.$$

Example 6

Consider again the problem of screening blood donors for HIV. One test used to screen blood donors for HIV antibody is an enzyme-linked immunoassay (EIA). So that the performance of the EIA can be measured, the test is performed on 400 patients; the hypothetical results are shown in the 2×2 table in Table 3.4.[5]

[5]This example assumes that we have a perfect method (different from EIA) for determining the presence or absence of antibody. We discuss the idea of gold-standard tests in Sect. 3.3.4. We have chosen the numbers in the example to simplify the calculations. In practice, the sensitivity and specificity of the HIV EIAs are greater than 99 %.

Table 3.4 A 2×2 contingency table for HIV antibody EIA

EIA test result	Antibody present	Antibody absent	Total
Positive EIA	98	3	101
Negative EIA	2	297	299
	100	300	

EIA enzyme-linked immunoassay

To determine test performance, we calculate the TPR (sensitivity) and TNR (specificity) of the EIA antibody test. The TPR, as defined previously, is:

$$\frac{TP}{TP+FN} = \frac{98}{98+2} = 0.98.$$

Thus, the likelihood that a patient with the HIV antibody will have a positive EIA test is 0.98. If the test were performed on 100 patients who truly had the antibody, we would expect the test to be positive in 98 of the patients. Conversely, we would expect two of the patients to receive incorrect, negative results, for an FNR of 2 %. (You should convince yourself that the sum of TPR and FNR by definition must be 1: TPR + FNR = 1.)

And the TNR is:

$$\frac{TN}{TN+FP} = \frac{297}{297+3} = 0.99$$

The likelihood that a patient who has no HIV antibody will have a negative test is 0.99. Therefore, if the EIA test were performed on 100 individuals who had not been infected with HIV, it would be negative in 99 and incorrectly positive in 1. (Convince yourself that the sum of TNR and FPR also must be 1: TNR + FPR = 1.)

3.3.3 Implications of Sensitivity and Specificity: How to Choose Among Tests

It may be clear to you already that the calculated values of sensitivity and specificity for a continuous-valued test depend on the particular cutoff value chosen to distinguish normal and abnormal results. In Fig. 3.2, note that increasing the cutoff level (moving it to the right) would decrease significantly the number of FP tests but also would increase the number of FN tests. Thus, the test would have become more specific but less sensitive. Similarly, a lower cutoff value would increase the FPs and decrease the FNs, thereby increasing sensitivity while decreasing specificity. Whenever a decision is made about what cutoff to use in calling a test abnormal, an inherent philosophic decision is being made about whether it is better to tolerate FNs (missed cases) or FPs (nondiseased people inappropriately classified as diseased). The choice of cutoff depends on the disease in question and on the purpose of testing. If the disease is serious and if lifesaving therapy is available, we should try to minimize the number of FN results. On the other hand, if the disease in not serious and the therapy is dangerous, we should set the cutoff value to minimize FP results.

We stress the point that sensitivity and specificity are characteristics not of a test per se but rather of the test and a criterion for when to call that test abnormal. Varying the cutoff in Fig. 3.2 has no effect on the test itself (the way it is performed, or the specific values for any particular patient); instead, it trades off specificity for sensitivity. Thus, the best way to characterize a test is by the range of values of sensitivity and specificity that it can take on over a range of possible cutoffs. The typical way to show this relationship is to plot the test's sensitivity against 1 minus specificity (i.e., the TPR against the FPR), as the cutoff is varied and the two test characteristics are traded off against each other (Fig. 3.3). The resulting curve, known as a **receiver-operating characteristic** (**ROC**) **curve**, was originally described by researchers investigating methods of electromagnetic-signal detection and was later applied to the field of psychology (Peterson and Birdsall 1953; Swets 1973). Any given point along an ROC curve for a test corresponds to the test sensitivity and specificity for a given threshold of "abnormality." Similar curves can be drawn for any test used to associate observed clinical data with specific diseases or disease categories.

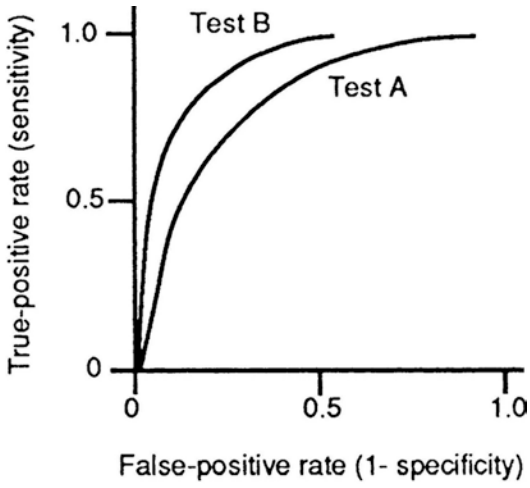

Fig. 3.3 Receiver operating characteristic (ROC) curves for two hypothetical tests. Test B is more discriminative than test A because its curve is higher (e.g., the false-positive rate (FPR) for test B is lower than the FPR for test A at any value of true-positive rate (TPR)). The more discriminative test may not always be preferred in clinical practice, however (see text)

Suppose a new test were introduced that competed with the current way of screening for the presence of a disease. For example, suppose a new radiologic procedure for assessing the presence or absence of pneumonia became available. This new test could be assessed for trade-offs in sensitivity and specificity, and an ROC curve could be drawn. As shown in Fig. 3.3, a test has better discriminating power than a competing test if its ROC curve lies above that of the other test. In other words, test B is more discriminating than test A when its specificity is greater than test A's specificity for any level of sensitivity (and when its sensitivity is greater than test A's sensitivity for any level of specificity).

Understanding ROC curves is important in understanding test selection and data interpretation. Clinicians should not necessarily, however, always choose the test with the most discriminating ROC curve. Matters of cost, risk, discomfort, and delay also are important in the choice about what data to collect and what tests to perform. When you must choose among several available tests, you should select the test that has the highest sensitivity and specificity, provided that other factors, such as cost and risk to the patient, are equal. The higher the sensitivity and specificity of a test, the more the results of that test will reduce uncertainty about probability of disease.

3.3.4 Design of Studies of Test Performance

In Sect. 3.3.2, we discussed measures of test performance: a test's ability to discriminate disease from no disease. When we classify a test result as TP, TN, FP, or FN, we assume that we know with certainty whether a patient is diseased or healthy. Thus, the validity of any test's results must be measured against a gold standard: a test that reveals the patient's true disease state, such as a biopsy of diseased tissue or a surgical operation. A **gold-standard test** is a procedure that is used to define unequivocally the presence or absence of disease. The test whose discrimination is being measured is called the **index test**. The gold-standard test usually is more expensive, riskier, or more difficult to perform than is the index test (otherwise, the less precise test would not be used at all).

The performance of the index test is measured in a small, select group of patients enrolled in a study. We are interested, however, in how the test performs in the broader group of patients in which it will be used in practice. The test may perform differently in the two groups, so we make the following distinction: the **study population** comprises those patients (usually a subset of the clinically relevant population) in whom test discrimination is measured and reported; the **clinically relevant population** comprises those patients in whom a test typically is used.

3.3.5 Bias in the Measurement of Test Characteristics

We mentioned earlier the problem of referral bias. Published estimates of disease prevalence (derived from a study population) may differ from the prevalence in the clinically relevant population because diseased patients are more likely to be included in studies than are nondiseased patients. Similarly, published values of

sensitivity and specificity are derived from study populations that may differ from the clinically relevant populations in terms of average level of health and disease prevalence. These differences may affect test performance, so the reported values may not apply to many patients in whom a test is used in clinical practice.

Example 7
In the early 1970s, a blood test called the carcinoembryonic antigen (CEA) was touted as a screening test for colon cancer. Reports of early investigations, performed in selected patients, indicated that the test had high sensitivity and specificity. Subsequent work, however, proved the CEA to be completely valueless as a screening blood test for colon cancer. Screening tests are used in unselected populations, and the differences between the study and clinically relevant populations were partly responsible for the original miscalculations of the CEA's TPR and TNR (Ransohoff and Feinstein 1978).

The experience with CEA has been repeated with numerous tests. Early measures of test discrimination are overly optimistic, and subsequent test performance is disappointing. Problems arise when the TPR and TNR, as measured in the study population, do not apply to the clinically relevant population. These problems usually are the result of bias in the design of the initial studies—notably spectrum bias, test referral bias, or test interpretation bias.

Spectrum bias occurs when the study population includes only individuals who have advanced disease ("sickest of the sick") and healthy volunteers, as is often the case when a test is first being developed. Advanced disease may be easier to detect than early disease. For example, cancer is easier to detect when it has spread throughout the body (metastasized) than when it is localized to, say, a small portion of the colon. In contrast to the study population, the clinically relevant population will contain more cases of early disease that are more likely to be missed by the index test (FNs). Thus, the study population will have an artifactually low FNR, which produces an artifactually high TPR (TPR = 1−FNR). In addition, healthy volunteers are less likely than are patients in the clinically relevant population to have other diseases that may cause FP results[6]; the study population will have an artificially low FPR, and therefore the specificity will be overestimated (TNR = 1−FPR). Inaccuracies in early estimates of the TPR and TNR of the CEA were partly due to spectrum bias.

Test-referral bias occurs when a positive index test is a criterion for ordering the gold standard test. In clinical practice, patients with negative index tests are less likely to undergo the gold standard test than are patients with positive tests. In other words, the study population, comprising individuals with positive index–test results, has a higher percentage of patients with disease than does the clinically relevant population. Therefore, both TN and FN tests will be underrepresented in the study population. The result is overestimation of the TPR and underestimation of the TNR in the study population.

Test-interpretation bias develops when the interpretation of the index test affects that of the gold standard test or vice versa. This bias causes an artificial concordance between the tests (the results are more likely to be the same) and spuriously increases measures of concordance—the sensitivity and specificity—in the study population. (Remember, the relative frequencies of TPs and TNs are the basis for measures of concordance). To avoid these problems, the person interpreting the index test should be unaware of the results of the gold standard test.

[6]Volunteers are often healthy, whereas patients in the clinically relevant population often have several diseases in addition to the disease for which a test is designed. These other diseases may cause FP test results. For example, patients with benign (rather than malignant) enlargement of their prostate glands are more likely than are healthy volunteers to have FP elevations of prostate-specific antigen (Meigs et al. 1996), a substance in the blood that is elevated in men who have prostate cancer. Measurement of prostate-specific antigen is often used to detect prostate cancer.

To counter these three biases, you may need to adjust the TPR and TNR when they are applied to a new population. All the biases result in a TPR that is higher in the study population than it is in the clinically relevant population. Thus, if you suspect bias, you should adjust the TPR (sensitivity) downward when you apply it to a new population.

Adjustment of the TNR (specificity) depends on which type of bias is present. Spectrum bias and test interpretation bias result in a TNR that is higher in the study population than it will be in the clinically relevant population. Thus, if these biases are present, you should adjust the specificity downward when you apply it to a new population. Test-referral bias, on the other hand, produces a measured specificity in the study population that is lower than it will be in the clinically relevant population. If you suspect test referral bias, you should adjust the specificity upward when you apply it to a new population.

3.3.6 Meta-Analysis of Diagnostic Tests

Often, there are many studies that evaluate the sensitivity and specificity of the same diagnostic test. If the studies come to similar conclusions about the sensitivity and specificity of the test, you can have increased confidence in the results of the studies. But what if the studies disagree? For example, by 1995, over 100 studies had assessed the sensitivity and specificity of the PCR for diagnosis of HIV (Owens et al. 1996a, b); these studies estimated the sensitivity of PCR to be as low as 10 % and to be as high as 100 %, and they assessed the specificity of PCR to be between 40 and 100 %. Which results should you believe? One approach that you can use is to assess the quality of the studies and to use the estimates from the highest-quality studies.

For evaluation of PCR, however, even the high-quality studies did not agree. Another approach is to perform a **meta-analysis**: a study that combines quantitatively the estimates from individual studies to develop a **summary ROC curve** (Moses et al. 1993; Owens et al. 1996a, b; Hellmich et al. 1999; Leeflang et al. 2008). Investigators develop a summary ROC curve by using estimates from many studies, in contrast to the type of ROC curve discussed in Sect. 3.3.3, which is developed from the data in a single study. Summary ROC curves provide the best available approach to synthesizing data from many studies.

Section 3.3 has dealt with the second step in the diagnostic process: acquisition of further information with diagnostic tests. We have learned how to characterize the performance of a test with sensitivity (TPR) and specificity (TNR). These measures reveal the probability of a test result given the true state of the patient. They do not, however, answer the clinically relevant question posed in the opening example: Given a positive test result, what is the probability that this patient has the disease? To answer this question, we must learn methods to calculate the post-test probability of disease.

3.4 Post-test Probability: Bayes' Theorem and Predictive Value

The third stage of the diagnostic process (see Fig. 3.1a) is to adjust our probability estimate to take into account the new information gained from diagnostic tests by calculating the post-test probability.

3.4.1 Bayes' Theorem

As we noted earlier in this chapter, a clinician can use the disease prevalence in the patient population as an initial estimate of the pretest risk of disease. Once clinicians begin to accumulate information about a patient, however, they revise their estimate of the probability of disease. The revised estimate (rather than the disease prevalence in the general population) becomes the pretest probability for the test that they perform. After they have gathered more information with a diagnostic test, they can calculate the post-test probability of disease with Bayes' theorem.

Bayes' theorem is a quantitative method for calculating post-test probability using the pretest probability and the sensitivity and specificity of the test. The theorem is derived from the definition of conditional probability and from the properties of probability (see the Appendix to this chapter for the derivation).

Recall that a conditional probability is the probability that event A will occur given that event B is known to occur (see Sect. 3.2). In general, we want to know the probability that disease is present (event A), given that the test is known to be positive (event B). We denote the presence of disease as D, its absence as −D, a test result as R, and the pretest probability of disease as p[D]. The probability of disease, given a test result, is written p[D|R]. Bayes' theorem is:

$$p[D \mid R] = \frac{p[D] \times p[R \mid D]}{p[D] \times p[R \mid D] + p[-D] \times p[R \mid -D]}$$

We can reformulate this general equation in terms of a positive test, (+), by substituting p[D|+] for p[D|R], p[+|D] for p[R|D], p[+|−D] for p[R|−D], and 1−p[D] for p[− D]. From Sect. 3.3, recall that p[+|D] = TPR and p[+|−D] = FPR. Substitution provides Bayes' theorem for a positive test:

$$p[D \mid +] = \frac{p[D] \times TPR}{p[D] \times TPR + (1 - p[D]) \times FPR}$$

We can use a similar derivation to develop Bayes' theorem for a negative test:

$$p[D \mid -] = \frac{p[D] \times FNR}{p[D] \times FNR + (1 - p[D]) \times TNR}$$

Example 8

We are now able to calculate the clinically important probability in Example 4: the post-test probability of heart disease after a positive exercise test. At the end of Sect. 3.2.2, we estimated the pretest probability of heart disease as 0.95, based on the prevalence of heart disease in men who have typical symptoms of heart disease and on the prevalence in people with a family history of heart disease. Assume that the TPR and FPR of the exercise stress test are 0.65 and 0.20, respectively. Substituting in Bayes' formula for a positive test, we obtain the probability of heart disease given a positive test result:

$$p[D \mid +] = \frac{0.95 \times 0.65}{0.95 \times 0.65 + 0.05 \times 0.20} = 0.98$$

Thus, the positive test raised the post-test probability to 0.98 from the pretest probability of 0.95. The change in probability is modest because the pretest probability was high (0.95) and because the FPR also is high (0.20). If we repeat the calculation with a pretest probability of 0.75, the post-test probability is 0.91. If we assume the FPR of the test to be 0.05 instead of 0.20, a pretest probability of 0.95 changes to 0.996.

3.4.2 The Odds-Ratio Form of Bayes' Theorem and Likelihood Ratios

Although the formula for Bayes' theorem is straightforward, it is awkward for mental calculations. We can develop a more convenient form of Bayes' theorem by expressing probability as odds and by using a different measure of test discrimination. Probability and odds are related as follows:

$$\text{odds} = \frac{p}{1-p},$$

$$p = \frac{\text{odds}}{1 + \text{odds}}.$$

Thus, if the probability of rain today is 0.75, the odds are 3:1. Thus, on similar days, we should expect rain to occur three times for each time it does not occur.

A simple relationship exists between pretest odds and post-test odds:

$$\text{post-test odds} = \text{pretest odds} \times \text{likelihood ratio}$$

or

$$\frac{p[\mathrm{D}\,|\,\mathrm{R}]}{p[-\mathrm{D}\,|\,\mathrm{R}]} = \frac{p[\mathrm{D}]}{p[-\mathrm{D}]} \times \frac{p[\mathrm{R}\,|\,\mathrm{D}]}{p[\mathrm{R}\,|\,-\mathrm{D}]}.$$

This equation is the **odds-ratio form** of Bayes' theorem.[7] It can be derived in a straightforward fashion from the definitions of Bayes' theorem and of conditional probability that we provided earlier. Thus, to obtain the post-test odds, we simply multiply the pretest odds by the **likelihood ratio** (LR) for the test in question.

The LR of a test combines the measures of test discrimination discussed earlier to give one number that characterizes the discriminatory power of a test, defined as:

$$\mathrm{LR} = \frac{p[\mathrm{R}|\mathrm{D}]}{p[\mathrm{R}\,|\,-\mathrm{D}]}$$

or

$$\mathrm{LR} = \frac{\text{probability of result in diseased people}}{\text{probability of result in nondiseased people}}$$

The LR indicates the amount that the odds of disease change based on the test result. We can use the LR to characterize clinical findings (such as a swollen leg) or a test result. We describe the performance of a test that has only two possible outcomes (e.g., positive or negative) by two LRs: one corresponding to a positive test result and the other corresponding to a negative test. These ratios are abbreviated LR+ and LR−, respectively.

$$\mathrm{LR}+ = \left(\frac{\begin{array}{c}\text{probability that test}\\ \text{is positive in diseased people}\end{array}}{\begin{array}{c}\text{probability that test}\\ \text{is positive in nondiseased people}\end{array}} \right) = \frac{\mathrm{TPR}}{\mathrm{FPR}}$$

[7]Some authors refer to this expression as the odds-likelihood form of Bayes' theorem.

In a test that discriminates well between disease and nondisease, the TPR will be high, the FPR will be low, and thus LR+ will be much greater than 1. An LR of 1 means that the probability of a test result is the same in diseased and nondiseased individuals; the test has no value. Similarly,

$$\mathrm{LR}- = \left(\frac{\begin{array}{c}\text{probability that test}\\ \text{is negative in diseased people}\end{array}}{\begin{array}{c}\text{probability that test}\\ \text{is negative in nondiseased people}\end{array}} \right) = \frac{\mathrm{FNR}}{\mathrm{TNR}}$$

A desirable test will have a low FNR and a high TNR; therefore, the LR− will be much less than 1.

Example 9

We can calculate the post-test probability for a positive exercise stress test in a 60 year-old man whose pretest probability is 0.75. The pretest odds are:

$$\text{odds} = \frac{p}{1-p} = \frac{0.75}{1-0.75} = \frac{0.75}{0.25} = 3, \text{ or } 3:1$$

The LR for the stress test is:

$$\mathrm{LR}+ = \frac{\mathrm{TPR}}{\mathrm{FPR}} = \frac{0.65}{0.20} = 3.25$$

We can calculate the post-test odds of a positive test result using the odds-ratio form of Bayes' theorem:

$$\text{post-test odds} = 3 \times 3.25 = 9.75:1$$

We can then convert the odds to a probability:

$$\mathrm{p} = \frac{\text{odds}}{1+\text{odds}} = \frac{9.75}{1+9.75} = 0.91$$

As expected, this result agrees with our earlier answer (see the discussion of Example 8).

The odds-ratio form of Bayes' theorem allows rapid calculation, so you can determine the probability at, for example, your patient's bedside.

The LR is a powerful method for characterizing the operating characteristics of a test: if you know the pretest odds, you can calculate the post-test odds in one step. The LR demonstrates that a useful test is one that changes the odds of disease.

3.4.3 Predictive Value of a Test

An alternative approach for estimation of the probability of disease in a person who has a positive or negative test is to calculate the predictive value of the test. The **positive predictive value** (PV+) of a test is the likelihood that a patient who has a positive test result also has disease. Thus, PV+ can be calculated directly from a 2×2 contingency table:

$$\text{PV+} = \frac{\text{number of diseased patients with positive test}}{\text{total number of patients with a positive test}}$$

From the 2×2 contingency table in Table 3.3,

$$\text{PV+} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The **negative predictive value** (PV−) is the likelihood that a patient with a negative test does not have disease:

$$\text{PV−} = \frac{\text{number of nondiseased patients with negative test}}{\text{Total number of patients with a negative test}}$$

From the 2×2 contingency table in Table 3.3,

$$\text{PV−} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

Example 10

We can calculate the PV of the EIA test from the 2×2 table that we constructed in Example 6 (see Table 3.4) as follows:

$$\text{PV+} = \frac{98}{98+3} = 0.97$$

$$\text{PV−} = \frac{297}{297+2} = 0.99$$

The probability that antibody is present in a patient who has a positive index test (EIA) in this study is 0.97; about 97 of 100 patients with a positive test will have antibody. The likelihood that a patient with a negative index test does not have antibody is about 0.99.

It is worth reemphasizing the difference between PV and sensitivity and specificity, given that both are calculated from the 2×2 table and they often are confused. The sensitivity and specificity give the probability of a particular test result in a patient who has a particular disease state. The PV gives the probability of true disease state once the patient's test result is known.

The PV+ calculated from Table 3.4 is 0.97, so we expect 97 of 100 patients with a positive index test actually to have antibody. Yet, in Example 1, we found that fewer than one of ten patients with a positive test were expected to have antibody. What explains the discrepancy in these examples? The sensitivity and specificity (and, therefore, the LRs) in the two examples are identical. The discrepancy is due to an extremely important and often overlooked characteristic of PV: the PV of a test depends on the prevalence of disease in the study population (the prevalence can be calculated as TP+FN divided by the total number of patients in the 2×2 table). The PV cannot be generalized to a new population because the prevalence of disease may differ between the two populations.

The difference in PV of the EIA in Example 1 and in Example 6 is due to a difference in the prevalence of disease in the examples. The prevalence of antibody was given as 0.001 in Example 1 and as 0.25 in Example 6. These examples should remind us that the PV+ is not an intrinsic property of a test. Rather, it represents the post-test probability of disease only when the prevalence is identical to that in the 2×2 contingency table from which the PV+ was calculated. Bayes' theorem provides a method for calculation of the post-test probability of disease for any prior

probability. For that reason, we prefer the use of Bayes' theorem to calculate the post-test probability of disease.

3.4.4 Implications of Bayes' Theorem

In this section, we explore the implications of Bayes' theorem for test interpretation. These ideas are extremely important, yet they often are misunderstood.

Figure 3.4 illustrates one of the most essential concepts in this chapter: The post-test probability of disease increases as the pretest probability of disease increases. We produced Fig. 3.4a by calculating the post-test probability after a positive test result for all possible pretest probabilities of disease. We similarly derived Fig. 3.4b for a negative test result.

The 45-degree line in each figure denotes a test in which the pretest and post-test probability are equal (LR = 1), indicating a test that is useless. The curve in Fig. 3.4a relates pretest and post-test probabilities in a test with a sensitivity and specificity of 0.9. Note that, at low pretest probabilities, the post-test probability after a positive test result is much higher than is the pretest probability. At high pretest probabilities, the post-test probability is only slightly higher than the pretest probability.

Figure 3.4b shows the relationship between the pretest and post-test probabilities after a negative test result. At high pretest probabilities, the post-test probability after a negative test result is much lower than is the pretest probability. A negative test, however, has little effect on the post-test probability if the pretest probability is low.

This discussion emphasizes a key idea of this chapter: the interpretation of a test result depends on the pretest probability of disease. If the pretest probability is low, a positive test result has a large effect, and a negative test result has a small effect. If the pretest probability is high, a positive test result has a small effect, and a negative test result has a large effect. In other words, when the clinician is almost certain of the diagnosis before testing (pretest probability nearly 0 or nearly 1), a confirmatory test has little effect on the posterior probability (see Example 8). If the pretest probability is intermediate or if the result contradicts a strongly held clinical impression, the test result will have a large effect on the post-test probability.

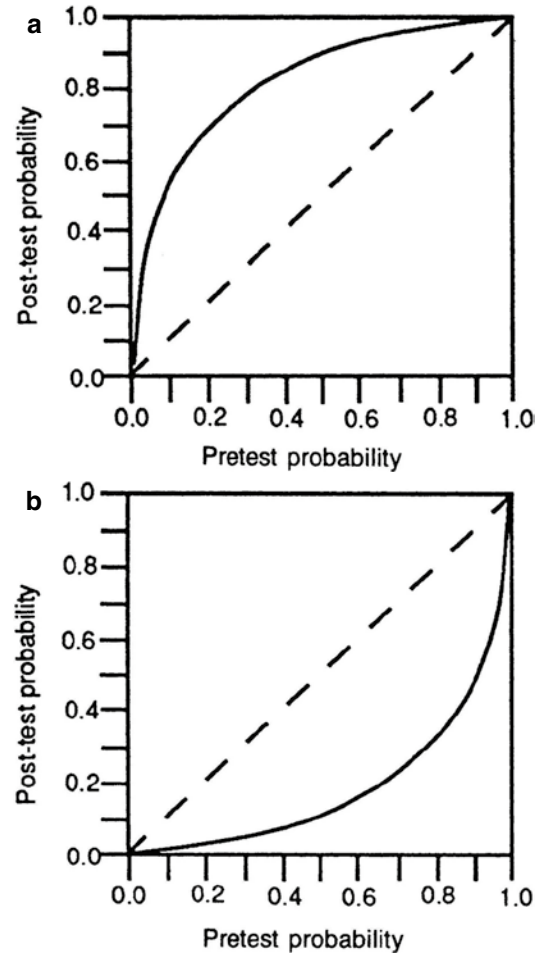

Fig. 3.4 Relationship between pretest probability and post-test probability of disease. The *dashed lines* correspond to a test that has no effect on the probability of disease. Sensitivity and specificity of the test were assumed to be 0.90 for the two examples. (**a**) The post-test probability of disease corresponding to a positive test result (*solid curve*) was calculated with Bayes' theorem for all values of pretest probability. (**b**) The post-test probability of disease corresponding to a negative test result (*solid curve*) was calculated with Bayes' theorem for all values of pretest probability (Source: Adapted from Sox, H.C. (1987). Probability theory in the use of diagnostic tests: Application to critical study of the literature. In: Sox H.C. (Ed.), *Common diagnostic tests: Use and interpretation* (pp. 1–17). American College of Physicians, with permission)

Note from Fig. 3.4a that, if the pretest probability is very low, a positive test result can raise the post-test probability into only the intermediate

range. Assume that Fig. 3.4a represents the relationship between the pretest and post-test probabilities for the exercise stress test. If the clinician believes the pretest probability of coronary artery disease is 0.1, the post-test probability will be about 0.5. Although there has been a large change in the probability, the post-test probability is in an intermediate range, which leaves considerable uncertainty about the diagnosis. Thus, if the pretest probability is low, it is unlikely that a positive test result will raise the probability of disease sufficiently for the clinician to make that diagnosis with confidence. An exception to this statement occurs when a test has a very high specificity (or a large LR+); e.g., HIV antibody tests have a specificity greater than 0.99, and therefore a positive test is convincing. Similarly, if the pretest probability is very high, it is unlikely that a negative test result will lower the post-test probability sufficiently to exclude a diagnosis.

Figure 3.5 illustrates another important concept: test specificity affects primarily the interpretation of a positive test; test sensitivity affects primarily the interpretation of a negative test. In both parts (a) and (b) of Fig. 3.5, the top family of curves corresponds to positive test results and the bottom family to negative test results. Figure 3.5a shows the post-test probabilities for tests with varying specificities (TNR). Note that changes in the specificity produce large changes in the top family of curves (positive test results) but have little effect on the lower family of curves (negative test results). That is, an increase in the specificity of a test markedly changes the post-test probability if the test is positive but has relatively little effect on the post-test probability if the test is negative. Thus, if you are trying to rule in a diagnosis,[8] you should choose a test with high specificity or a high LR+. Figure 3.5b shows the post-test probabilities for tests with varying sensitivities. Note that changes in sensitivity produce

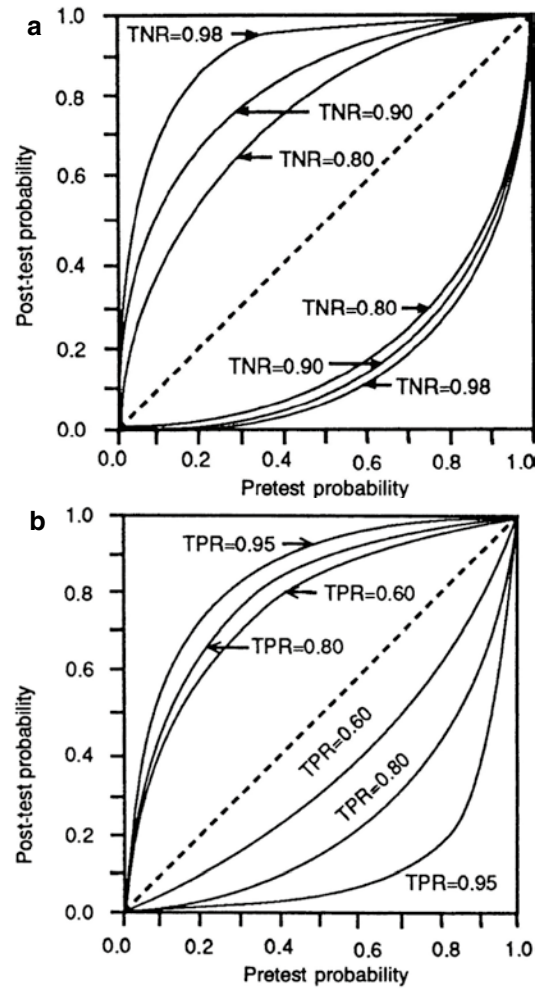

Fig. 3.5 Effects of test sensitivity and specificity on post-test probability. The *curves* are similar to those shown in Fig. 3.4 except that the calculations have been repeated for several values of the sensitivity (*TPR* true-positive rate) and specificity (*TNR* true-negative rate) of the test. (**a**) The sensitivity of the test was assumed to be 0.90, and the calculations were repeated for several values of test specificity. (**b**) The specificity of the test was assumed to be 0.90, and the calculations were repeated for several values of the sensitivity of the test. In both panels, the top family of curves corresponds to positive test results, and the bottom family of curves corresponds to negative test results (Source: Adapted from Sox (1987). Probability theory in the use of diagnostic tests: Application to critical study of the literature. In: Sox (Ed.), *Common diagnostic tests: Use and interpretation* (pp. 1–17), American College of Physicians, with permission)

[8]In medicine, to *rule in* a disease is to confirm that the patient *does* have the disease; to *rule out* a disease is to confirm that the patient *does not* have the disease. A doctor who strongly suspects that his or her patient has a bacterial infection orders a culture to *rule in* his or her diagnosis. Another doctor is almost certain that his or her patient has a simple sore throat but orders a culture to rule out streptococcal infection (strep throat). This terminology oversimplifies a diagnostic process that is probabilistic. Diagnostic tests rarely, if ever, rule in or rule out a disease; rather, the tests raise or lower the probability of disease.

large changes in the bottom family of curves (negative test results) but have little effect on the top family of curves. Thus, if you are trying to exclude a disease, choose a test with a high sensitivity or a high LR−.

3.4.5 Cautions in the Application of Bayes' Theorem

Bayes' theorem provides a powerful method for calculating post-test probability. You should be aware, however, of the possible errors you can make when you use it. Common problems are inaccurate estimation of pretest probability, faulty application of test-performance measures, and violation of the assumptions of conditional independence and of mutual exclusivity.

Bayes' theorem provides a means to adjust an estimate of pretest probability to take into account new information. The accuracy of the calculated post-test probability is limited, however, by the accuracy of the estimated pretest probability. Accuracy of estimated prior probability is increased by proper use of published prevalence rates, heuristics, and clinical prediction rules. In a decision analysis, as we shall see, a range of prior probability often is sufficient. Nonetheless, if the pretest probability assessment is unreliable, Bayes' theorem will be of little value.

A second potential mistake that you can make when using Bayes' theorem is to apply published values for the test sensitivity and specificity, or LRs, without paying attention to the possible effects of bias in the studies in which the test performance was measured (see Sect. 3.3.5). With certain tests, the LRs may differ depending on the pretest odds in part because differences in pretest odds may reflect differences in the spectrum of disease in the population.

A third potential problem arises when you use Bayes' theorem to interpret a sequence of tests. If a patient undergoes two tests in sequence, you can use the post-test probability after the first test result, calculated with Bayes' theorem, as the pretest probability for the second test. Then, you use Bayes' theorem a second time to calculate the post-test probability after the second test. This approach is valid, however, only if the two tests are conditionally independent. Tests for the same disease are **conditionally independent** when the probability of a particular result on the second test does not depend on the result of the first test, given (conditioned on) the disease state. Expressed in conditional probability notation for the case in which the disease is present,

$$
\begin{aligned}
&p\left[\begin{array}{l}\text{second test positive} \mid \text{first test positive}\\ \text{and disease present}\end{array}\right]\\
&= p\left[\begin{array}{l}\text{second test positive} \mid \text{first test negative}\\ \text{and disease present}\end{array}\right]\\
&= p[\text{second test positive} \mid \text{disease present}].
\end{aligned}
$$

If the conditional independence assumption is satisfied, the post-test odds = pretest odds × LR1 × LR2. If you apply Bayes' theorem sequentially in situations in which conditional independence is violated, you will obtain inaccurate post-test probabilities (Gould 2003).

The fourth common problem arises when you assume that all test abnormalities result from one (and only one) disease process. The Bayesian approach, as we have described it, generally presumes that the diseases under consideration are **mutually exclusive**. If they are not, Bayesian updating must be applied with great care.

We have shown how to calculate post-test probability. In Sect. 3.5, we turn to the problem of decision making when the outcomes of a clinician's actions (e.g., of treatments) are uncertain.

3.5 Expected-Value Decision Making

Medical decision-making problems often cannot be solved by reasoning based on pathophysiology. For example, clinicians need a method for choosing among treatments when the outcome of the treatments is uncertain, as are the results of a surgical operation. You can use the ideas developed in the preceding sections to solve such difficult decision problems. Here we discuss two methods: the decision tree, a method for representing and comparing the expected outcomes of each decision alternative; and the threshold probability, a method for deciding whether new information can change a management decision.

These techniques help you to clarify the decision problem and thus to choose the alternative that is most likely to help the patient.

3.5.1 Comparison of Uncertain Prospects

Like those of most biological events, the outcome of an individual's illness is unpredictable. How can a clinician determine which course of action has the greatest chance of success?

Which of the two therapies is preferable? Example 11 demonstrates a significant fact: a choice among therapies is a choice among gambles (i.e., situations in which chance determines the outcomes). How do we usually choose among gambles? More often than not, we rely on hunches or on a sixth sense. How should we choose among gambles? We propose a method

Example 11

There are two available therapies for a fatal illness. The length of a patient's life after either therapy is unpredictable, as illustrated by the frequency distribution shown in Fig. 3.6 and summarized in Table 3.5. Each therapy is associated with uncertainty: regardless of which therapy a patient receives, he will die by the end of the fourth year, but there is no way to know which year will be the patient's last. Figure 3.6 shows that survival until the fourth year is more likely with therapy B, but the patient might die in the first year with therapy B or might survive to the fourth year with therapy A.

Table 3.5 Distribution of probabilities for the two therapies in Fig. 3.7

	Probability of death	
Years after therapy	Therapy A	Therapy B
1	0.20	0.05
2	0.40	0.15
3	0.30	0.45
4	0.10	0.35

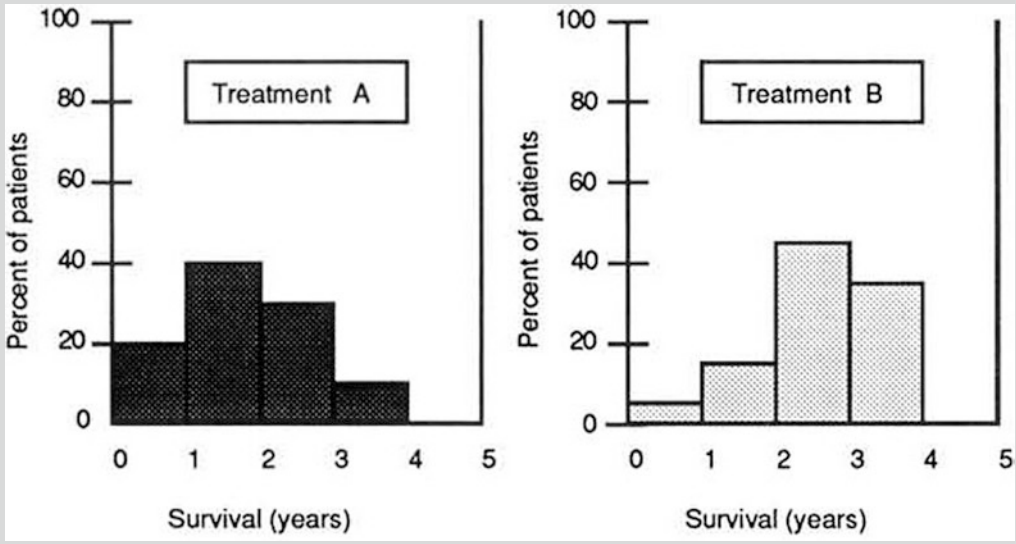

Fig. 3.6 Survival after therapy for a fatal disease. Two therapies are available; the results of either are unpredictable

for choosing called expected-value decision making: we characterize each gamble by a number, and we use that number to compare the gambles.[9] In Example 11, therapy A and therapy B are both gambles with respect to duration of life after therapy. We want to assign a measure (or number) to each therapy that summarizes the outcomes such that we can decide which therapy is preferable.

The ideal criterion for choosing a gamble should be a number that reflects preferences (in medicine, often the patient's preferences) for the outcomes of the gamble. **Utility** is the name given to a measure of preference that has a desirable property for decision making: the gamble with the highest utility should be preferred. We shall discuss utility briefly (Sect. 3.5.4), but you can pursue this topic and the details of decision analysis in other textbooks (see Suggested Readings at the end of this chapter).[10] We use the average duration of life after therapy (survival) as a criterion for choosing among therapies; remember that this model is oversimplified, used here for discussion only. Later, we consider other factors, such as the quality of life.

Because we cannot be sure of the duration of survival for any given patient, we characterize a therapy by the mean survival (average length of life) that would be observed in a large number of patients after they were given the therapy. The first step we take in calculating the mean survival for a therapy is to divide the population receiving the therapy into groups of patients who have similar survival rates. Then, we multiply the survival time in each group[11] by the fraction of the total population in that group. Finally, we sum these products over all possible survival values.

We can perform this calculation for the therapies in Example 11. Mean survival for therapy $A = (0.2 \times 1.0) + (0.4 \times 2.0) + (0.3 \times 3.0) + (0.1 \times 4.0) = 2.3$ years. Mean survival for therapy $B = (0.05 \times 1.0) + (0.15 \times 2.0) + (0.45 \times 3.0) + (0.35 \times 4.0) = 3.1$ years.

Survival after a therapy is under the control of chance. Therapy A is a gamble characterized by an average survival equal to 2.3 years. Therapy B is a gamble characterized by an average survival of 3.1 years. If length of life is our criterion for choosing, we should select therapy B.

3.5.2 Representation of Choices with Decision Trees

The choice between therapies A and B is represented diagrammatically in Fig. 3.7. Events that are under the control of chance can be represented by a **chance node**. By convention, a chance node is shown as a circle from which several lines emanate. Each line represents one of the possible outcomes. Associated with each line is the probability of the outcome occurring. For a single patient, only one outcome can occur. Some physicians object to using probability for just this reason: "You cannot rely on population data, because each patient is an individual." In fact, we often must use the frequency of the outcomes of many patients experiencing the same event to inform our opinion about what might happen to an individual. From these frequencies, we can make patient-specific adjustments and thus estimate the probability of each outcome at a chance node.

A chance node can represent more than just an event governed by chance. The outcome of a chance event, unknowable for the individual, can be represented by the **expected value** at the chance node. The concept of expected value is important and is easy to understand. We can calculate the mean survival that would be expected based on the probabilities depicted by the chance node in Fig. 3.7. This average length of life is called the expected survival or, more generally, the expected value of the chance node. We calculate the expected value at a chance node by the process just described: we multiply the survival value associated with each possible outcome by

[9] Expected-value decision making had been used in many fields before it was first applied to medicine.

[10] A more general term for expected-value decision making is expected utility decision making. Because a full treatment of utility is beyond the scope of this chapter, we have chosen to use the term expected value.

[11] For this simple example, death during an interval is assumed to occur at the end of the year.

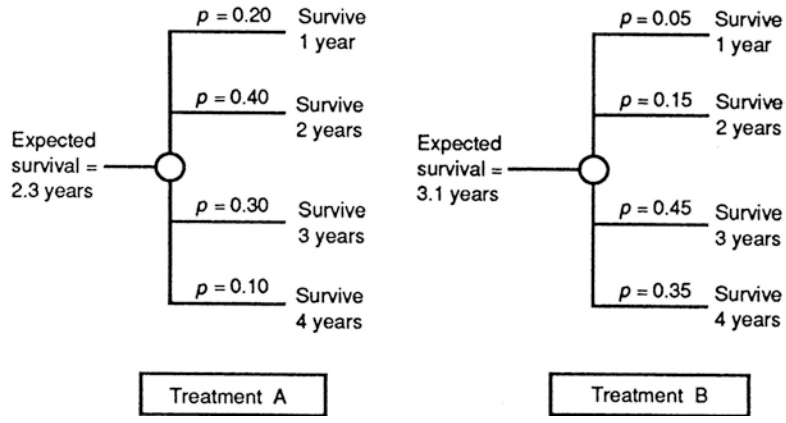

Fig. 3.7 A chance-node representation of survival after the two therapies in Fig. 3.6. The probabilities times the corresponding years of survival are summed to obtain the total expected survival

the probability that that outcome will occur. We then sum the product of probability times survival over all outcomes. Thus, if several hundred patients were assigned to receive either therapy A or therapy B, the expected survival would be 2.3 years for therapy A and 3.1 years for therapy B.

We have just described the basis of expected-value decision making. The term expected value is used to characterize a chance event, such as the outcome of a therapy. If the outcomes of a therapy are measured in units of duration of survival, units of sense of well-being, or dollars, the therapy is characterized by the expected duration of survival, expected sense of well-being, or expected monetary cost that it will confer on, or incur for, the patient, respectively.

To use expected-value decision making, we follow this strategy when there are therapy choices with uncertain outcomes: (1) calculate the expected value of each decision alternative and then (2) pick the alternative with the highest expected value.

3.5.3 Performance of a Decision Analysis

We clarify the concepts of expected-value decision making by discussing an example. There are four steps in decision analysis:

1. Create a decision tree; this step is the most difficult, because it requires formulating the decision problem, assigning probabilities, and measuring outcomes.
2. Calculate the expected value of each decision alternative.
3. Choose the decision alternative with the highest expected value.
4. Use sensitivity analysis to test the conclusions of the analysis.

Many health professionals hesitate when they first learn about the technique of decision analysis, because they recognize the opportunity for error in assigning values to both the probabilities and the utilities in a decision tree. They reason that the technique encourages decision making based on small differences in expected values that are estimates at best. The defense against this concern, which also has been recognized by decision analysts, is the technique known as sensitivity analysis. We discuss this important fourth step in decision analysis in Sect. 3.5.5.

The first step in decision analysis is to create a **decision tree** that represents the decision problem. Consider the following clinical problem.

Example 12

The patient is Mr. Danby, a 66-year-old man who has been crippled with arthritis of both knees so severely that, while he can get about the house with the aid of two canes, he must otherwise use a wheelchair. His other major health problem is emphysema, a disease in which the lungs lose their ability to exchange oxygen and carbon dioxide between blood and air, which in turn causes shortness of breath (dyspnea). He is able to breathe comfortably

when he is in a wheelchair, but the effort of walking with canes makes him breathe heavily and feel uncomfortable. Several years ago, he seriously considered knee replacement surgery but decided against it, largely because his internist told him that there was a serious risk that he would not survive the operation because of his lung disease. Recently, however, Mr. Danby's wife had a stroke and was partially paralyzed; she now requires a degree of assistance that the patient cannot supply given his present state of mobility. He tells his doctor that he is reconsidering knee replacement surgery.

Mr. Danby's internist is familiar with decision analysis. She recognizes that this problem is filled with uncertainty: Mr. Danby's ability to survive the operation is in doubt, and the surgery sometimes does not restore mobility to the degree required by such a patient. Furthermore, there is a small chance that the prosthesis (the artificial knee) will become infected, and Mr. Danby then would have to undergo a second risky operation to remove it. After removal of the prosthesis, Mr. Danby would never again be able to walk, even with canes. The possible outcomes of knee replacement include death from the first procedure and death from a second mandatory procedure if the prosthesis becomes infected (which we will assume occurs in the immediate postoperative period, if it occurs at all). Possible functional outcomes include recovery of full mobility or continued, and unchanged, poor mobility. Should Mr. Danby choose to undergo knee replacement surgery, or should he accept the status quo?

Using the conventions of decision analysis, the internist sketches the decision tree shown in Fig. 3.8. According to these conventions, a square box denotes a **decision node**, and each line emanating from a decision node represents an action that could be taken.

According to the methods of expected-value decision making, the internist first must assign a probability to each branch of each chance node. To accomplish this task, the internist asks several orthopedic surgeons for their estimates of the chance of recovering full function after surgery (p[full recovery] = 0.60) and the chance of developing infection in the prosthetic joint (p[infection] = 0.05). She uses her subjective estimate of the probability that the patient will die during or immediately after knee surgery (p[operative death] = 0.05).

Next, she must assign a value to each outcome. To accomplish this task, she first lists the outcomes. As you can see from Table 3.6, the outcomes differ in two dimensions: length of life (survival) and quality of life (functional status). To characterize each outcome accurately, the internist must develop a measure that takes into account these two dimensions. Simply using duration of survival is inadequate because Mr. Danby values 5 years of good health more than he values 10 years of poor health. The internist can account for this trade-off factor by converting outcomes with two dimensions into outcomes with a single dimension: duration of survival in good health. The resulting measure is called a **quality-adjusted life year** (QALY).[12]

She can convert years in poor health into years in good health by asking Mr. Danby to indicate the shortest period in good health (full mobility) that he would accept in return for his full expected lifetime (10 years) in a state of poor health (status quo). Thus, she asks Mr. Danby: "Many people say they would be willing to accept a shorter life in excellent health in preference to a longer life with significant disability. In your case, how many years with normal mobility do you feel is equivalent in value to 10 years in your current state of disability?" She asks him this question for each outcome. The patient's responses are shown in the third column of Table 3.6. The patient decides that 10 years of limited mobility are equivalent to 6 years of normal mobility,

[12]QALYs commonly are used as measures of utility (value) in medical decision analysis and in health policy analysis.

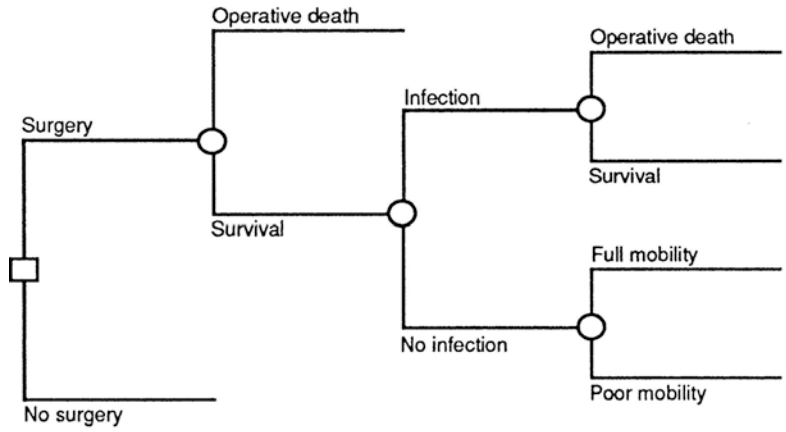

Fig. 3.8 Decision tree for knee replacement surgery. The *box* represents the decision node (whether to have surgery); the *circles* represent chance nodes

Table 3.6 Outcomes for Example 12

Survival (years)	Functional status	Years of full function equivalent to outcome
10	Full mobility (successful surgery)	10
10	Poor mobility (status quo or unsuccessful surgery)	6
10	Wheelchair-bound (the outcome if a second surgery is necessary)	3
0	Death	0

whereas 10 years of wheelchair confinement are equivalent to only 3 years of full function. Figure 3.9 shows the final decision tree—complete with probability estimates and utility values for each outcome.[13]

The second task that the internist must undertake is to calculate the expected value, in healthy years, of surgery and of no surgery. She calculates the expected value at each chance node, moving from right (the tips of the tree) to left (the root of the tree). Let us consider, for example, the expected value at the chance node representing the outcome of surgery to remove an infected prosthesis (Node A in Fig. 3.9). The calculation requires three steps:

1. Calculate the expected value of operative death after surgery to remove an infected prosthesis. Multiply the probability of operative death (0.05) by the QALY of the outcome—death (0 years): $0.05 \times 0 = 0$ QALY.
2. Calculate the expected value of surviving surgery to remove an infected knee prosthesis. Multiply the probability of surviving the operation (0.95) by the number of healthy years equivalent to 10 years of being wheelchair-bound (3 years): $0.95 \times 3 = 2.85$ QALYs.
3. Add the expected values calculated in step 1 (0 QALY) and step 2 (2.85 QALYs) to obtain the expected value of developing an infected prosthesis: $0 + 2.85 = 2.85$ QALYs.

Similarly, the expected value at chance node B is calculated: $(0.6 \times 10) + (0.4 \times 6) = 8.4$ QALYs. To obtain the expected value of surviving knee replacement surgery (Node C), she proceeds as follows:

1. Multiply the expected value of an infected prosthesis (already calculated as 2.85 QALYs) by the probability that the prosthesis will become infected (0.05): $2.85 \times 0.05 = 0.143$ QALYs.
2. Multiply the expected value of never developing an infected prosthesis (already calculated as 8.4 QALYs) by the probability that the prosthesis will not become infected (0.95): $8.4 \times 0.95 = 7.98$ QALYs.
3. Add the expected values calculated in step 1 (0.143 QALY) and step 2 (7.98 QALYs) to get the expected value of surviving knee replacement surgery: $0.143 + 7.98 = 8.123$ QALYs.

[13] In a more sophisticated decision analysis, the clinician also would adjust the utility values of outcomes that require surgery to account for the pain and inconvenience associated with surgery and rehabilitation.

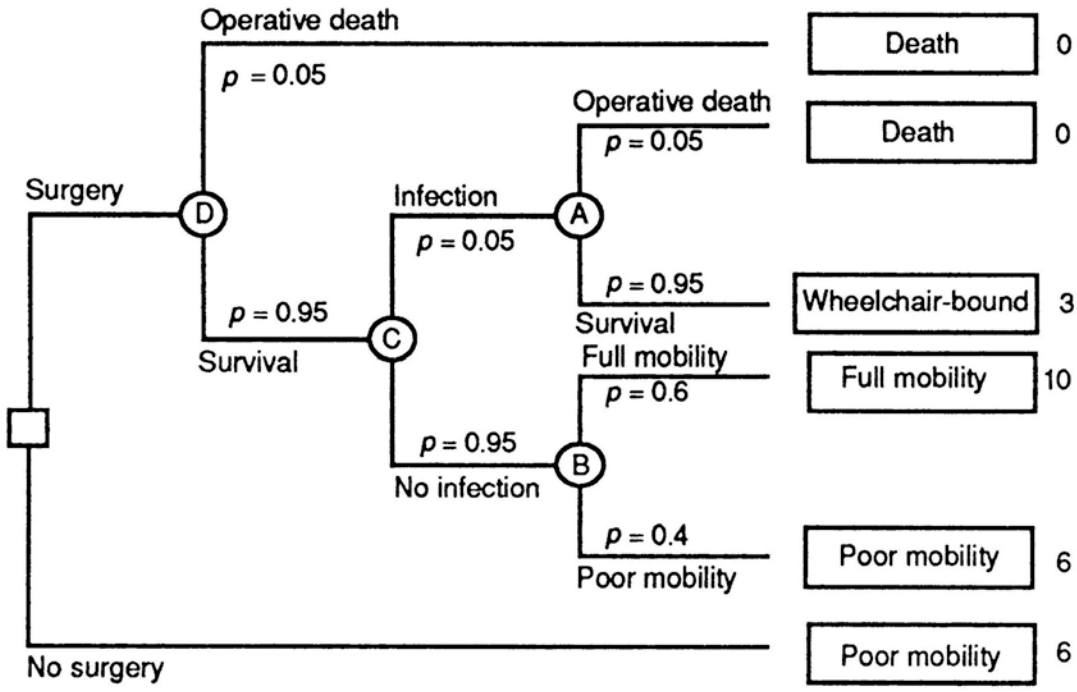

Fig. 3.9 Decision tree for knee-replacement surgery. Probabilities have been assigned to each branch of each chance node. The patient's valuations of outcomes (measured in years of perfect mobility) are assigned to the tips of each branch of the tree

The clinician performs this process, called **averaging out at chance nodes**, for node D as well, working back to the root of the tree, until the expected value of surgery has been calculated. The outcome of the analysis is as follows. For surgery, Mr. Danby's average life expectancy, measured in years of normal mobility, is 7.7. What does this value mean? It does not mean that, by accepting surgery, Mr. Danby is guaranteed 7.7 years of mobile life. One look at the decision tree will show that some patients die in surgery, some develop infection, and some do not gain any improvement in mobility after surgery. Thus, an individual patient has no guarantees. If the clinician had 100 similar patients who underwent the surgery, however, the average number of mobile years would be 7.7. We can understand what this value means for Mr. Danby only by examining the alternative: no surgery.

In the analysis for no surgery, the average length of life, measured in years of normal mobility, is 6.0, which Mr. Danby considered equivalent to 10 years of continued poor mobility. Not all patients will experience this outcome; some who have poor mobility will live longer than, and some will live less than, 10 years. The average length of life, however, expressed in years of normal mobility, will be 6. Because 6.0 is less than 7.7, on average the surgery will provide an outcome with higher value to the patient. Thus, the internist recommends performing the surgery.

The key insight of expected-value decision making should be clear from this example: given the unpredictable outcome in an individual, the best choice for the individual is the alternative that gives the best result on the average in similar patients. Decision analysis can help the clinician to identify the therapy that will give the best results when averaged over many similar patients. The decision analysis is tailored to a specific patient in that both the utility functions and the probability estimates are adjusted to the individual. Nonetheless, the results of the analysis represent the outcomes that would occur on average in a population of patients who have similar utilities and for whom uncertain events have similar probabilities.

3.5.4 Representation of Patients' Preferences with Utilities

In Sect. 3.5.3, we introduced the concept of QALYs, because length of life is not the only outcome about which patients care. Patients' preferences for a health outcome may depend on the length of life with the outcome, on the quality of life with the outcome, and on the risk involved in achieving the outcome (e.g., a cure for cancer might require a risky surgical operation). How can we incorporate these elements into a decision analysis? To do so, we can represent patients' preferences with utilities. The utility of a health state is a quantitative measure of the desirability of a health state from the patient's perspective. Utilities are typically expressed on a 0 to 1 scale, where 0 represents death and 1 represents ideal health. For example, a study of patients who had chest pain (angina) with exercise rated the utility of mild, moderate, and severe angina as 0.95, 0.92, and 0.82 (Nease et al. 1995), respectively. There are several methods for assessing utilities.

The **standard-gamble** technique has the strongest theoretical basis of the various approaches to utility assessment, as shown by Von Neumann and Morgenstern and described by Sox et al. (1988). To illustrate use of the standard gamble, suppose we seek to assess a person's utility for the health state of asymptomatic HIV infection. To use the standard gamble, we ask our subject to compare the desirability of asymptomatic HIV infection to those of two other health states whose utility we know or can assign. Often, we use ideal health (assigned a utility of 1) and immediate death (assigned a utility of 0) for the comparison of health states. We then ask our subject to choose between asymptomatic HIV infection and a gamble with a chance of ideal health or immediate death. We vary the probability of ideal health and immediate death systematically until the subject is indifferent between asymptomatic HIV infection and the gamble. For example, a subject might be indifferent when the probability of ideal health is 0.8 and the probability of death is 0.2. At this point of indifference, the utility of the gamble and that of asymptomatic HIV infection are equal. We calculate the utility of the gamble as the weighted average of the utilities of each outcome of the gamble $[(1 \times 0.8) + (0 \times 0.2)] = 0.8$. Thus in this example, the utility of asymptomatic HIV infection is 0.8. Use of the standard gamble enables an analyst to assess the utility of outcomes that differ in length or quality of life. Because the standard gamble involves chance events, it also assesses a person's willingness to take risks—called the person's **risk attitude**.

A second common approach to utility assessment is the **time-trade-off** technique (Sox et al. 1988; Torrance and Feeny 1989). To assess the utility of asymptomatic HIV infection using the time-trade-off technique, we ask a person to determine the length of time in a better state of health (usually ideal health or best attainable health) that he or she would find equivalent to a longer period of time with asymptomatic HIV infection. For example, if our subject says that 8 months of life with ideal health was equivalent to 12 months of life with asymptomatic HIV infection, then we calculate the utility of asymptomatic HIV infection as $8 \div 12 = 0.67$. The time-trade-off technique provides a convenient method for valuing outcomes that accounts for gains (or losses) in both length and quality of life. Because the time trade-off does not include gambles, however, it does not assess a person's risk attitude. Perhaps the strongest assumption underlying the use of the time trade-off as a measure of utility is that people are risk neutral. A **risk-neutral** decision maker is indifferent between the expected value of a gamble and the gamble itself. For example, a risk-neutral decision maker would be indifferent between the choice of living 20 years (for certain) and that of taking a gamble with a 50 % chance of living 40 years and a 50 % chance of immediate death (which has an expected value of 20 years). In practice, of course, few people are risk-neutral. Nonetheless, the time-trade-off technique is used frequently to value health outcomes because it is relatively easy to understand.

Several other approaches are available to value health outcomes. To use the **visual analog scale**, a person simply rates the quality of life with a health outcome (e.g., asymptomatic HIV infection) on a scale from 0 to 100. Although the

visual analog scale is easy to explain and use, it has no theoretical justification as a valid measure of utility. Ratings with the visual analog scale, however, correlate modestly well with utilities assessed by the standard gamble and time trade-off. For a demonstration of the use of standard gambles, time trade-offs, and the visual analog scale to assess utilities in patients with angina, see Nease et al. (1995); in patient living with HIV, see Joyce et al. (2009) and (2012). Other approaches to valuing health outcomes include the Quality of Well-Being Scale, the Health Utilities Index, and the EuroQoL (see Gold et al. 1996, ch. 4). Each of these instruments assesses how people value health outcomes and therefore may be appropriate for use in decision analyses or cost-effectiveness analyses.

In summary, we can use utilities to represent how patients value complicated health outcomes that differ in length and quality of life and in riskiness. Computer-based tools with an interactive format have been developed for assessing utilities; they often include text and multimedia presentations that enhance patients' understanding of the assessment tasks and of the health outcomes (Sumner et al. 1991; Nease and Owens 1994; Lenert et al. 1995).

3.5.5 Performance of Sensitivity Analysis

Sensitivity analysis is a test of the validity of the conclusions of an analysis over a wide range of assumptions about the probabilities and the values, or utilities. The probability of an outcome at a chance node may be the best estimate that is available, but there often is a wide range of reasonable probabilities that a clinician could use with nearly equal confidence. We use sensitivity analysis to answer this question: Do my conclusions regarding the preferred choice change when the probability and outcome estimates are assigned values that lie within a reasonable range?

The knee-replacement decision in Example 12 illustrates the power of sensitivity analysis. If the conclusions of the analysis (surgery is preferable to no surgery) remain the same despite a wide range of assumed values for the probabilities and outcome measures, the recommendation is trustworthy. Figures 3.10 and 3.11 show the expected survival in healthy years with surgery and without surgery under varying assumptions of the probability of operative death and the probability of attaining perfect mobility, respectively. Each point (value) on these lines represents one calculation of expected survival using the tree in Fig. 3.8. Figure 3.10 shows that expected survival is higher with surgery over a wide range of operative mortality rates. Expected survival is lower with surgery, however, when the operative mortality rate exceeds 25 %. Figure 3.11 shows the effect of varying the probability that the operation will lead to perfect mobility. The expected survival, in healthy years, is higher for surgery as long as the probability of perfect mobility exceeds 20 %, a much lower figure than is expected from previous experience with the operation. (In Example 12, the consulting orthopedic surgeons estimated the chance of full recovery at 60 %.) Thus, the internist can proceed with confidence

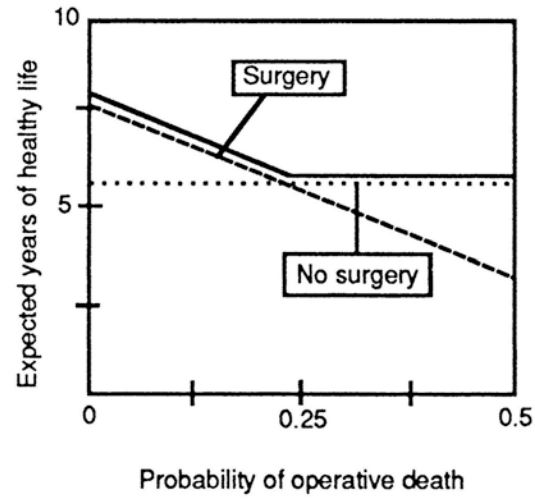

Fig. 3.10 Sensitivity analysis of the effect of operative mortality on length of healthy life (Example 12). As the probability of operative death increases, the relative values of surgery versus no surgery change. The point at which the two lines cross represents the probability of operative death at which no surgery becomes preferable. The *solid line* represents the preferred option at a given probability

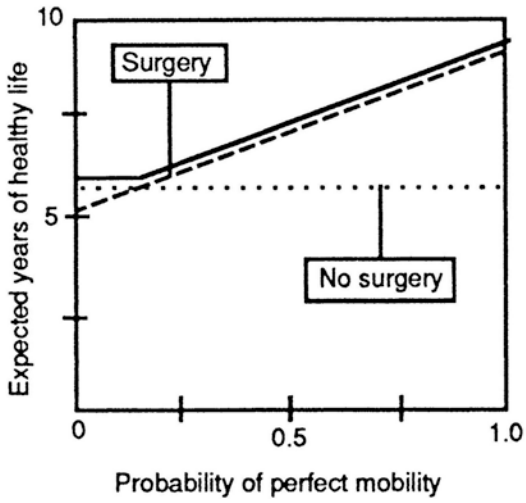

Fig. 3.11 Sensitivity analysis of the effect of a successful operative result on length of healthy life (Example 12). As the probability of a successful surgical result increases, the relative values of surgery versus no surgery change. The point at which the two lines cross represents the probability of a successful result at which surgery becomes preferable. The *solid line* represents the preferred option at a given probability

to recommend surgery. Mr. Danby cannot be sure of a good outcome, but he has valid reasons for thinking that he is more likely to do well with surgery than he is without it.

Another way to state the conclusions of a sensitivity analysis is to indicate the range of probabilities over which the conclusions apply. The point at which the two lines in Fig. 3.10 cross is the probability of operative death at which the two therapy options have the same expected survival. If expected survival is to be the basis for choosing therapy, the internist and the patient should be indifferent between surgery and no surgery when the probability of operative death is 25 %.[14] When the probability is lower, they should select surgery. When it is higher, they should select no surgery.

[14]An operative mortality rate of 25 % may seem high; however, this value is correct when we use QALYs as the basis for choosing treatment. A decision maker performing a more sophisticated analysis could use a utility function that reflects the patient's aversion to risking death.

The approach to sensitivity analyses we have described enables the analyst to understand how uncertainty in one, two, or three parameters affects the conclusions of an analysis. But in a complex problem, a decision tree or decision model may have a 100 or more parameters. The analyst may have uncertainty about many parameters in a model. **Probabilistic sensitivity analysis** is an approach for understanding how the uncertainty in all (or a large number of) model parameters affects the conclusion of a decision analysis. To perform a probabilistic sensitivity analysis, the analyst must specify a probability distribution for each model parameter. The analytic software then chooses a value for each model parameter randomly from the parameter's probability distribution. The software then uses this set of parameter values and calculates the outcomes for each alternative. For each evaluation of the model, the software will determine which alternative is preferred. The process is usually repeated 1,000–10,000 times. From the probabilistic sensitivity analysis, the analyst can determine the proportion of times an alternative is preferred, accounting for all uncertainty in model parameters simultaneously. For more information on this advanced topic, see the article by Briggs and colleagues referenced at the end of the chapter.

3.5.6 Representation of Long-Term Outcomes with Markov Models

In Example 12, we evaluated Mr. Danby's decision to have surgery to improve his mobility, which was compromised by arthritis. We assumed that each of the possible outcomes (full mobility, poor mobility, death, etc.) would occur shortly after Mr. Danby took action on his decision. But what if we want to model events that might occur in the distant future? For example, a patient with HIV infection might develop AIDS 10–15 years after infection; thus, a therapy to prevent or delay the development of AIDS could affect events that occur 10–15 years, or more, in the future. A similar

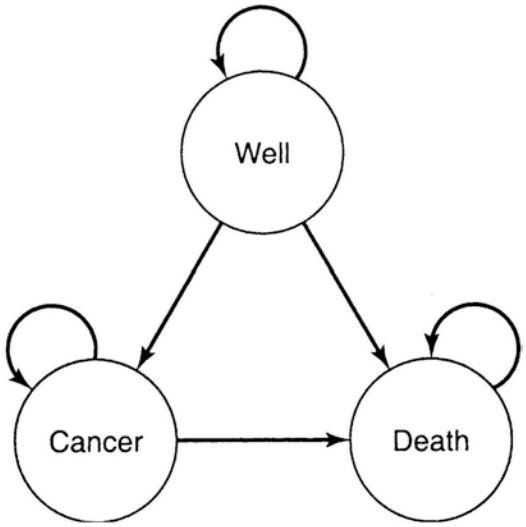

Fig. 3.12 A simple Markov model. The states of health that a person can experience are indicated by the *circles*; *arrows* represent allowed transitions between health states

problem arises in analyses of decisions regarding many chronic diseases: we must model events that occur over the lifetime of the patient. The decision tree representation is convenient for decisions for which all outcomes occur during a short time horizon, but it is not always sufficient for problems that include events that could occur in the future. How can we include such events in a decision analysis? The answer is to use Markov models (Beck and Pauker 1983; Sonnenberg and Beck 1993; Siebert et al. 2012).

To build a **Markov model**, we first specify the set of health states that a person could experience (e.g., Well, Cancer, and Death in Fig. 3.12). We then specify the **transition probabilities**, which are the probabilities that a person will transit from one of these health states to another during a specified time period. This period—often 1 month or 1 year—is the length of the **Markov cycle**. The Markov model then simulates the transitions among health states for a person (or for a hypothetical cohort of people) for a specified number of cycles; by using a Markov model, we can calculate the probability that a person will be in each of the health states at any time in the future. As an illustration, consider a simple

Table 3.7 Transition probabilities for the Markov model in Fig. 3.13

Health state transition	Annual probability
Well to well	0.9
Well to cancer	0.06
Well to death	0.04
Cancer to well	0.0
Cancer to cancer	0.4
Cancer to death	0.6
Death to well	0.0
Death to cancer	0.0
Death to death	1.0

Markov model that has three health states: Well, Cancer, and Death (see Fig. 3.12). We have specified each of the transition probabilities in Table 3.7 for the cycle length of 1 year. Thus, we note from Table 3.7 that a person who is in the well state will remain well with probability 0.9, will develop cancer with probability 0.06, and will die from non-cancer causes with probability 0.04 during 1 year. The calculations for a Markov model are performed by computer software. Based on the transition probabilities in Table 3.7, the probabilities that a person remains well, develops cancer, or dies from non-cancer causes over time is shown in Table 3.8. We can also determine from a Markov model the expected length of time that a person spends in each health state. Therefore, we can determine life expectancy, or quality-adjusted life expectancy, for any alternative represented by a Markov model.

In decision analyses that represent long-term outcomes, the analysts will often use a Markov model in conjunction with a decision tree to model the decision (Owens et al. 1995; Salpeter et al. 1997; Sanders et al. 2005). The analyst models the effect of an intervention as a change in the probability of going from one state to another. For example, we could model a cancer-prevention intervention (such as screening for breast cancer with mammography) as a reduction in the transition probability from Well to Cancer in Fig. 3.12. (See the articles by Beck and Pauker (1983) and Sonnenberg and Beck (1993) for further explanation of the use of Markov models.)

Table 3.8 Probability of future health states for the Markov model in Fig. 3.12

Health state	Probability of health state at end of year						
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Well	0.9000	0.8100	0.7290	0.6561	0.5905	0.5314	0.4783
Cancer	0.0600	0.0780	0.0798	0.0757	0.0696	0.0633	0.0572
Death	0.0400	0.1120	0.1912	0.2682	0.3399	0.4053	0.4645

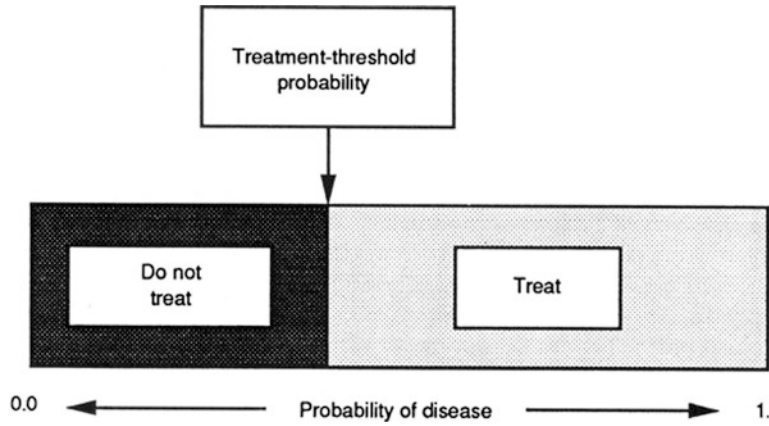

Fig. 3.13 Depiction of the treatment threshold probability. At probabilities of disease that are less than the treatment threshold probability, the preferred action is to withhold therapy. At probabilities of disease that are greater than the treatment threshold probability, the preferred action is to treat

3.6 The Decision Whether to Treat, Test, or Do Nothing

The clinician who is evaluating a patient's symptoms and suspects a disease must choose among the following actions:

1. Do nothing further (neither perform additional tests nor treat the patient).
2. Obtain additional diagnostic information (test) before choosing whether to treat or do nothing.
3. Treat without obtaining more information.

When the clinician knows the patient's true state, testing is unnecessary, and the doctor needs only to assess the trade-offs among therapeutic options (as in Example 12). Learning the patient's true state, however, may require costly, time-consuming, and often risky diagnostic procedures that may give misleading FP or FN results. Therefore, clinicians often are willing to treat a patient even when they are not absolutely certain about a patient's true state. There are risks in this course: the clinician may withhold therapy from a person who has the disease of concern, or he may administer therapy to someone who does not have the disease yet may suffer undesirable side effects of therapy.

Deciding among treating, testing, and doing nothing sounds difficult, but you have already learned all the principles that you need to solve this kind of problem. There are three steps:

1. Determine the treatment threshold probability of disease.
2. Determine the pretest probability of disease.
3. Decide whether a test result could affect your decision to treat.

The **treatment threshold probability** of disease is the probability of disease at which you should be indifferent between treating and not treating (Pauker and Kassirer 1980). Below the treatment threshold, you should not treat. Above the treatment threshold, you should treat (Fig. 3.13). Whether to treat when the diagnosis is not certain is a problem that you can solve with a decision tree, such as the one shown in Fig. 3.14.

You can use this tree to learn the treatment threshold probability of disease by leaving the probability of disease as an unknown, setting the expected value of surgery equal to the expected value for medical (i.e., nonsurgical, such as drugs or physical therapy) treatment, and solving for the probability of disease. (In this example, surgery corresponds to the "treat" branch of the tree

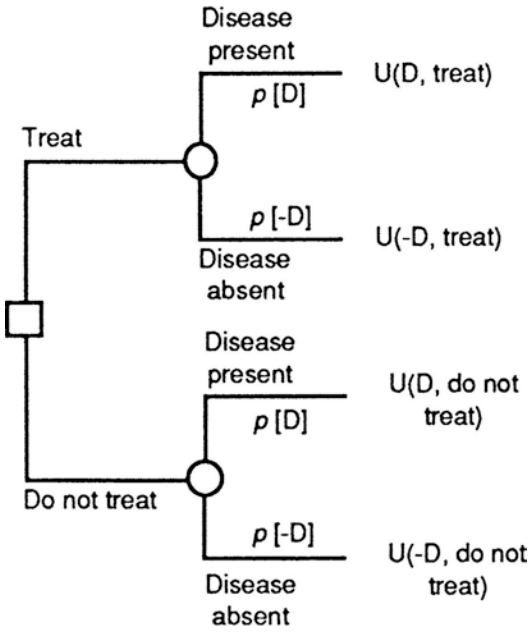

Fig. 3.14 Decision tree with which to calculate the treatment threshold probability of disease. By setting the utilities of the treat and do not treat choices to be equal, we can compute the probability at which the clinician and patient should be indifferent to the choice. Recall that p [−D] = 1 − p [D]

in Fig. 3.14, and nonsurgical intervention corresponds to the "do not treat" branch.) Because you are indifferent between medical treatment and surgery at this probability, it is the treatment threshold probability. Using the tree completes step 1. In practice, people often determine the treatment threshold intuitively rather than analytically.

An alternative approach to determination of the treatment threshold probability is to use the equation:

$$p^* = \frac{\mathrm{H}}{\mathrm{H} + \mathrm{B}},$$

where p^* = the treatment threshold probability, H = the harm associated with treatment of a nondiseased patient, and B = the benefit associated with treatment of a diseased patient (Pauker and Kassirer 1980; Sox et al. 1988). We define B as the difference between the utility (U) of diseased patients who are treated and diseased patients who are not treated (U[D, treat] − U[D, do not treat], as shown in Fig. 3.14). The utility of diseased patients who are treated should be greater than that of diseased patients who are not treated; therefore, B is positive. We define H as the difference in utility of nondiseased patients who are not treated and nondiseased patients who are treated (U[−D, do not treat] − U[−D, treat], as shown in Fig. 3.14). The utility of nondiseased patients who are not treated should be greater than that of nondiseased patients who are treated; therefore, H is positive. The equation for the treatment threshold probability fits with our intuition: if the benefit of treatment is small and the harm of treatment is large, the treatment threshold probability will be high. In contrast, if the benefit of treatment is large and the harm of treatment is small, the treatment threshold probability will be low.

Once you know the pretest probability, you know what to do in the absence of further information about the patient. If the pretest probability is below the treatment threshold, you should not treat the patient. If the pretest probability is above the threshold, you should treat the patient. Thus you have completed step 2.

One of the guiding principles of medical decision making is this: do not order a test unless it could change your management of the patient. In our framework for decision making, this principle means that you should order a test only if the test result could cause the probability of disease to cross the treatment threshold. Thus, if the pretest probability is above the treatment threshold, a negative test result must lead to a post-test probability that is below the threshold. Conversely, if the pretest probability is below the threshold probability, a positive result must lead to a post-test probability that is above the threshold. In either case, the test result would alter your decision of whether to treat the patient. This analysis completes step 3.

To decide whether a test could alter management, we simply use Bayes' theorem. We calculate the post-test probability after a test result that would move the probability of disease toward the treatment threshold. If the pretest probability is above the treatment threshold, we calculate the probability of disease if the test result is negative.

If the pretest probability is below the treatment threshold, we calculate the probability of disease if the test result is positive.

Example 13
You are a pulmonary medicine specialist. You suspect that a patient of yours has a pulmonary embolus (blood clot lodged in the vessels of the lungs). One approach is to do a computed tomography angiography (CTA) scan, a test in which a computed tomography (CT) of the lung is done after a radiopaque dye is injected into a vein. The dye flows into the vessels of the lung. The CT scan can then assess whether the blood vessels are blocked. If the scan is negative, you do no further tests and do not treat the patient.

To decide whether this strategy is correct, you take the following steps:

1. Determine the treatment threshold probability of pulmonary embolus.
2. Estimate the pretest probability of pulmonary embolus.
3. Decide whether a test result could affect your decision to treat for an embolus.

First, assume you decide that the treatment threshold should be 0.10 in this patient. What does it mean to have a treatment threshold probability equal to 0.10? If you could obtain no further information, you would treat for pulmonary embolus if the pretest probability was above 0.10 (i.e., if you believed that there was greater than a 1 in 10 chance that the patient had an embolus), and would withhold therapy if the pretest probability was below 0.10. A decision to treat when the pretest probability is at the treatment threshold means that you are willing to treat nine patients without pulmonary embolus to be sure of treating one patient who has pulmonary embolus. A relatively low treatment threshold is justifiable because treatment of a pulmonary embolism with blood-thinning medication substantially reduces the high mortality of pulmonary embolism, whereas there is only a relatively small danger (mortality of less than 1 %) in treating someone who does not have pulmonary embolus. Because the benefit of treatment is high and the harm of treatment is low, the treatment threshold probability will be low, as discussed earlier. You have completed step 1.

You estimate the pretest probability of pulmonary embolus to be 0.05, which is equal to a pretest odds of 0.053. Because the pretest probability is lower than the treatment threshold, you should do nothing unless a positive CTA scan result could raise the probability of pulmonary embolus to above 0.10. You have completed step 2.

To decide whether a test result could affect your decision to treat, you must decide whether a positive CTA scan result would raise the probability of pulmonary embolism to more than 0.10, the treatment threshold. You review the literature and learn that the LR for a positive CTA scan is approximately 21 (Stein et al. 2006).

A negative CTA scan result will move the probability of disease away from the treatment threshold and will be of no help in deciding what to do. A positive result will move the probability of disease toward the treatment threshold and could alter your management decision if the post-test probability were above the treatment threshold. You therefore use the odds-ratio form of Bayes' theorem to calculate the post-test probability of disease if the lung scan result is reported as high probability.

$$\text{Post-test odds} = \text{pretest odds} \times \text{LR}$$
$$= 0.053 \times 21 = 1.11.$$

A post-test odds of 1.1 is equivalent to a probability of disease of 0.53. Because the post-test probability of pulmonary embolus is higher than the treatment threshold, a positive CTA scan result would change your management of the patient, and you should order the lung scan. You have completed step 3.

This example is especially useful for two reasons: first, it demonstrates one method for making decisions and second, it shows how the concepts that were introduced in this chapter all fit together in a clinical example of medical decision making.

3.7 Alternative Graphical Representations for Decision Models: Influence Diagrams and Belief Networks

In Sects. 3.5 and 3.6, we used decision trees to represent decision problems. Although decision trees are the most common graphical representation for decision problems, **influence diagrams** are an important alternative representation for such problems (Nease and Owens 1997; Owens et al. 1997).

As shown in Fig. 3.15, influence diagrams have certain features that are similar to decision trees, but they also have additional graphical elements. Influence diagrams represent decision nodes as squares and chance nodes as circles. In contrast to decision trees, however, the influence diagram also has arcs between nodes and a diamond-shaped value node. An arc between two chance nodes indicates that a probabilistic relationship may exist between the chance nodes (Owens et al. 1997). A **probabilistic relationship** exists when the occurrence of one chance event affects the probability of the occurrence of another chance event. For example, in Fig. 3.15, the probability of a positive or negative PCR test result (PCR result) depends on whether a person has HIV infection (HIV status); thus, these nodes have a probabilistic relationship, as indicated by the arc. The arc points from the **conditioning event** to the **conditioned event** (PCR test result is conditioned on HIV status in Fig. 3.15). The absence of an arc between two chance nodes, however, always indicates that the nodes are independent or conditionally independent. Two events are conditionally independent, given a third event, if the occurrence of one of the events does not affect the probability of the other event conditioned on the occurrence of the third event.

Unlike a decision tree, in which the events usually are represented from left to right in the order in which the events are observed, influence diagrams use arcs to indicate the timing of events. An arc from a chance node to a decision node indicates that the chance event has been observed at the time the decision is made. Thus, the arc from PCR result to Treat? in Fig. 3.15 indicates that the decision maker knows the PCR test result (positive, negative, or not obtained) when he or she decides whether to treat. Arcs between decision nodes indicate the timing of decisions: the arc points from an initial decision to subsequent decisions. Thus, in Fig. 3.15, the decision maker must decide whether to obtain a PCR test before deciding whether to treat, as indicated by the arc from Obtain PCR? to Treat?

The probabilities and utilities that we need to determine the alternative with the highest expected value are contained in tables associated with chance nodes and the value node (Fig. 3.16). These tables contain the same information that we would use in a decision tree. With a decision tree, we can determine the expected value of each alternative by averaging out at chance nodes and folding back the tree (Sect. 3.5.3). For influence diagrams, the calculation of expected value is more complex (Owens et al. 1997), and generally must be performed with computer software. With the appropriate software, we can use influence diagrams to perform the same analyses that we would perform with a decision tree. Diagrams that have only chance nodes are called **belief networks**; we use them to perform probabilistic inference.

Why use an influence diagram instead of a decision tree? Influence diagrams have both advantages and limitations relative to decision trees. Influence diagrams represent graphically the probabilistic relationships among variables (Owens et al. 1997). Such representation is advantageous for problems in which probabilistic conditioning is complex or in which communication of such conditioning is important (such as may occur in large models). In an influence diagram, probabilistic conditioning is indicated by the arcs, and thus the conditioning is apparent immediately by inspection. In a decision tree, probabilistic conditioning is revealed by the probabilities in the branches of the tree. To determine whether events are conditionally independent in a decision tree requires that the analyst compare probabilities of events between branches of the tree. Influence diagrams also are particularly useful for discussion with content experts who can help to structure a problem but who are not familiar with decision analysis. In contrast, problems that have decision

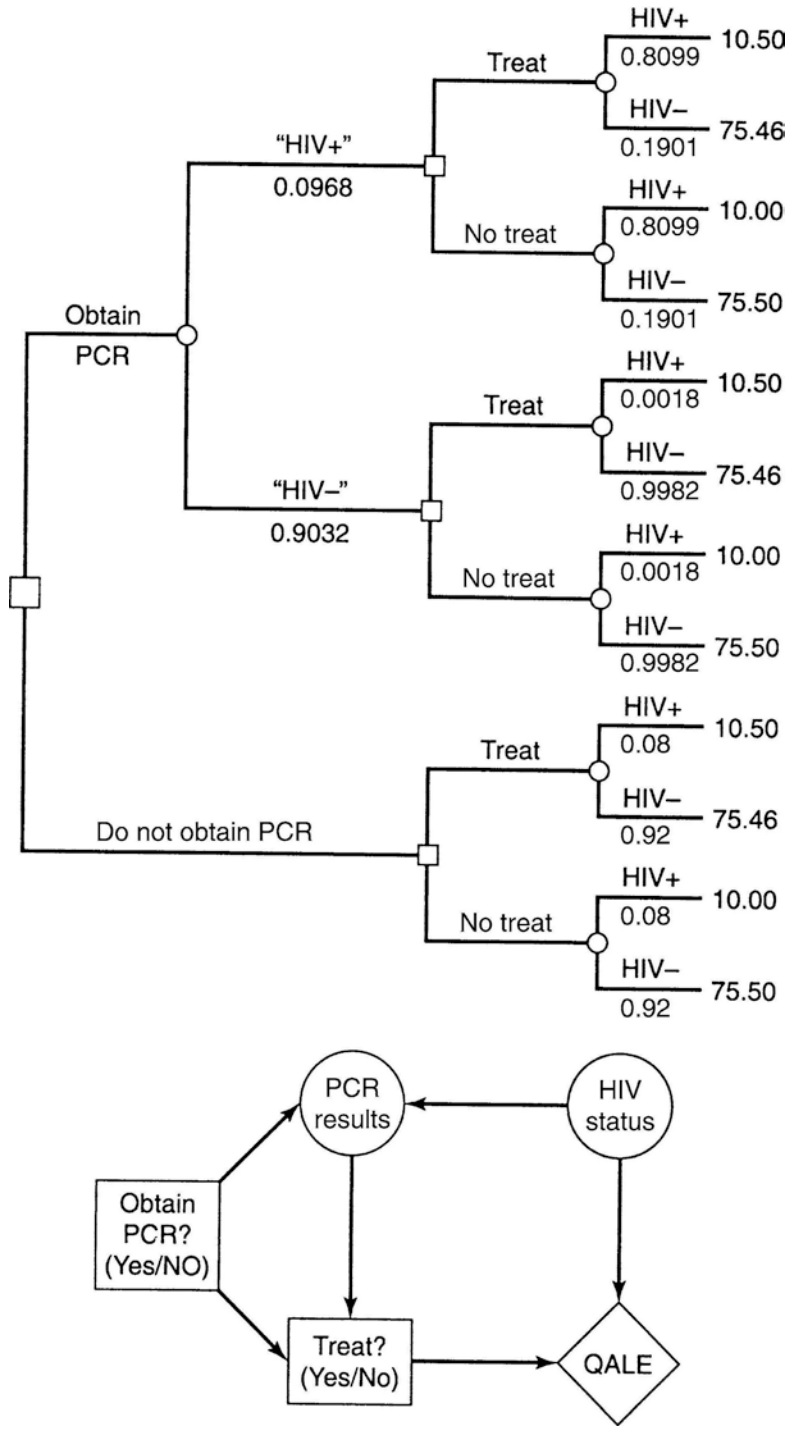

Fig. 3.15 A decision tree (*top*) and an influence diagram (*bottom*) that represent the decisions to test for, and to treat, HIV infection. The structural asymmetry of the alternatives is explicit in the decision tree. The influence diagram highlights probabilistic relationships. *HIV* human immunodeficiency virus, *HIV+* HIV infected, *HIV–* not infected with HIV, *QALE* quality-adjusted life expectancy, *PCR* polymerase chain reaction. Test results are shown in quotation marks ("HIV+"), whereas the true disease state is shown without quotation marks (HIV+) (Source: Owens et al. (1997). Reproduced with permission)

alternatives that are structurally different may be easier for people to understand when represented with a decision tree, because the tree shows the structural differences explicitly, whereas the influence diagram does not. The choice of whether to use a decision tree or an influence diagram depends

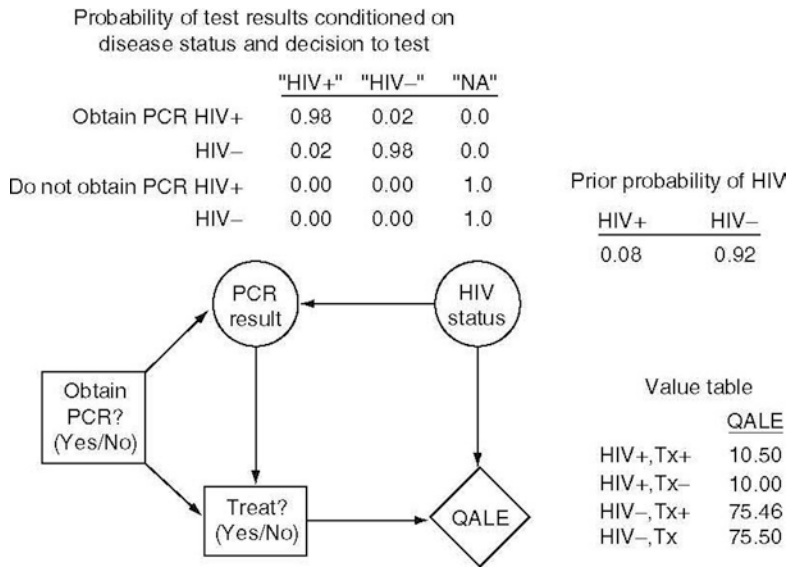

Fig. 3.16 The influence diagram from Fig. 3.15, with the probability and value tables associated with the nodes. The information in these tables is the same as that associated with the branches and endpoints of the decision tree in Fig. 3.15. *HIV* human immunodeficiency virus, *HIV+* HIV infected, *HIV–* not infected with HIV, *QALE* quality-adjusted life expectancy, *PCR* polymerase chain reaction, *NA* not applicable, *TX+* treated, *TX–* not treated. Test results are shown in quotation marks ("HIV+"), and the true disease state is shown without quotation marks (HIV+) (Source: Owens et al.. (1997). Reproduced with permission)

on the problem being analyzed, the experience of the analyst, the availability of software, and the purpose of the analysis. For selected problems, influence diagrams provide a powerful graphical alternative to decision trees.

3.8 Other Modeling Approaches

We have described decision trees, Markov models and influence diagrams. An analyst also can choose several other approaches to modeling. The choice of modeling approach depends on the problem and the objectives of the analysis. Although how to choose and design such models is beyond our scope, we note other type of models that analysts use commonly for medical decision making. **Microsimulation models** are individual-level health state transition models, similar to Markov models, that provide a means to model very complex events flexibly over time. They are useful when the clinical history of a problem is complex, such as might occur with cancer, heart disease, and other chronic diseases. **Dynamic transmission models** are particularly well-suited for assessing the outcomes of infectious diseases. These models divide a population into compartments (for example, uninfected, infected, recovered, dead), and transitions between compartments are governed by differential or difference equations. The rate of transition between compartments depends in part on the number of individuals in the compartment, an important feature for infectious diseases in which the transmission may depend on the number of infected or susceptible individuals. **Discrete event simulation models** also are often used to model interactions between people. These models are composed of entities (a patient) that have attributes (clinical history), and that experience events (a heart attack). An entity can interact with other entities and use resources. For more information on these types of models, we suggest a recent series of papers on best modeling practices; the paper by Caro and colleagues noted in the suggested readings at the end of the chapter is an overview of this series of papers.

3.9 The Role of Probability and Decision Analysis in Medicine

You may be wondering how probability and decision analysis might be integrated smoothly into medical practice. An understanding of probability and measures of test performance will prevent any number of misadventures. In Example 1, we discussed a hypothetical test that, on casual inspection, appeared to be an accurate way to screen blood donors for previous exposure to the AIDS virus. Our quantitative analysis, however, revealed that the hypothetical test results were misleading more often than they were helpful because of the low prevalence of HIV in the clinically relevant population. Fortunately, in actual practice, much more accurate tests are used to screen for HIV.

The need for knowledgeable interpretation of test results is widespread. The federal government screens civil employees in "sensitive" positions for drug use, as do many companies. If the drug test used by an employer had a sensitivity and specificity of 0.95, and if 10 % of the employees used drugs, one-third of the positive tests would be FPs. An understanding of these issues should be of great interest to the public, and health professionals should be prepared to answer the questions of their patients.

Although we should try to interpret every kind of test result accurately, decision analysis has a more selective role in medicine. Not all clinical decisions require decision analysis. Some decisions depend on physiologic principles or on deductive reasoning. Other decisions involve little uncertainty. Nonetheless, many decisions must be based on imperfect data, and they will have outcomes that cannot be known with certainty at the time that the decision is made. Decision analysis provides a technique for managing these situations.

For many problems, simply drawing a tree that denotes the possible outcomes explicitly will clarify the question sufficiently to allow you to make a decision. When time is limited, even a "quick and dirty" analysis may be helpful. By using expert clinicians' subjective probability estimates and asking what the patient's utilities might be, you can perform an analysis quickly and learn which probabilities and utilities are the important determinants of the decision.

Health care professionals sometimes express reservations about decision analysis because the analysis may depend on probabilities that must be estimated, such as the pretest probability. A thoughtful decision maker will be concerned that the estimate may be in error, particularly because the information needed to make the estimate often is difficult to obtain from the medical literature. We argue, however, that uncertainty in the clinical data is a problem for any decision-making method and that the effect of this uncertainty is explicit with decision analysis. The method for evaluating uncertainty is sensitivity analysis: we can examine any variable to see whether its value is critical to the final recommended decision. Thus, we can determine, for example, whether a change in pretest probability from 0.6 to 0.8 makes a difference in the final decision. In so doing, we often discover that it is necessary to estimate only a range of probabilities for a particular variable rather than a precise value. Thus, with a sensitivity analysis, we can decide whether uncertainty about a particular variable should concern us.

The growing complexity of medical decisions, coupled with the need to control costs, has led to major programs to develop clinical practice guidelines. Decision models have many advantages as aids to guideline development (Eddy 1992): they make explicit the alternative interventions, associated uncertainties, and utilities of potential outcomes. Decision models can help guideline developers to structure guideline-development problems (Owens and Nease 1993), to incorporate patients' preferences (Nease and Owens 1994; Owens 1998), and to tailor guidelines for specific clinical populations (Owens and Nease 1997). In addition, Web-based interfaces for decision models can provide distributed decision support for guideline developers and users by making the decision model available for analysis to anyone who has access to the Web (Sanders et al. 1999).

We have not emphasized computers in this chapter, although they can simplify many aspects of decision analysis (see Chap. 22). MEDLINE and other bibliographic retrieval systems (see Chap. 21) make it easier to obtain published estimates of

disease prevalence and test performance. Computer programs for performing statistical analyses can be used on data collected by hospital information systems. Decision analysis software, available for personal computers, can help clinicians to structure decision trees, to calculate expected values, and to perform sensitivity analyses. Researchers continue to explore methods for computer-based automated development of practice guidelines from decision models and use of computer-based systems to implement guidelines (Musen et al. 1996). With the growing maturity of this field, there are now companies that offer formal analytical tools to assist with clinical outcome assessment and interpretation of population datasets.[15]

Medical decision making often involves uncertainty for the clinician and risk for the patient. Most health care professionals would welcome tools that help them make decisions when they are confronted with complex clinical problems with uncertain outcomes. There are important medical problems for which decision analysis offers such aid.

Suggested Readings

Briggs, A., Weinstein, M., Fenwick, E., Karnon, J., Sculpher, M., & Paltiel, A. (2012). Model parameter estimation and uncertainty analysis: A report of the ISPOR-SMDM modeling good research practices task force-6. *Medical Decision Making, 32*(5), 722–732. This article describes best practices for estimating model parameters and for performing sensitivity analyses, including probabilistic sensitivity analysis.

Caro, J., Briggs, A., Siebert, U., & Kuntz, K. (2012). Modeling good research practices – overview: A report of the ISPOR-SMDM modeling good research practices task force-1. *Value in Health, 15*, 796–803. This paper is an introduction to a series of papers that describe best modeling practices.

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost effectiveness in health and medicine*. New York: Oxford University Press. This book provides authoritative guidelines for the conduct of cost-effectiveness analyses. Chapter 4 discusses approaches for valuing health outcomes.

Hunink, M., Glasziou, P., Siegel, J., Weeks, J., Pliskin, J., Einstein, A., & Weinstein, M. (2001). *Decision making in health and medicine*. Cambridge: Cambridge University Press. This textbook addresses in detail most of the topics introduced in this chapter.

Nease, R. F., Jr., & Owens, D. K. (1997). Use of influence diagrams to structure medical decisions. *Medical Decision Making, 17*(13), 263–275. This article provides a comprehensive introduction to the use of influence diagrams.

Owens, D. K., Schacter, R. D., & Nease, R. F., Jr. (1997). Representation and analysis of medical decision problems with influence diagrams. *Medical Decision Making, 17*(3), 241–262. This article provides a comprehensive introduction to the use of influence diagrams.

Raiffa, H. (1970). *Decision analysis: Introductory lectures on choices under uncertainty*. Reading: Addison-Wesley. This now classic book provides an advanced, nonmedical introduction to decision analysis, utility theory, and decision trees.

Sox, H. C. (1986). Probability theory in the use of diagnostic tests. *Annals of Internal Medicine, 104*(1), 60–66. This article is written for clinicians; it contains a summary of the concepts of probability and test interpretation.

Sox, H. C., Higgins, M. C., & Owens, D. K. (2013). *Medical decision making*. Chichester: Wiley-Blackwell. This introductory textbook covers the subject matter of this chapter in greater detail, as well as discussing many other topics.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124. This now classic article provides a clear and interesting discussion of the experimental evidence for the use and misuse of heuristics in situations of uncertainty.

Questions for Discussion

1. Calculate the following probabilities for a patient about to undergo CABG surgery (see Example 2):
   (a) The only possible, mutually exclusive outcomes of surgery are death, relief of symptoms (angina and dyspnea), and continuation of symptoms. The probability of death is 0.02, and the probability of relief of symptoms is 0.80. What is the probability that the patient will continue to have symptoms?
   (b) Two known complications of heart surgery are stroke and heart attack, with probabilities of 0.02 and 0.05, respectively. The patient asks what chance he or she has of having both

[15]See, for example, the Archimedes tools described at http://archimedesmodel.com/.

Table 3.9 A 2×2 contingency table for the hypothetical study in problem 2

PCR test result	Gold standard test positive	Gold standard test negative	Total
Positive PCR	48	8	56
Negative PCR	2	47	49
Total	50	55	105

PCR polymerase chain reaction

complications. Assume that the complications are conditionally independent, and calculate your answer.

(c) The patient wants to know the probability that he or she will have a stroke given that he or she has a heart attack as a complication of the surgery. Assume that 1 in 500 patients has both complications, that the probability of heart attack is 0.05, and that the events are independent. Calculate your answer.

2. The results of a hypothetical study to measure test performance of a diagnostic test for HIV are shown in the 2×2 table in Table 3.9.
   (a) Calculate the sensitivity, specificity, disease prevalence, PV+, and PV–.
   (b) Use the TPR and TNR calculated in part (a) to fill in the 2×2 table in Table 3.10. Calculate the disease prevalence, PV+, and PV–.
3. You are asked to interpret the results from a diagnostic test for HIV in an asymptomatic man whose test was positive when he volunteered to donate blood. After taking his history, you learn that he has a history of intravenous-drug use. You know that the overall prevalence of HIV infection in your community is 1 in 500 and that the prevalence in people who have injected drugs is 20 times as high as in the community at large.
   (a) Estimate the pretest probability that this man is infected with HIV.

Table 3.10 A 2×2 contingency table to complete for problem 2b

PCR test result	Gold standard test positive	Gold standard test negative	Total
Positive PCR	x	x	x
Negative PCR	100	99,900	x
Total	x	x	x

PCR polymerase chain reaction
x quantities that the question ask students to calculate

   (b) The man tells you that two people with whom he shared needles subsequently died of AIDS. Which heuristic will be useful in making a subjective adjustment to the pretest probability in part (a)?
   (c) Use the sensitivity and specificity that you worked out in 2(a) to calculate the post-test probability of the patient having HIV after a positive and negative test. Assume that the pretest probability is 0.10.
   (d) If you wanted to increase the post-test probability of disease given a positive test result, would you change the TPR or TNR of the test?
4. You have a patient with cancer who has a choice between surgery or chemotherapy. If the patient chooses surgery, he or she has a 2 % chance of dying from the operation (life expectancy = 0), a 50 % chance of being cured (life expectancy = 15 years), and a 48 % chance of not being cured (life expectancy = 1 year). If the patient chooses chemotherapy, he or she has a 5 % chance of death (life expectancy = 0), a 65 % chance of cure (life expectancy = 15 years), and a 30 % chance that the cancer will be slowed but not cured (life expectancy = 2 years). Create a decision tree.

Calculate the expected value of each option in terms of life expectancy.

5. You are concerned that a patient with a sore throat has a bacterial infection that would require antibiotic therapy (as opposed to a viral infection, for which no treatment is available). Your treatment threshold is 0.4, and based on the examination you estimate the probability of bacterial infection as 0.8. A test is available (TPR = 0.75, TNR = 0.85) that indicates the presence or absence of bacterial infection. Should you perform the test? Explain your reasoning. How would your analysis change if the test were extremely costly or involved a significant risk to the patient?
6. What are the three kinds of bias that can influence measurement of test performance? Explain what each one is, and state how you would adjust the post-test probability to compensate for each.
7. How could a computer system ease the task of performing a complex decision analysis?
8. When you search the medical literature to find probabilities for patients similar to one you are treating, what is the most important question to consider? How should you adjust probabilities in light of the answer to this question?
9. Why do you think clinicians sometimes order tests even if the results will not affect their management of the patient? Do you think the reasons that you identify are valid? Are they valid in only certain situations? Explain your answers. See the January 1998 issue of Medical Decision Making for articles that discuss this question.
10. Explain the differences in three approaches to assessing patients' preferences for health states: the standard gamble, the time trade-off, and the visual analog scale.

Appendix: Derivation of Bayes' Theorem

Bayes' theorem is derived as follows. We denote the conditional probability of disease, D, given a test result, R, p[D|R]. The prior (pretest) probability of D is p[D]. The definition of conditional probability is:

$$p[\mathrm{D}\,|\,\mathrm{R}] = \frac{p[\mathrm{R},\mathrm{D}]}{p[\mathrm{R}]} \tag{3.1}$$

The probability of a test result (p[R]) is the sum of its probability in diseased patients and its probability in nondiseased patients:

$$p[\mathrm{R}] = \mathrm{p}[\mathrm{R},\mathrm{D}] + p[\mathrm{R},-\mathrm{D}].$$

Substituting into Equation 3.1, we obtain:

$$p[\mathrm{D}\,|\,\mathrm{R}] = \frac{p[\mathrm{R},\mathrm{D}]}{p[\mathrm{R},\mathrm{D}] + p[\mathrm{R}-\mathrm{D}]} \tag{3.2}$$

Again, from the definition of conditional probability,

$$p[\mathrm{R}|\mathrm{D}] = \frac{p[\mathrm{R},\mathrm{D}]}{p[\mathrm{D}]} \text{ and } p[\mathrm{R}\,|\,-\mathrm{D}] = \frac{p[\mathrm{R},-\mathrm{D}]}{p[-\mathrm{D}]}$$

These expressions can be rearranged:

$$p[\mathrm{R},\mathrm{D}] = p[\mathrm{D}] \times p[\mathrm{R}\,|\mathrm{D}], \tag{3.3}$$

$$p[\mathrm{R},-\mathrm{D}] = p[-\mathrm{D}] \times p[\mathrm{R}\,|\,-\mathrm{D}]. \tag{3.4}$$

Substituting Eqs. 3.3 and 3.4 into Eq. 3.2, we obtain Bayes' theorem:

$$p[\mathrm{D}\,|\,\mathrm{R}] = \frac{p[\mathrm{D}] \times p[\mathrm{R}\,|\,\mathrm{D}]}{p[\mathrm{D}] \times p[\mathrm{R}\,|\,\mathrm{D}] + \mathrm{p}[-\mathrm{D}] \times p[\mathrm{R}\,|\,-\mathrm{D}]}$$