2
Assessment of Surveillance Test Performance and Cost

Katherine S. Virgo

Surveillance test performance and costs are important considerations for a book summarizing the state of the art in patient management after implantation of prosthetic devices for several reasons. First is the growing concern about more efficient use of limited resources, made ever more apparent by the ongoing health care reform debate. Second is the expanding role of “gatekeepers” wielding increasing control over the total cost of health care. Third is the push for clinicians to develop guidelines, driven by the idea that reimbursement could subsequently be tied to adherence to these guidelines.

Given such issues and the increasing cost of high-tech prosthetic devices, one would expect to see a plethora of articles assessing various follow-up strategies for efficacy, efficiency, and cost-effectiveness, but few exist. One reason for the shortage of articles may be that the appropriate patient-management strategy after implantation is not well delineated for many devices. Much of the existing literature on patient follow-up after implantation of prosthetic devices consists of articles that suggest strategies based on either inadequate sample size or data from a single institution. Very few proposed patient-management strategies are based on the results of large retrospective analyses of secondary data sets or prospective, randomized clinical trials.

Another reason for the shortage of articles may be that few clinicians sufficiently understand epidemiological and cost-analysis methodologies. Therefore, this chapter provides a review of the tools needed to weigh alternative follow-up strategies against one another. The epidemiology section describes how to determine whether screening for disease is appropriate in a given population, assess the performance of individual diagnostic tests, compare performance across diagnostic tests, and determine whether further diagnostic testing is required. The economics section specifies how to calculate and compare the costs and benefits of individual diagnostic tests or entire follow-up strategies.

From: *The Bionic Human: Health Promotion for People With Implanted Prosthetic Devices*
Edited by: F. E. Johnson and K. S. Virgo © Humana Press Inc., Totowa, NJ

EPIDEMIOLOGICAL PRINCIPLES FOR EVALUATING DIAGNOSTIC TEST PERFORMANCE IN SCREENING FOR DEVICE MALFUNCTION OR INFECTION

Overview

Epidemiology is the study of the frequency and determinants of disease and injury in human populations *(1)*. Whereas clinical medicine focuses on the delivery of medical care to patients, epidemiology analyzes why different populations have differing incidences and prevalences of disease. Incidence refers to the probability that individuals without disease will develop disease over a given period of time and is calculated as the number of new cases of disease divided by the population at risk. Prevalence refers to the number of people in a population who already have the disease and is calculated as the number of existing cases of disease divided by the total population. Clinical epidemiology focuses on the application of epidemiological principles to the practice of clinical medicine. This section uses basic principles of epidemiology to assess diagnostic test performance in screening for device malfunction or infection.

Screening

Asymptomatic patients rarely seek care unless participating in a regular surveillance program. When symptoms do appear, patients often delay seeking care for an extended period, during which time the condition worsens. It is generally believed that early detection through the use of screening tests improves the probability of repairing device malfunction or treating infection and reduces the probabilities of both death and disability.

Screening is the use of tests or examinations to distinguish asymptomatic individuals with a high probability of disease from asymptomatic individuals with a low probability of disease. Some screening programs are designed to identify individuals who might not have disease now but who have a high probability of developing it in the future. Screening tests are usually quick, minimally invasive, and inexpensive. Usually, screening or surveillance is performed among populations who have not previously been diagnosed with the disease under evaluation. However, the term *surveillance* can also be used to refer to the follow-up of patients after implantation of prosthetic devices to detect device malfunction or infection. Although the term *screening* is used throughout this section, the same concepts apply to both screening and surveillance.

DIAGNOSTIC TEST CHARACTERISTICS

Validity

Important characteristics of screening tests are validity, reliability, and yield. Validity refers to the ability of a test to distinguish between those who have disease and those who do not. The two measures of validity are sensitivity and specificity. Sensitivity, often referred to as the true positive rate (TPR), measures the ability of the test to correctly identify those who actually have disease. Sensitivity is calculated as the percentage of all patients with disease who screen positive for disease. Specificity, also referred to as the true negative rate (TNR), measures the ability of the test to correctly identify those who do not have disease. Specificity is calculated as the percentage of all patients without disease who screen negative for disease *(1)*.

Table 1
Derivation of Sensitivity and Specificity

	Disease category	
Screening test result	Disease present	Disease absent
Positive	*a* (TP)	*b* (FP)
Negative	*c* (FN)	*d* (TN)

Sensitivity = *a* / (*a* + *c*) = TP / (TP + FN)
Specificity = *d* / (*b* + *d*) = TN / (FP + TN)
TP, true-positive; FN, false-negative; TN, true-negative; FP, false-positive.

Table 1 depicts the derivation of sensitivity and specificity *(2)*. Patients who are correctly predicted by the diagnostic test of interest to have disease are referred to as true-positives (TP). Similarly, those patients who are correctly predicted to be disease free are referred to as true-negatives (TN). Those patients falsely predicted to have disease are false-positives (FP). Those patients falsely predicted to be disease free are false-negatives (FN). Once the sensitivity and specificity of a diagnostic test are known, clinicians can use these estimates to revise original estimates of the probability of disease made prior to the ordering of a diagnostic test (pretest probability). According to a principle known as Bayes' theorem, posttest probability can be calculated as:

$$\mathrm{P_r} = \frac{(P_i)(\text{sensitivity})}{(P_i)(\text{sensitivity}) + (1-P_i)(100\%-\text{specificity})}$$

where P_r is the posttest probability and P_i is the pretest probability. Tabular and graphic expressions of Bayes' theorem are also available, such as Bayes' nomogram and Benish's tables, which permit one to look up the posttest probability once the pretest probability, sensitivity, and specificity are known *(3,4)*.

Sensitivity and specificity are derived by comparing the results from the test in question (the index test) with those of a definitive test (a gold standard test). Irrespective of the results of the screen (positive or negative), in most cases, every person screened must be tested using the gold standard to establish or rule out disease *(5,6)*. The optimal test would be 100% specific and 100% sensitive. Unfortunately, this will not be observed in practice because sensitivity and specificity are usually inversely related. In other words, sensitivity can be improved, but only at the expense of specificity, and specificity can be improved, but only at the expense of sensitivity.

To understand this, consider that range of diagnostic test results that can be considered either normal or abnormal, as depicted by the overlapping bell-shaped curves in Fig. 1. If the range of overlapping values is 20 to 40 and the line distinguishing normal from abnormal test results is drawn so that 20 and above is considered abnormal, the screening intervention will have high sensitivity because all patients with diagnostic test results in the 20 and above range will be treated as TP. However, the intervention will have low specificity because many of the results treated as positive will turn out to be FP. Alternatively, if the line distinguishing normal from abnormal test results is

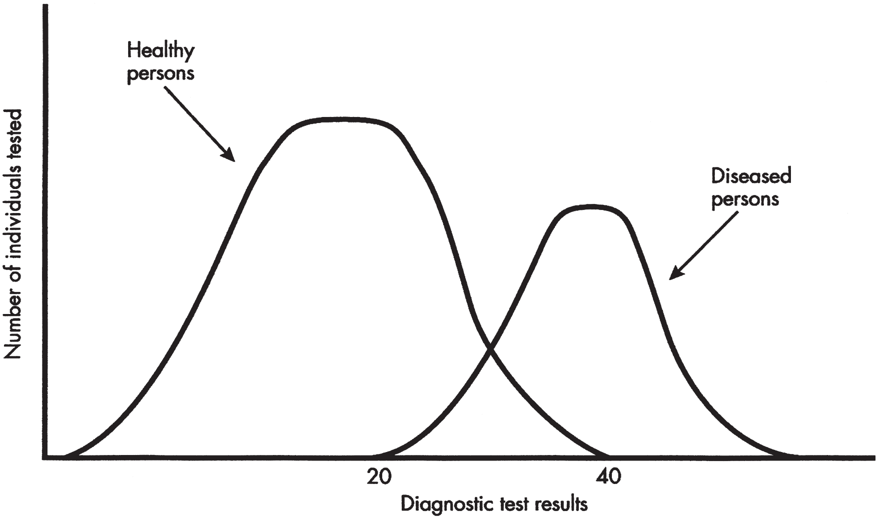

Fig. 1. Distribution of diagnostic test results.

redrawn so that 40 and above is considered abnormal and all other values are considered normal, the screening intervention will have low sensitivity and high specificity because all patients with diagnostic test results in the 39 and below range will be treated as TN.

Other factors that influence measurement of the validity of a test are severity of disease and the presence of co-morbid conditions. With some diagnostic tests, such as the serological test for syphilis, the probability of FN is very high in the early or very late stages of disease *(1)*. The presence of co-morbid conditions and drugs taken for these conditions can also greatly influence diagnostic test results.

The ability of a diagnostic test to correctly discriminate between the presence or absence of disease is also dependent on the prevalence of disease, in addition to a test's specificity and sensitivity. The greater the prevalence of disease, the greater the predictive value (PV) of a positive test, which is the probability that a positive test result is accurately predicting disease. As prevalence approaches zero, the PV of a positive test approaches zero. The PV of a positive test is calculated as $a / (a + b)$ or TP / (TP + FP). According to Bayes' theorem of conditional probabilities, the PV of a positive test can also be calculated as (sensitivity × prevalence) / [(sensitivity × prevalence) + (1 – specificity) × (1 – prevalence)] *(7)*. The PV of a negative test is $d / (c + d)$ or TN / (FN + TN). According to Vecchio *(8)*, the prevalence of a disease must be at least 15 to 20% to reach an acceptable PV (70–80%).

Reliability

The second important characteristic of screening tests is reliability or precision. Reliability measures whether the same test administered more than once to the same person will produce the same results repetitively. The two types of variation that can occur are variation in the method itself and variation related to the person(s) interpreting the results. Variation in method can be the result of mechanical fluctuations (fluctuations in the testing apparatus) or fluctuations in the substance being measured by the diagnostic test. Variation related to the interpretation of the results can be classified into two types. Intraobserver variation is variation caused by one person interpreting the results differently on different occasions. Interobserver variation is variation across different persons interpreting the results *(9)*. Such variation can be substantially reduced through training seminars and the use of independent observations on a subsample of cases.

Yield

The third important characteristic of screening tests is yield, which refers to the number of cases with previously undiagnosed disease that are detected and treated as a result of the screen. Yield is affected by the sensitivity of the diagnostic test, the prevalence of unrecognized disease, whether the screening is multiphasic (multiple diagnostic tests were administered), screening frequency, and the number of positive screens who actually receive treatment *(1)*. The effect of sensitivity on yield is that, if few TPs are identified, the other factors become immaterial, because yield will be low. If the prevalence of unrecognized disease is low, because of such factors as high medical care availability or a recent screen of the population, the yield will be low.

The ability to identify risk factors for the disease and narrow down the number of individuals who must be screened will increase yield. Another way to increase yield is through multiphasic screening in which a variety of tests are used to screen for multiple conditions during one visit.

Frequency of Screening

On the issue of frequency of screening, the literature is not very clear in many instances. Frequency should be dictated by the natural history of the disease, the incidence of disease, and risk factors. Whether a patient with identified disease will consent to treatment is determined by whether the patient views there to be a serious threat to health, whether the patient feels vulnerable, and whether the patient decides that seeking treatment will be beneficial *(1)*.

LIKELIHOOD RATIOS

Another way to measure the performance of a diagnostic test, which has not yet been discussed, is through the use of likelihood ratios. To understand likelihood ratios, which are a type of odds ratio, the difference between probability and odds must be clear. Probability ranges from zero to 1 and measures the likelihood that a particular outcome will occur. A value close to zero indicates little chance of occurrence; a value close to 1 indicates a large chance of occurrence. If an experiment is conducted *n* times and the event of interest occurs *m* times, the probability of that event occurring is calculated as *m* / *n (10)*. Sensitivity and specificity are both measures of the probabilities of specific events occurring.

Odds are ratios of two probabilities and are calculated as the probability of an event / (1 – the probability of an event) *(7)*. One can also work backward and calculate probability from odds using the following equation: odds / (1 + odds). Likelihood ratios measure how much more likely it is that a diagnosis will be made in the presence of disease as in the absence of disease and can be defined for any number of test results over the entire range of possible values. For positive and negative test results, the respective likelihood ratios are sensitivity / (1 – specificity) and (1 – sensitivity) / specificity *(11)*. Use of likelihood ratios has the advantage of placing more weight on very high and very low test results than on borderline results when attempting to determine the odds that a disease is really present. In comparison to sensitivity and specificity measures, another advantage of likelihood ratios is that diagnostic test performance is quantified as one measure rather than two. A disadvantage of likelihood ratios is that the conversion from probability to odds and back again can be difficult.

REQUIREMENTS FOR ESTABLISHING A SCREENING PROGRAM

There are several major issues identified by Wilson and Jungner *(12)* that should serve as a prerequisite for the establishment of a screening program. Among these are that the health problem must be important, the disease should have either a latent stage or an early treatable stage, a diagnostic test acceptable to the population should be available, the natural history of the condition should be sufficiently understood, treatment should be available for identified cases, there should be clarity regarding which cases can be curatively treated, and screening should be cost effective.

RECEIVER OPERATING CHARACTERISTIC CURVE

Once the need for a screening program is established and the appropriate diagnostic test is selected, the next step is to clarify how test results will be interpreted. Complicating the situation is that factors such as age, sex, race, and nutrition can all impact laboratory test results. For example, what is normal for a 70-year-old male may not be normal for a 25-year-old female. Although what is considered normal can vary by patient, the distribution of clinical measurements for an individual is generally represented by a normally distributed (bell-shaped) curve *(13)*. The dispersion of values around the mean in a normal distribution is due to random variation alone.

In addition to variation across subjects in terms of what is normal and abnormal, the cutoff between normal and abnormal for a given diagnostic test can be varied given the goals of the particular screening intervention. If the goal is to correctly identify, for example, 95% of all cases of disease, the range of values constituting an abnormal test result can be expanded until this goal is reached. Unfortunately, doing so causes the number of FP to increase, thus decreasing specificity, because sensitivity and specificity are inversely related. Similarly, if the goal is to correctly identify 95% of all cases without disease, the range of values constituting a normal test result can be expanded until this goal is reached. However, increasing specificity is achieved at the expense of decreasing sensitivity.

When diagnostic test results by patient are depicted graphically for both healthy and diseased individuals (Fig.1), there is usually a range of values that is clearly normal and another range of values that is clearly abnormal. However, there is also a range of values that could easily represent either normal or abnormal results, as depicted by the overlapping bell-shaped curves. Figure 1 depicts how the selected cutoff between normal and abnormal determines the sensitivity and specificity of a test.

Receiver operating characteristic (ROC) curves are used to depict the trade-off between TPR, or sensitivity, and FPR, or 1 – specificity *(14)*. Unlike the limited information provided by a single estimate of sensitivity and specificity for one possible cutoff point between normal and abnormal, ROC curves are more useful, because they depict the complete range of all possible TPR/FPR trade-offs corresponding to all pos-

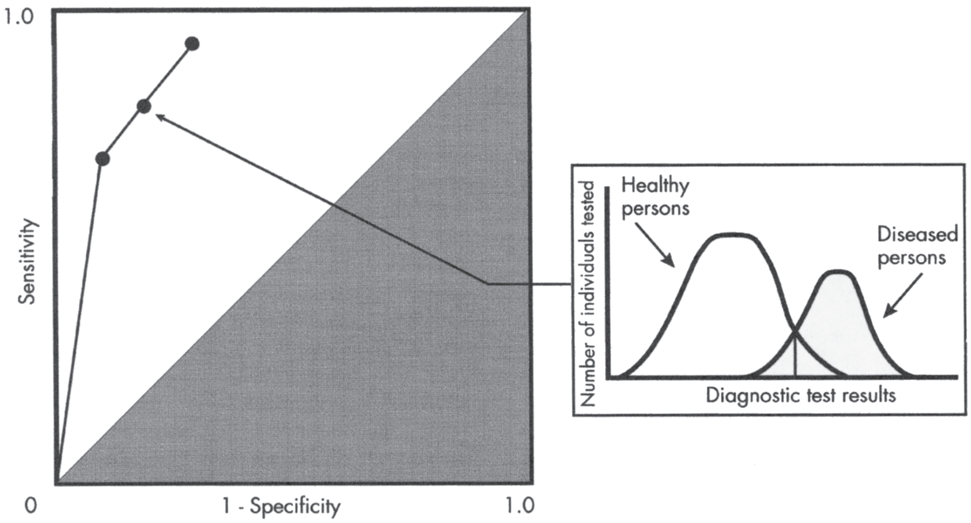

Fig. 2. Receiver operating characteristics curve depicting the trade-off between sensitivity and 1 – specificity. (From ref. *14a.*)

sible cutoffs between normal and abnormal. Derived from signal detection theory, ROC curves plot sensitivity on the vertical axis and 1 – specificity on the horizontal axis (Fig. 2) *(15)*. For all points along the 45-degree line, sensitivity equals 1 – specificity. Points on this line have no impact on the probability of disease. The probability of disease increases for points above the 45-degree line and decreases for points below the line. The points on the curve are calculated as sensitivity / (1 – specificity). Each point on the curve represents a different selected cutoff point between normal and abnormal. The perfect curve would extend straight from the origin to the upper left-hand corner and then over to the upper right-hand corner, maximizing the area under the curve (AUC) *(16)*. The AUC is considered an index of diagnostic performance *(17)*. If two tests are being compared statistically, the test with the greater AUC is considered the better test *(18–23)*. A perfect diagnostic test has an AUC of 1.0. Some authors consider the AUC concept, and ROC curve analysis in general, not very useful because prevalence is not incorporated *(2)*.

ROC curve analysis can also be used to identify the appropriate cutoff point between normal and abnormal *(24,25)*. Clinicians generally use the “upper limit of normal” provided by the laboratory. Sox et al. *(24)* suggest that the ROC curve method is better but its use is severely limited by the time needed to perform the analysis.

THRESHOLDS FOR TREATMENT

Once a decision has been made regarding whether a test is normal or abnormal, the next issue to be dealt with is whether sufficient testing has been completed to make a diagnosis. If no further testing is required, treatment can begin. The goal here is to determine at what point the acquisition of additional information would have no effect on the diagnosis. Major determining factors in this decision are the probability of disease and the penalty for being wrong. If the probability of disease is high and the margin for error is wide, the willingness to make a diagnosis will also be high. However, if the probability of disease is low and the margin for error is slim, the need for more information will be high and the willingness to make a diagnosis will be low.

A term commonly used to describe the dividing line between a decision to treat or not to treat is the treatment threshold. The treatment threshold, p^*, is the probability of disease at which the clinician is indifferent between treating and withholding treatment *(24)*. If the probability of disease for a given patient is above the treatment threshold, treatment will be selected because the acquisition of more information will not change the diagnosis. If the probability of disease is below the treatment threshold, treatment will be withheld and more testing may be ordered (or no action may be taken).

The treatment threshold can be depicted graphically with the probability of disease on the horizontal axis and the expected utility of the treatment on the vertical axis (Fig. 3). Utility is defined here as the value or the level of well-being an individual assigns to a given option. The treatment threshold is calculated by solving for p^* in the following equation:

$$\frac{p*}{1-p*} = \frac{U\left[D-T-\right]-U\left[D-T+\right]}{U\left[D+T+\right]-U\left[D+T-\right]}$$

where U = utility, D- = absence of disease, D+ = presence of disease, T- = withholding treatment, T+ = providing treatment, $U\,[D-T–]$ = the utility of withholding treatment in the absence of disease, $U\,[D-T+]$ = the utility of providing treatment in the absence of disease, $U\,[D+ T+]$ = the utility of providing treatment in the presence of disease, and $U\,[D + T–]$ = the utility of withholding treatment in the presence of disease. The line defined by the points A, C, and E represents the utility of withholding treatment irrespective of whether disease is present or absent. The line defined by the points B, C, and D represents the utility of providing treatment irrespective of whether disease is present or absent. The point of intersection between these two lines is the treatment threshold. At this point, the utilities of the two choices are equal.

The above equation can also be rephrased in terms of costs and benefits, still solving for the treatment threshold, p^*. The difference in utility between treating and not treating patients without disease can be considered a cost, C, because no benefit derives from treating these patients. Similarly, the difference between treating and not treating patients with disease can be considered a benefit, B. The previous equation would then be simplified to $p^* = C / (C + B)$ *(24)*.

THRESHOLDS FOR TESTING

Up to this point, only the threshold between treating and withholding treatment has been discussed. There are two other thresholds: the no-treatment–test threshold and the treatment–test threshold *(26)*. The no-treatment–test threshold, p_1, is the probability of disease at which there is indifference between no treatment and further diagnostic testing. The treatment–test threshold, p_2, is the probability of disease at which there is indifference between treatment and further diagnostic testing. A third line can be plotted on Fig. 3 to depict these testing thresholds (Fig. 4). The testing thresholds are calculated as follows:

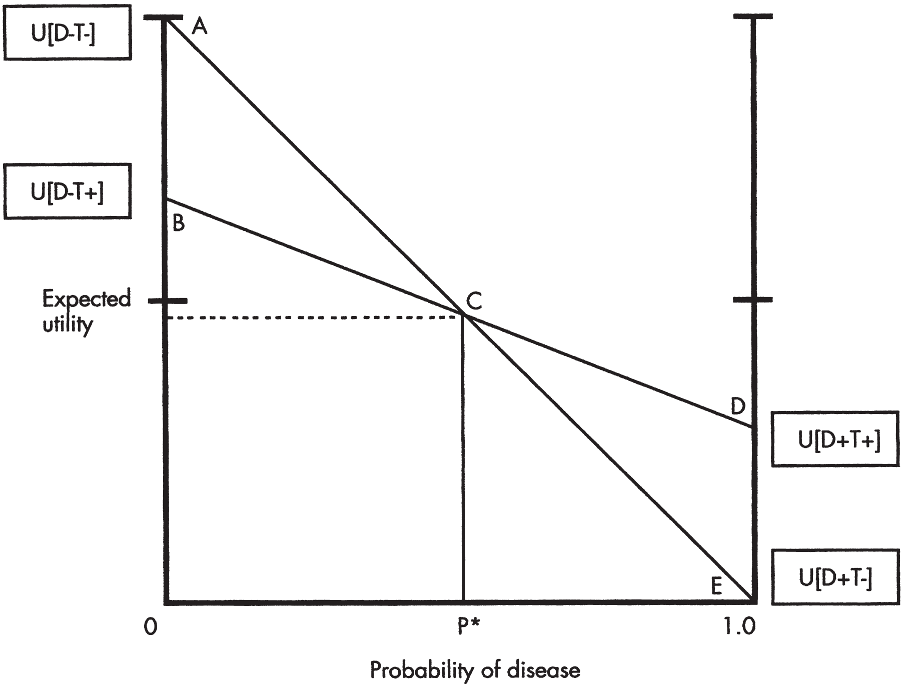

Fig. 3. Treatment threshold or the point of intersection between providing treatment and withholding treatment, whether disease is present or absent. (From ref. *14a.*)

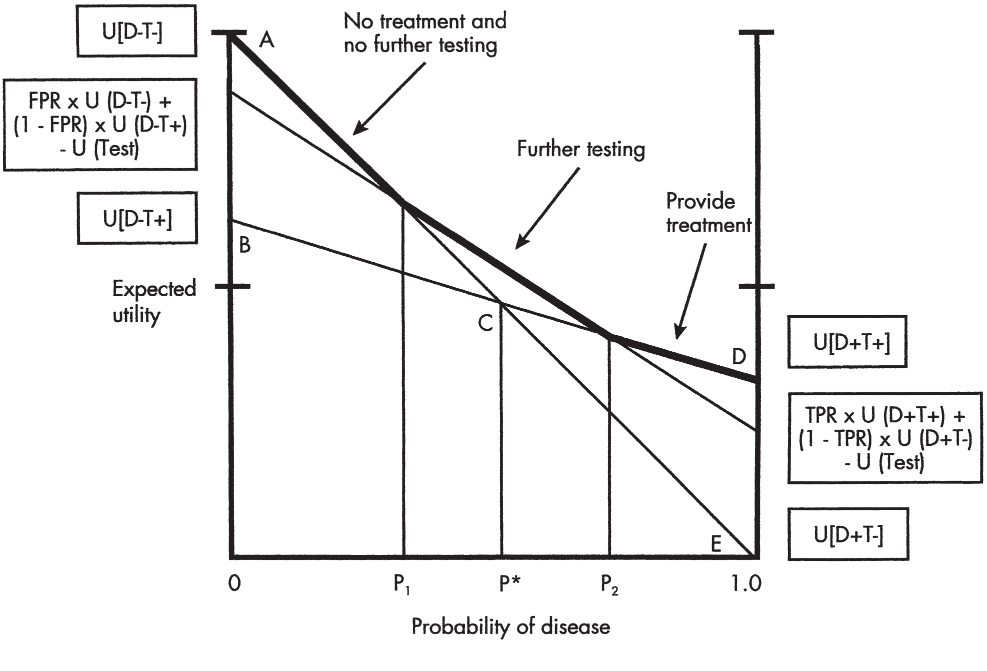
U[D-T-]
A
No treatment and no further testing
FPR x U (D-T-) + (1 - FPR) x U (D-T+) - U (Test)
Further testing
U[D-T+]
B
Provide treatment
Expected utility
C
D
U[D+T+]
TPR x U (D+T+) + (1 - TPR) x U (D+T-) - U (Test)
E
U[D+T-]
0
P₁
P*
P₂
1.0
Probability of disease

Fig. 4. Depiction of the no-treatment–test threshold, p_1, and the treatment–test threshold, p_2. (From ref. *14a.*)

$$p_1 = \frac{p^* \times \text{FPR} - p^* \times \left(U\left[Test\right] / \left(U\left[D-T-\right] - U\left[D-T+\right]\right)\right)}{p^* \times \text{FPR} + \left(1 - p^*\right) \times \text{TPR}}$$

$$p_2 = \frac{p^* \times \text{TNR} + p^* \times \left(U\left[Test\right] / \left(U\left[D-T-\right] - U\left[D-T+\right]\right)\right)}{p^* \times \text{TNR} + \left(1 - p^*\right) \times \text{FNR}}$$

where FPR = the false positive rate or 1 – specificity, FNR = the false-negative rate or 1 –sensitivity, TNR = the true-negative rate or specificity, TPR = the true-positive rate or sensitivity, and *U* [Test] represents the net utility of the diagnostic test as determined by the patient's assessment of the test regarding such factors as cost, potential side effects, the unpleasantness of the test itself, and any reassurance having the test performed provides to the patient *(24)*. (The remaining variables have already been defined.)

Below p_1, treatment is never preferred because the information to be gained from additional testing would not increase the probability of disease sufficiently to cross the treatment threshold. Above p_2 treatment is always preferred because the information gained from additional testing would not decrease the probability of disease sufficiently to cross the treatment threshold. Only for the range of disease probabilities between p_1 and p_2 could an abnormal test result have enough of an influence on disease probability to cross the treatment threshold and change patient management.

The same analysis developed in the discussion of treatment thresholds and expanded in the discussion of testing thresholds can be expanded still further to permit choosing among two or more diagnostic tests or selecting combinations of diagnostic tests. For a more in-depth discussion of these topics, refer to Sox et al. *(24)*. For an application of threshold analysis to cases where a single diagnostic test provides information about more than one event, see Nease et al. *(27)*. For an adaptation of Bayes' nomogram to threshold analysis, see Glasziou *(28)*.

ECONOMIC PRINCIPLES OF FOLLOW-UP EVALUATION

Overview of Cost Analysis

Clinicians, health administrators, and decision makers in general are constantly faced with questions regarding how to appropriate limited resources to cover what seem to be an ever-growing number of health needs. For what illnesses should every patient be automatically screened? How much follow-up is sufficient after primary treatment of a condition? If personnel dollars are short, where should cuts be made and what tradeoffs should decision makers be willing to make? The need to clarify the decision-making process and promote efficiency are the reasons economic evaluation (also known as efficiency evaluation) methodologies were developed. The three most widely used methods for assessing the relative merit of alternative courses of action are cost-effectiveness analysis (CEA), cost–benefit analysis (CBA), and utility analysis *(29,30)*. This section of the chapter provides an overview of these methods, followed by a discussion of concepts common to all three. Each of these methods is then discussed in greater detail in subsequent sections, restricting the discussion to differences across methods.

Briefly, CEA places priorities on alternative expenditures without requiring that the dollar value of life and health be assessed *(31)*. Some benefits are measured in nonmonetary units. CEA is usually the method of choice unless there is no single quantifiable unit by which alternatives can be compared, in which case CBA is the method of choice.

CBA requires the valuation of all outcomes in economic terms, including lives or years of life and morbidity *(32–34)*. This analysis, which is often viewed as a subset of CEA, assumes a goal of economic efficiency. Economic efficiency is defined as providing each unit of output at minimum possible cost *(35)*. In CBA, total costs minus total benefits equals net benefit.

Utility analysis or cost–utility analysis is very similar to CEA and is often treated as a special type of CEA. The main difference between the two is that benefits must be converted into quality-adjusted life years (QALYs) for utility analysis. QALYs is a measure of years of life gained from a procedure or intervention that is then weighted to reflect the quality of life in that year *(36)*.

Subjectivity is an important issue in any type of economic evaluation. Different analysts performing basically the same analysis can easily reach very different conclusions. This can be confusing to the novice. However, the variation in conclusions is tied directly to differences in the assumptions made in the design of the evaluation. Different conclusions do not imply that one analysis is correct and the other incorrect. They just imply that different assumptions were made.

Selection of Perspective

A first step in any economic evaluation is the selection of the perspective from which the analysis will be performed. Such analyses are generally performed from a societal perspective, but other, narrower perspectives may often apply, such as the provider's perspective, the payor's perspective, or the patient's perspective. Which perspective is selected guides the identification of costs and benefits. For example, an insurance company will evaluate a new health prevention from a payor's perspective with respect to the change in total future costs, whereas a societal perspective would assign some inherent value to illness prevented.

Specification of the Problem

The second step in any economic evaluation is specification of the problem, objective(s), and alternatives. This would seem to be rather obvious at first glance. However, if the problem is not well delineated, the range of alternatives selected to address the problem may be too narrow, ignoring important alternatives. For example, if, in the case of patients undergoing implantation of prosthetic devices, the problem is defined as the suffering undergone by current patients, only treatment alternatives will be considered. However, if the problem is defined as affecting both current and future patients, options such as delaying or preventing the onset of device malfunction or infection will also be considered in the analysis.

Production Function

The third step in economic evaluation is describing the production function. In economics, the production function is the relationship between the output of a good or service and the inputs required to produce it. The goal of this step is to specify the resources that would be utilized under each of the alternatives, the way in which the resources would be combined, and the expected result. Completion of this step allows the analyst to begin calculating costs and benefits. Complexity is not synonymous with accuracy, though models often become quite complex rather quickly.

Warner and Luce *(29)* mention six issues that must be considered in the development of the production model. First, economies of scale may exist that cause fewer and fewer inputs to be required as sample size increases to produce the same level of output per person. For example, an intervention that cost $25,000 for 1000 patients may cost only $35,000 for 5000 patients. Second, if technological change is occurring or is expected, this must be built into the model. If, while projecting future costs of follow-up after implantation of prosthetic devices, using currently available diagnostic tests, preliminary results are published of a new follow-up methodology that may replace one or more of the existing tests, the effect of substituting this test must be factored into the analysis. Third, market characteristics may affect the inputs required to produce a given output, causing the required inputs to vary by such factors as geographic location. For example, geographic differences in rates of pay for health care personnel or variation in the supply of personnel may cause an intervention to be more expensive in one city than in another. Fourth, different populations may respond differently to the same intervention, specifically in terms of compliance. More costly follow-up mechanisms may then be required to achieve the desired effect. Fifth, efficiency cannot always be assumed. The fact that a task has always been done one way does not mean that it is the most efficient. Sixth, some inputs are unique to a particular facility. If attempting to model an intervention at a new facility after an existing one at a different facility, one must examine carefully all inputs to ensure that these inputs are available or can be made available at the new facility.

Once the production function is specified, costs and benefits can be calculated. Costs can be direct, indirect, or intangible. Direct costs are defined as variable costs plus fixed (overhead) costs. In the health sector, the terms direct medical costs and direct nonmedical costs are often used. Direct medical costs are the costs directly related to the provision of care and usually involve monetary transactions, such as physicians' fees, nurses' salaries, drug purchases, equipment purchases, and independent laboratory processing fees. Direct nonmedical costs are costs incurred in the process of seeking care, such as the patient's costs of transportation to the hospital or clinic, parking costs, hotel costs if the patient cannot return home each evening because of distance, the costs of special equipment to modify one's home to accommodate a disabled family member, and child-care costs *(37)*.

Indirect costs are defined as the costs of foregone opportunities. These include the costs of morbidity and mortality. The indirect costs of morbidity are typically measured as time lost from work and the resulting wages foregone or production losses. In addition, morbidity would include the costs associated with an increased risk of complications. Similarly, the indirect costs of mortality can also be measured as time lost from work because premature death causes permanent removal from the workforce. Intangible costs are defined as the psychological costs of illness such as pain, suffering, and grief. These are the most difficult costs to measure.

Benefits can also be divided into the three categories of direct, indirect, or intangible. Benefits are often phrased as savings in costs. Direct benefits are tangible savings in health resource utilization, such as decreased length of stay or diagnostic test utilization. Indirect benefits are earnings not lost owing to avoidance of premature death or disability. Intangible benefits include pain, discomfort, and grief averted not only by the patient, but by family and friends as well. Depending on the perspective from which the analysis is performed, an item may be a cost in one analysis but a benefit (i.e., a cost averted) in the next.

The next three sections describe in greater depth each of the three types of economic evaluation. CEA is presented first.

Cost-Effectiveness Analysis

To understand CEA, one first needs to understand the term *cost-effective*. For an intervention to be cost-effective, it must be worth the money required to conduct it. Cost-effective is not always synonymous with the terms inexpensive or technically efficient, although the term is often used in this fashion. Cost-effectiveness is based on the concept of opportunity cost. The real cost of an intervention or treatment is the value of the alternative uses of the same resources *(29)*.

The goal in CEA is to determine which alternative intervention or treatment yields the greatest benefits for the lowest cost. There is no requirement that costs and benefits be measured in the same units. Some benefits are measured in nonmonetary units, such as years of life saved or disability days avoided. Indirect economic benefits are generally ignored. Unlike CBA, CEA does not allow a comparison of interventions or treatments with different outcome measures. In addition, it does not generate sufficient results to determine what dollar value per year of life saved is an acceptable level of investment.

The steps in CEA are as follows:

1. Define the problem and the objective(s) to be attained.
2. Identify alternative solutions.
3. Identify the costs of solving the problem under each alternative and all relevant benefits.
4. Compare the alternatives on the basis of prespecified criteria and select the best alternative.

Although CEA is considered a simpler approach than CBA because benefits do not need to be expressed in monetary terms, this does not mean that CEA is without its share of methodological problems. The first difficulty arises if there is more than one benefit and different units of measure apply to each benefit. The benefits are not additive and, therefore, must be analyzed separately. The next problem is how to interpret the results if the separate analyses produce contrary results. A third difficulty in CEA arises when costs and benefits accrue over a period longer than 1 year. Both costs and benefits would need to be discounted to present value. This can easily be achieved for costs, as explained in the CBA section. The problem arises when discounting benefits, because these are not measured in monetary terms.

Cost–Benefit Analysis

Cost–Benefit Analysis vs Cost-Effectiveness Analysis

CBA was previously considered a superior analysis to CEA because of its simplicity in valuing all costs and benefits in dollars. CBA is now considered inferior to CEA by some researchers because it ignores the noneconomic aspects of a program or intervention.

The steps in CBA are as follows:

1. Define the problem and the objective(s) to be attained.
2. Identify alternative solutions.
3. Identify the costs of solving the problem under each alternative and all relevant benefits.
4. Assign monetary values to the costs and benefits.
5. Discount future streams of costs and benefits to net present value if costs and benefits accrue over a period longer than 1 year.
6. Compare total present values.
7. Interpret the results *(38)*.

The first three steps are identical to those in CEA. However, the assignment of monetary values to all costs and benefits represents a major difference between CBA and CEA. Discussed in more detail in a separate section, CBA requires that a dollar value be assigned to life years saved. This issue has been quite controversial over the years.

The next step is to discount all future streams of costs and benefits to their net present value if costs and benefits accrue over a period longer than 1 year. Discounting is particularly important if one of the alternatives being compared has future costs and benefits and the other does not. The concept of discounting derives from the fact that time makes a difference. A dollar received today is worth more than a dollar to be received next year. This premise is true because a dollar received today could be invested and earn interest so that, by next year, it would be worth \$1.05, assuming a 5% interest rate. On the other hand, a dollar to be received next year is only worth \$0.95 today, assuming the same rate of interest. The equation for calculating the present value is:

$$PV = FV / \left(1 + r\right)^{t}$$

where PV = present value, FV = costs or benefits to be incurred in the future, r = discount rate, and t = the number of years into the future when the costs or benefits are expected to be incurred. The selection of the appropriate discount rate should be a function of the rate of inflation, the perspective of the analysis, and the political process as a means of reflecting social values *(39)*.

The last two steps in the analysis are to compare present values and interpret the results. A benefit–cost ratio can be calculated as the present value of total benefits divided by the present value of total costs. In comparing two interventions, the intervention with the highest benefit–cost ratio would be considered as returning greater benefit per dollar of cost. The difference between the present value of total benefits and the present value of total costs (the net benefit) is another measure commonly used to compare interventions.

Valuation of Life

A controversial issue that often comes up in CBA is how to estimate the value of human life in dollar terms *(40,41)*. This is an extremely difficult task. There are a number of suggested methodologies in the literature, with no single method considered the most correct. The two major approaches for valuing life are the willingness-to-pay approach and the human capital approach. Factors to consider in valuing life include income potential, age, quality of life, number of dependents, productivity, personal preference (religion), and personal habits.

The willingness-to-pay approach is based on how much one is willing to pay to avoid sacrificing lives. The two methods of valuing lives under the willingness-to-pay approach are the questionnaire method and the risk premium method. The questionnaire method is self-explanatory in that it entails surveying individuals to determine their willingness to pay. The problem with the questionnaire method is that respondents have little incentive to answer truthfully. However, it is easy to obtain data in the format required for analysis because the analyst has control over design of the instrument.

In contrast to the questionnaire method in which individuals are surveyed regarding their willingness to pay, the risk premium method entails observing actual behavior. For example, if people work in riskier jobs, do they really get paid more and what does that say about how they value life? Often people assume high-risk jobs because they have few job opportunities elsewhere. Other examples of high-risk behavior include smoking, drinking alcohol, and eating hazardous foods. Although activities that definitely increase the risk of death are generally intolerable to individuals, activities that may or may not increase the risk of death are apparently not *(42).*

An alternative method for valuing life is the human capital approach, which is productivity-based and ignores the costs associated with the pain and suffering avoided by averting illness and prolonging life. The term *human capital* refers to the fact that individuals, like capital equipment, can be expected to yield productive activity over their lifetimes that can be valued at their wage rate *(43)*. There are three methods of valuing human life under the human capital approach: discounted future earnings, discounted consumption, and discounted net production. The discounted future earnings method involves discounting to present value all earnings that would be realized as a result of the prolongation of life or the avoidance of a disabling illness. The advantages of the discounted earnings method are that it is reasonably objective and easy to compute. The discounted consumption method calculates a person's value of life by estimating a person's lifetime consumption of goods and services and discounting it back, resulting in a conservative estimate of the value of life. The discounted net production method, a method often used in malpractice suits, combines discounted consumption with discounted earnings. The problem with this method is that the result may be a negative number, because the present value of an individual's future consumption may be more than the present value of an individual's future earnings. Of all the human capital approach methodologies, the discounted net production method clearly results in the lowest estimate of the value of life.

Utility Analysis

Utility Analysis vs CEA

Utility analysis or cost–utility analysis is very similar to CEA and is often treated as a special type of CEA. Relevant benefits include final outcomes, such as years of life saved or days of disability averted. The difference is that in utility analysis these benefits must be converted into QALYs or, as some have suggested, healthy-years equivalents *(44,45)*. Therefore, some benefits that would be included in a CEA, such as cases found or patients correctly treated, cannot be considered in a utility analysis, because they cannot be converted into QALYs. The difference between years of life saved and

QALYs saved relates to the quality of life of the patient whose life was saved. To have one's life saved and be in a wheelchair should be valued much differently than to have one's life saved and be healthy. The results of utility analyses are usually expressed as cost per QALY gained.

There are several circumstances in which utility analysis has particular applicability. These circumstances are first, if quality of life is an important outcome; second, if both morbidity and mortality are affected by the intervention and the preference is for a single outcome combining both effects; and, third, if multiple alternative programs are being compared with a wide variety of outcome measures, the use of utility analysis would simplify the evaluation by converting all outcomes to one unit of measure *(30).*

Utility Values for Health States

The most time-consuming task in a utility analysis is determining utility values for health states. Utility is broadly defined in economics as the value an individual assigns to a given option. In health care it is generally defined as the level of well-being experienced in a given health state. Although one could estimate or possibly obtain these values from the literature, the best way to determine utility values for health states is to measure them directly *(46).* There are different schools of thought on what populations should be used to measure utilities. One approach is to identify a population with the condition of interest and measure the population's utility for the condition. The analyst needs to keep in mind that patients have a tendency to exaggerate the disutility of their condition. The second approach is to identify a population without the condition, provide a scenario of what the life of a patient with the condition is like, and measure the population's utility for the condition. The methodological difficulty with the second approach is determining how much detail to provide, what media to use to describe the condition, and how to describe the condition without biasing the result. It is suggested that the level of detail be kept to a minimum and that a balanced presentation of the condition be provided, showing both positive and negative implications of the condition.

Utilities are generally measured on a scale from 0 to 1, with 1 representing healthy and 0 representing dead. Health states often viewed as worse than death, such as dementia and coma, are assigned negative values *(47).* The three methods currently in use for measuring utility values are the rating scale, standard gamble, and time tradeoff *(48).* The rating scale method is normally depicted as a line on a page segmented into gradations by multiples of 10. A single chronic health state or multiple chronic health states and a single age of onset of illness are described to the individual whose utility for the various health states is being measured. A state of perfect health and a state of death are also described to the individual as points of reference. The individual is asked to select from among the various health states the most preferred and least preferred, which become the ends of the scale. The individual is then asked to locate the chronic health states relative to each other on the scale.

The standard gamble method is generally displayed as two circles, one representing a chronic health state and the other representing the gamble as a pie graph *(49–51).* The individual is given a choice between two alternatives. The first alternative is a definite probability of living in a particular chronic health state for life. The second alternative, or the gamble, depicts the individual returning to normal health and living for an additional number of years or dying immediately. For the gamble, the probability of perfect health is initially set at 100% and the probability of death is set at 0%. After the individual makes a choice, the probabilities are changed to 100% probability of death and 0% probability of perfect health, the pie graph is changed accordingly, and the question is posed again. This process continues until the individual is indifferent between living in the chronic health state for life and the gamble. This indifference point is the individual's utility for the health state.

In the time trade-off method, the patient is given a choice between living in a given health state for a given period followed by death vs being healthy for a shorter period of time followed by death *(52)*. After the individual makes a choice, the times are varied and the process repeats continuously until an indifference point is reached. That indifference point is the individual's utility for the health state.

Converting Utilities to QALYs

Regardless of the method used to calculate utilities, the final step in a utility analysis is to convert these utilities into QALYs and interpret the results. Assume a group's utility for living with a pacemaker averages 0.45, and a new follow-up strategy has been demonstrated to result in a 1.5-year increase in survival over the status quo. The group's QALYs would be 0.68 (0.45 × 1.5). The analyst would then calculate the cost of the follow-up strategy per QALY, discount costs and benefits to net present value if costs and benefits accrue over a period longer than 1 year, and use these data to determine if the new strategy was worth the investment *(31,53,54)*.

The use of QALYs is not without criticism, however. It has been suggested that QALYs discriminate against the elderly, equity issues are disregarded, and the resulting quality-of-life scores are biased *(44,55–59)*.

Sensitivity Analysis

No matter which method of cost analysis is chosen, certain assumptions need to be made in relation to causation. These assumptions must be carefully delineated. In addition, the analyst should determine how sensitive the results of the cost analysis are to the assumptions made. For example, if there is known imprecision in any of the estimates used, both conservative as well as liberal alternative estimates should be constructed and the sensitivity of the results to the varying estimates should be tested.

There are three major forms of sensitivity analysis: simple sensitivity analysis, extreme scenarios, and probabilistic sensitivity analysis. Simple sensitivity analysis involves varying one or more of the assumptions on which the economic evaluation is based to determine the effect on the results. The extreme-scenarios approach consists of analyzing the extremes of the distribution of costs and effectiveness and determining whether the results hold up under the most optimistic and pessimistic assumptions. Probabilistic sensitivity analysis assigns ranges and distributions to variables, using computer programs to select values at random from each range and measure the effects. This approach can handle a large number of variables and basically generates confidence intervals for each option *(60)*. Irrespective of the type of sensitivity analysis conducted, the goal is to measure whether large variations in the assumptions result in significant variations in the results of the cost evaluation. If significant variations are not the result, more confidence can be placed in the study's results. If significant variations are the result, an attempt should be made to either reduce uncertainty or improve the accuracy of crucial variables *(30)*.

SUMMARY

This chapter provides a review of the tools needed to assess and compare the performance of diagnostic tests, to determine thresholds for diagnostic testing and treating, and to calculate the total costs of follow-up after implantation of prosthetic devices. These tools allow the clinician to gather data, thus permitting more informed decision making regarding the composition of the chosen strategy. The concepts presented here lay the groundwork for the next chapter, which applies cost-evaluation methodology to the management of patients with implanted prosthetic devices, assigning dollar values to the follow-up strategies suggested in subsequent chapters. This chapter also lays the groundwork for Chapter 4, which discusses clinical, legal, economic, and ethical issues that impact how decisions should be made regarding the composition of follow-up strategies.

REFERENCES

1. Mausner JS, Kramer S. Epidemiology—an introductory text (2nd ed.). Philadelphia, PA: W.B. Saunders, 1985.
2. Fisher LD, van Belle G. Biostatistics: a methodology for the health sciences. New York: John Wiley, 1993.
3. Benish WA. Graphic and tabular expressions of Bayes' theorem. Med Decis Making 1987;7:104–106.
4. Fagan TJ. Nomogram for Bayes' formula. N Engl J Med 1975;293:257.
5. Matchar DB, Simel DL, Geweke JF, Feussner JR. A Bayesian method for evaluating medical test operating characteristics when some patients' conditions fail to be diagnosed by the reference standard. Med Decis Making 1990;10:102–111.
6. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. Am J Epidemiol 1994;140:759–769.
7. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology: the essentials (2nd ed.). Baltimore, MD: Williams & Wilkins, 1988.
8. Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. N Engl J Med 1966;274:1171–1173.
9. Friedman GD. Primer of epidemiology (2nd ed.). New York: McGraw-Hill Book, 1980.
10. Woolson RF. Statistical methods for the analysis of biomedical data. New York: John Wiley, 1987.
11. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45:1143–1154.
12. Wilson JM, Jungner F. Principles and practice of screening for disease. Geneva: World Health Organization, Public Health Papers; 1968: No. 34.
13. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia, PA: W.B. Saunders, 1985.
14. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–298.

14a. Sox HC, Blatt MA, Higgins HC, Marton KI. Medical decision making. Boston, MA: Butterworths, 1988.

15. Swets JA. Signal detection and recognition by human observers. New York: John Wiley, 1964.
16. Centor RM. Signal detectability: the use of ROC curves and their analysis. Med Decis Making 1991;11:102–106.
17. van der Schouw YT, Straatman H, Verbeek AL. ROC curves and the areas under them for dichotomized tests: empirical findings for logistically and normally distributed diagnostic test results. Med Decis Making 1994;14:374–381.

18. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
19. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–843.
20. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psych 1980;22:218–243.
21. Moise A, Clement B, Ducimetiere P, Bourassa MG. Comparison of receiver operating curves derived from the same population: a bootstrapping approach. Comput Biomed Res 1985;18:125–131.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1985; 44:837–845.
23. McClish DK. Comparing the areas under more than two independent ROC curves. Med Decis Making 1987;7:149–155.
24. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Boston, MA: Butterworths, 1988.
25. Pasanen PA, Eskelinen M, Partanen K, Pikkarainen P, Penttila I, Alhava E. Receiver operating characteristic (ROC) curve analysis of the tumour markers CEA, CA 50, and CA 242 in pancreatic cancer: results from a prospective study. Br J Cancer 1993;67:852–855.
26. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302:1109–1117.
27. Nease RF, Owens DK, Sox HC. Threshold analysis using diagnostic tests with multiple results. Med Decis Making 1989;9:91–103.
28. Glasziou P. Threshold analysis via the Bayes' nomogram. Med Decis Making 1991;11:61–62.
29. Warner KE, Luce BR. Cost–benefit and cost-effectiveness analysis in health care. Ann Arbor, MI: Health Administration Press, 1982.
30. Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.
31. Weinstein MC. Foundations of cost-effectiveness analysis for health and medical practices. N Engl J Med 1977;296:716–721.
32. Dasgupta AK, Pearce DW. Cost–benefit analysis: theory and practice. London: Macmillan, 1972.
33. Mishan EJ. Cost–benefit analysis. London: George Allen and Unwin, 1975.
34. Sugden R, Williams AH. The principles of practical cost–benefit analysis. Oxford: Oxford University Press, 1979.
35. Eastaugh SR. Medical economics and health finance. Dover, MA: Auburn House, 1981.
36. Robinson R. Cost-utility analysis. BMJ 1993;307:859–862.
37. Strom BL. Pharmacoepidemiology. New York: Churchill Livingstone, 1989.
38. Rapoport J, Robertson RL, Stuart B. Understanding health economics. Rockville, MD:, 1982.
39. Krahn M, Gafni A. Discounting in the economic evaluation of health care interventions. Med Care 1993;31:403–418.
40. Muller A, Reutzel TJ. Willingness to pay for reduction in fatality risk: an exploratory survey. Am J Public Health 1984;74:808–812.
41. Zeckhauser R. Procedures for valuing lives. Public Policy 1975;23:419–464.
42. Jacobs P. The economics of health and medical care. Rockville, MD: Aspen, 1987.
43. Robinson R. Cost–benefit analysis. BMJ 1993;307(b):924–926.
44. Mehrez A, Gafni A. Quality adjusted life years and healthy year equivalents. Med Decis Making 1989;9:142–149.
45. Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years. Med Decis Making 1993;13:287–292.
46. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chron Dis 1978;31:697–704.

47. Patrick DL, Starks HE, Cain KC, Uhlmann, RF, Pearlman RA. Measuring preferences for health states worse than death. Med Decis Making 1994;14:9–18.
48. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. Socio Econ Plan Sci 1976;10:129–136.
49. von Neumann J, Morgenstern O. Theory of games and economic behavior. New York: John Wiley, 1953.
50. Sonnenberg FA. U-maker 1.0 [computer program]. Microcomputer utility assessment program. New Brunswick, NJ, 1993.
51. Gafni A. The standard gamble method: what is being measured and how it is interpreted. Health Serv Res 1994;29:207–224.
52. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programmes. Health Serv Res 1972;7:118–133.
53. Weinstein MC. Principles of cost-effective resource allocation in health care organizations. Int J Technol Assess Health Care 1990;6:93–103.
54. Johannesson M, Pliskin JS, Weinstein MC. A note on QALYs, time trade-off, and discounting. Med Decis Making 1994;14:188–193.
55. Carr-Hill R. Assumptions of the QALY procedure. Soc Sci Med 1989;29:469–477.
56. Carr-Hill R. Allocating resources to health care: is the QALY a technical solution to a political problem? Int J Health Serv 1991;21:351–363.
57. Loomes G, McKenzie L. The use of QALYs in health care decision making. Soc Sci Med 1989;28:299–308.
58. Wagstaff A. QALYs and the equity–efficiency trade-off. J Health Econ 1991;10:21–41.
59. Johannesson M, Pliskin JS, Weinstein MC. Are healthy-years equivalents an improvement over quality-adjusted life years. Med Decis Making 1993;13:281–286.
60. Robinson R. Cost-effectiveness analysis. BMJ 1993;307(c):793–795.