The New England
Journal of Medicine

Volume 293 JULY 31, 1975 Number 5

PRIMER ON CERTAIN ELEMENTS OF MEDICAL DECISION MAKING

BARBARA J. MCNEIL, M.D., PH.D., EMMETT KEELER, PH.D.,
AND S. JAMES ADELSTEIN, M.D., PH.D.

Abstract The value of a diagnostic test lies in its ability to detect patients with disease (its sensitivity) and to exclude patients without disease (its specificity). For tests with binary outcomes, these measures are fixed. For tests with a continuous scale of values, various cutoff points can be selected to adjust the sensitivity and specificity of the test to conform with the physician's goals. Principles of statistical decision theory and information theory suggest technics for objectively determining these cutoff points, depending upon whether the physician is concerned with health costs, with financial costs, or with the information content of the test. (N Engl J Med 293:211-215, 1975)

WHEN a doctor orders a test, he has, on the basis of his knowledge and experience, a certain impression of its reliability. Does the test have many falsely positive or negative results? Moreover, whatever the result, will the findings play a determining part in shaping the doctor's decision, or will they affect his diagnosis only in a minor and complementary way? The answer to such questions need not depend merely on impression. A number of critical methods are available to evaluate diagnostic (or, for that matter, therapeutic) procedures. In addition, critical evaluation is necessary so that use of given diagnostic procedures can be justified in these days of limited resources and spiralling costs for medical care.

This primer describes three methods to achieve such critical evaluation. Though the methods go by names foreign to most physicians, their basic principles are relatively simple. In essence, they consider the ability of a diagnostic procedure to detect patients with disease while simultaneously excluding patients without disease. They also take into account goals of the physician requesting the test — for example, is he concerned primarily with health or financial cost-benefit relations, or is he concerned only with the amount of diagnostic information contained in the test? The technical terms used for the three methods to be described are the decision matrix, the receiver operating characteristic (ROC) curve and information theory. Once a diagnostic procedure has been evaluated by one of these technics, simple algebraic manipulations can be performed so that the result of a test can be applied to a particular patient; a formula called Bayes's theorem is used for this purpose.

From the Department of Radiology, Harvard Medical School and Peter Bent Brigham Hospital, and the Center for the Analysis of Health Practices, Harvard School of Public Health (address reprint requests to Dr. McNeil at the Department of Radiology, Harvard Medical School, 25 Shattuck St., Boston, MA 02115).

Supported in part by grants (GM 02201 and GM 18674) from the U.S. Public Health Service and by a grant from the Robert Wood Johnson Foundation through the Center for the Analysis of Health Practices.

DECISION MATRIX

By use of a decision matrix we can logically relate the results of a diagnostic test to the clinical or pathologic outcome. This type of analysis is most easily applied to the simple decision of whether disease is present, D+, or absent, D−, when the test is abnormal (i.e., positive), T+, or normal (i.e., negative), T−. When, as shown in Table 1, these two binary results are plotted on a two-×-two table to show the four possible combinations (indicated by a, b, c and d), a decision matrix is formed.

Each of the four combinations can be used to evaluate the test by comparing its results to the actual presence or absence of disease (i.e., four ratios may be formed). The so-called true-positive (TP) ratio is the proportion of positive tests in all patients who actually have the disease, or $\frac{a}{a+b}$. This value expresses probability (P) that patients with the disease will have abnormal test results, and can be written as the "conditional probability" P(T+|D+)* — i.e., the probability that a patient with disease, D+, will have a positive test, T+. The true-positive ratio expresses the *sensitivity* of the examination. It measures the fraction of patients with disease that will be detected by the diagnostic test in question.

The false-positive (FP) ratio is the proportion of positive tests in all patients who do not have disease, or $\frac{c}{c+d}$. It is the probability that patients without disease will have abnormal test results, P(T+|D−).

The true-negative (TN) ratio is the proportion of negative tests in all patients who do not have the disease, or $\frac{d}{c+d}$. It is the probability that patients without disease

*A "conditional probability" is written, as a matter of convention, with a vertical bar before the given state or condition that is present or absent. It does not imply division.

Table 1. A General Decision Matrix.

Test Results	Presence of Disease		
	Present (D+)	Absent (D−)	Totals
Abnormal (T+)	a	c	a + c
Normal (T−)	b	d	b + d
Totals	a + b	c + d	

Table 2. Correlation of Liver Scan Data with Pathological Outcome.

Scan		Liver Disease Present (D+)	No Disease (D−)	Totals
Abnormal	T+	231	32	263
Normal	T−	27	54	81
Totals		258	86	344

will have negative test results, P(T−|D−). This ratio expresses the *specificity* of the examination. It measures the fraction of patients who will be correctly identified as having no disease. It is equal to (1 − FP ratio).

The false-negative (FN) ratio is the proportion of negative tests in all patients with disease, or $\frac{b}{a+b}$. It is the probability that patients with disease will have negative test results, P(T−|D+). It is equal to (1 − TP ratio).

Obviously, a good diagnostic examination has a high TP ratio and a low FP ratio; it correctly identifies a large portion of diseased patients without incorrectly including patients without disease. The ratio of the TP ratio to the FP ratio is known as the likelihood ratio, L. Obviously, tests with high likelihood ratios are better discriminators of disease than those with low ones.

These test characteristics may be illustrated with a specific example. In a study on the use of liver scans for detecting disease in 344 patients, the actual state of the liver was determined either by biopsy or at autopsy.[1] When the actual numbers as determined by the scans and by the morphologic examinations are put into the decision matrix, the following table emerges (Table 2).

We may calculate the characteristics of the liver scan to be as follows:

$$\text{True-positive ratio} = P(T+|D+) = \frac{231}{231+27} = 0.90.$$

$$\text{False-positive ratio} = P(T+|D-) = \frac{32}{32+54} = 0.37.$$

$$\text{True-negative ratio} = P(T-|D-) = \frac{54}{54+32} = 0.63 = (1-0.37).$$

$$\text{False-negative ratio} = P(T-|D+) = \frac{27}{27+231} = 0.10 = (1-0.90).$$

Thus, the liver scan is 90 per cent *sensitive* and 63 per cent *specific*. It will detect 90 per cent of patients with liver disease and will correctly classify 63 per cent of those without disease.

The ROC Curve

General Characteristics

Few tests have simple binary outcomes and thus cannot be classified as just positive or negative. Instead, most yield a continuous scale of values, of which one of several can be selected as a cutoff point to differentiate subjects with and without disease. The cutoff point chosen depends on the relative costs associated with classifying patients with disease as normal versus classifying normal patients as diseased. In screening for a potentially fatal disease with a fairly safe treatment, for example, we would be likely to accept a large proportion of false-positive diagnoses to ensure that our test discovers almost all diseased patients—i.e., that it has high sensitivity. Patients with hypothyroidism fall into this category because of the high morbidity associated with failing to diagnose and treat this disease and the low morbidity associated with treating euthyroid patients with replacement therapy. For less serious conditions or more dangerous treatments, on the other hand, we are willing to miss more diseased patients to reduce the number of false-positive diagnoses.

If we select a cutoff point that makes the test very sensitive to detect as many patients with actual disease as possible, the number of false-positive diagnoses unavoidably increases; in other words, the more sensitive the examination, the less specific it becomes. To help us determine the most advantageous cutoff point, we first construct a graph plotting true-positive (TP) ratios (i.e., the expression of sensitivity) against false-positive (FP) ratios. The resulting plot, which takes the shape of a smooth, concave curve, is know as an "ROC curve" (receiver-operating-characteristic curve).

A hypothetical ROC curve was constructed (Fig. 1) with the assumption that a laboratory examination has a range

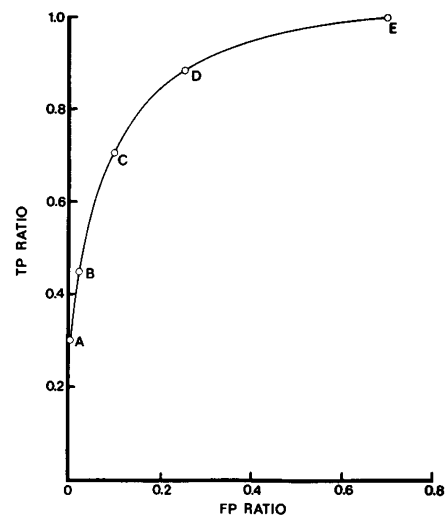

Figure 1. Hypothetical ROC Curve

The vertical scale is the TP ratio, and the horizontal scale the FP ratio. At one extreme point, A, the test has poor sensitivity (TP ratio = 0.30) but good specificity (FP ratio = 0.07). At the other extreme, E, the test has high sensitivity (TP ratio = 1) but poor specificity (FP ratio = 0.70).

of values, any one of which can be used to separate normal from diseased patients (Table 3). The proportion of subjects classified as normal or abnormal on the basis of this test then depends upon where this cutoff point is placed. When test results larger than those corresponding to the value at point A are considered abnormal, only 30 per cent of the diseased patients are detected, but there are no false-positive diagnoses. At this cutoff point the test has great specificity but very poor sensitivity. When test results greater than those corresponding to the value at point C are considered abnormal, 70 per cent of patients with disease are detected, but 10 per cent of subjects without disease have abnormal results. When a low cutoff value (point E) is used to separate those with disease from those without, all patients with disease are identified, but at the expense of including a large proportion of patients (70 per cent) without disease. The location of a cutoff point along an ROC curve is called an "operating position."

Table 3. Correlation of Cutoff Point and True-Positive and False-Positive Ratios for a Hypothetical ROC Curve.

Cutoff Point	Proportion of Patients Having Abnormal Test Results	
	Patients with disease (true-positive ratio)	Patients without disease (false-positive ratio)
A	0.30	0.00
B	0.45	0.02
C	0.70	0.10
D	0.90	0.25
E	1.00	0.70

Selection of a Cutoff Point

Selection of an appropriate cutoff point is aided by knowledge of the probability of disease in the patient population of interest. This probability of disease in any given member of the group as a whole is called the prior or pretest (that is, before the results of the test in the given member are obtained) probability and may be designated as P(D+). The prior or pretest probability of no disease may be designated as P(D−). For illustrative purposes in subsequent examples, we shall assume that the hypothetical test described in Table 3 was performed on a group of patients of whom 30 per cent have disease {P(D+) = 0.30} and 70 per cent do not {P(D−) = 0.70}.*

Selection of an appropriate cutoff point is also aided by knowledge of the costs associated with errors in diagnosis — both false-positive and false-negative errors. We are generally interested in the additional costs associated with these errors in comparison with the costs associated with an ideal or perfect test. Costs can be divided into those that pertain to health and those that pertain to money.

Health costs make use of two major indexes, mortality and morbidity, and are usually the basis of decisions involving patient management. Varying mortality patterns are most conveniently evaluated by computing and comparing, for various diagnostic tests and therapeutic regimens, an index called "person-years"; a person-year is defined as one person surviving for one year.† Person-years do not take into account the morbidity, pain and anxiety associated with the diagnostic tests and therapeutic regimens. It is possible, therefore, that the morbidity associated with the best treatment regimen (measured in terms of person-years) would be sufficiently high so that a given patient would prefer a less effective regimen. Ideally, physicians should take patients' preferences regarding mortality and morbidity into account when making decisions in their behalf. Some weighting of mortality and morbidity should probably be performed, and the resulting index called "healthy person-years of life" rather than person-years of life alone.

Financial costs have two major components. There are the medical bills themselves, paid by the patient or some third party. In addition, since death or disability may prevent patients from supporting themselves or others, there may be additional costs to society and insurers for support of patients or their dependents. Even if these support costs cannot be exactly determined, it is clear that use of medical bills alone to determine the cost of diagnosis and treatment underestimates the total financial costs.

When health costs are most important and are used to select a cutoff point between normal and abnormal results, we want to minimize differences in person-years (or, ideally, healthy person-years) between our diagnosis and treatment and that existing for perfect diagnosis and optimal treatment. If we let the additional cost in person-years associated with a false negative diagnosis be AC_{fn} and with a false-positive diagnosis be AC_{fp} we can determine from statistical decision theory‡ that the optimal operating position on the ROC curve occurs where the slope of the ROC curve equals

$$\frac{AC_{fp}\,P(D-)}{AC_{fn}\,P(D+)} \quad (1).$$

For example, on the average, if the cost of missing a diagnosis is high and the cost of mistakenly treating patients is low, intuition tells us to operate at a point near E on the ROC curve (Fig. 1) where we treat all patients with disease. The formula supports this estimate because under these conditions, the ratio AC_{fp}/AC_{fn} is small, and the slope of the curve changes only slightly near E. On the other hand, if the therapeutic results of treating a disease are of marginal value, and the health costs of treating a patient without disease are high, intuition tells us to oper-

*The ratio of the probability of disease to the probability of no disease, $\frac{P(D+)}{P(D-)}$, is known as the prior odds and is usually designated by the Greek letter omega, Ω. In this example the prior odds are 3:7. The prior probability of disease can be calculated from the prior odds by the following relation[5]: $P(D+) = \frac{\Omega}{1+\Omega}$.

†This definition implies that one year of life for two individuals is the same as two years of life for one person and that all years of life are valued equally.

‡Derivation of expression (1) can be found in *Signal Detection and Recognition by Human Observers,* edited by John A. Swets, New York, John Wiley and Sons, 1964. It is based on considerations related to expected value. Expected value is defined here as the sum, considering all potential outcomes of a decision, of the products of the probabilities of outcome, and the value attached to each outcome.

ate at a position near point A where the slope, like AC_{fp}/AC_{fn}, is steep. Finally, if the likelihood of disease in the patient is very small (that is, the ratio of the probability of no disease to disease is large), we again choose a point near A; this situation occurs in screening programs.

When financial costs are our criterion for selection of a cutoff point, the same principles apply except that now the additional costs relate to money. AC_{fp}, for example, is the extra cost associated with unnecessary diagnostic and perhaps therapeutic regimens, whereas AC_{fn} is the additional cost caused by the progression of untreated disease.

In many cases we do not have accurate estimates of the additional health or financial costs associated with errors in diagnosis. One approach to this problem is to choose a cutoff point that minimizes our mistakes. This position may be designated by M_{min} and occurs where the slope of the ROC curve equals P(D−)/P(D+). For example, for the population where P(D−) = 0.70 and P(D+) = 0.30 the position occurs where the slope equals

$$\frac{P(D-)}{P(D+)} = \frac{0.70}{0.30} = 2.33.$$

This position is near point C of the ROC curve in Figure 1.

It is important to emphasize that the cutoff point chosen is best only for the measure of costs selected. Thus, if costs are based on mortality, the resulting operating position should yield the lowest average mortality. If costs are based on finances, the lowest average financial cost results. This resulting cutoff point need not be the same as that associated with the lowest average mortality. In addition, the cutoff point that minimizes mistakes may differ from both these points.

In evaluating various therapeutic or diagnostic procedures, these health and financial costs are initially estimated independently. To relate diagnostic or treatment (health) costs to financial costs we may calculate an average financial cost required to achieve a given unit of health (e.g., person-years) for each procedure; this is an *average cost.* In comparing various therapeutic or diagnostic procedures, a cost called a *marginal cost* may be calculated. This cost is the financial cost of achieving one additional unit of health (e.g., one more person-year) by one procedure over another. For example, if one treatment costing $15,000 yields five person-years of life and another costing $8,000 yields three person-years of health, the marginal cost of each additional person-year resulting from the first treatment is $3,500 {(15,000-8,000)/(5-3)}. For diagnostic procedures the same principles apply except that now the *average* cost relates to finding a patient with disease and the *marginal* cost relates to finding an additional patient with disease using one diagnostic procedure as compared with another. Such diagnostic procedures can be either different tests or different cutoff points applied to the results of a single test.

Information Theory

General Characteristics

Information theory has been used as one of several possible means for selecting a cutoff point along the ROC curve.[2] In this context, information is defined as a reduction in uncertainty; thus, the greater the difference between the certainty of a diagnosis after a test is performed and the certainty before it is performed, the greater the information content of test. Accordingly, if the certainty about a given diagnosis is already high, little information is gained from an additional diagnostic test. For example, results of a serum ceruloplasmin level in a patient with Kayser–Fleischer rings and ataxia provide less information than results of a similar test in a brother of a child with Wilson's disease. In the former instance we are fairly certain of the diagnosis on the basis of physical findings — that is, our pretest probability estimate is already close to 1.0. In the latter instance, on the other hand, we are less certain; our pretest probability is lower, around 0.25, and we therefore gain much more information from a serum ceruloplasmin measurement. Obviously, if we think that one disease is as likely to be present as not, our pretest probability is 0.50; in this case, we gain information from a test that will help us go from a completely uncertain 50:50 state to one of greater certainty.

Selection of a Cutoff Point

A theoretical relation exists between the maximum information content obtainable from a perfect test (TP ratio = 1, FP ratio = 0) and the frequency of the disease in question; this relation is described by a smooth curve having a continuous range of values (Fig. 2). Because most tests are not perfect, however, the theoretical maximum value is seldom achieved. The actual value depends upon the TP and FP ratios. Tests that have a continuous scale of values and thus a number of possible discrete cutoff points have different amounts of information associated with each cutoff point as well as with each prior probability. For example, we can calculate* the information content for each of the cutoff points on the ROC curve presented in Figure 1, and to show that for prior probabilities of 0.25, 0.50 and 0.75, the information content at point D is higher than that for points A, B, C and E. The cutoff point of a diagnostic test having a value closest to the theoretical one is the one that maximizes the information content of the test. This point is said to have an information content of I_{max}.

Bayes's Theorem

Once a diagnostic test has been evaluated so that its characteristics (i.e., sensitivity and specificity) are known, it is possible to formulate new probability statements about the presence or absence of disease in a particular patient examined by the diagnostic test. These probability statements are called posterior or post-test probabilities because they reflect the test results. If a patient has an abnormal test result the probability of disease is written as P(D+|T+) and if he has a normal test result, it is written as

*We calculated these results by evaluating a complex algebraic expression derived and discussed in detail by Metz.[2] This expression makes use of the TP and FP ratios of the test (or, in our case, of the ratios corresponding to varying cutoff points) and of the prior probability with which we are concerned.

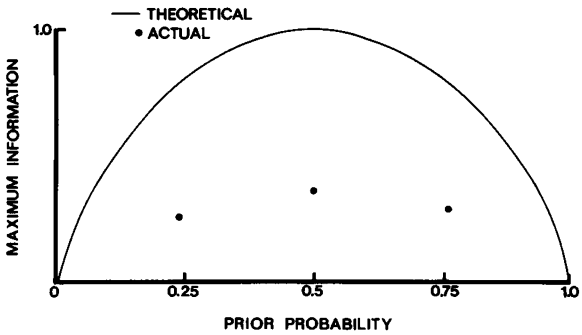

Figure 2. Information Content of a Test as a Function of the Prior Probability of Disease.

The maximum information content theoretically obtainable occurs with a perfect test (TP ratio = 1, and FP ratio = 0); its maximum value is 1.0 and occurs where the prior probability is 0.5 (solid line). The circles represent the maximum information content (I_{max}) for the hypothetical ROC curive (Fig. 1) at three prior probabilities; these three values of I_{max} are associated with point D on the ROC curve.

P(D+|T−). Bayes's* theorem is a technic that allows us to calculate these posterior probabilities that we wish to know from information that we already know beforehand ("a priori") about the implications of a diagnostic test. For example, if we wish to estimate the probability of disease in a patient with an abnormal test result we must know the probabilities that the diagnostic test will be positive in patients with and without disease — the TP and FP ratios — and an estimate of the prior probabilities, P(D+) and P(D−). The following formula is used:

$$P(D+|T+) = \frac{P(T+|D+)\ P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)} \quad (2).$$

Alternatively, if we wish to know the probability that a patient with a normal test result has disease, we need to know the TN and FN ratios as well as P(D+) and P(D−). The relevant formula is:

$$P(D+|T-) = \frac{P(T-|D+)\ P(D+)}{P(T-|D+)P(D+) + P(T-|D-)P(D-)} \quad (3).$$

*The Reverend Thomas Bayes (1702-1761) was the author of the first treatise on one type of inductive inference. He is believed to be responsible for the following statement written in 1736: "It is not the business of a Mathematician to show that a strait line or circle can be drawn, but he tells you what he means by these; and if you understand him, you may proceed further with him; and it would not be to the purpose to object that there is no such thing in nature as a true strait line or perfect circle, for this is none of his concern; he is not inquiring how things are in matter of fact, but supposing things to be in a certain way, what are the consequences to be deduced from them; and all that is to be demanded of him is, that his suppositions be intelligible, and his inferences just from the suppositions he makes."

As a specific example illustrating these formulas consider the hypothetical test (Table 3, Fig. 1) performed on a group of patients 30 per cent of whom are estimated to have disease. Let us assume that we have used point D as our cutoff point. The probability of disease in a patient with an abnormal test is calculated from equation (2) and is

$$P(D+|T+) = \frac{(0.90)\ (0.30)}{(0.90)(0.30) + (0.25)(0.70)} = 0.61.$$

Thus, the abnormal test has changed the probability of disease in a patient from 0.30 to 0.61, a factor of two. If, on the other hand, the patient has a normal test his probability of disease is calculated from equation (3) and becomes

$$P(D+|T-) = \frac{(0.10)\ (0.30)}{(0.10)(0.30) + (0.75)(0.70)} = 0.05.$$

A negative test has reduced the probability of disease from 0.30 to 0.05, a factor of six. In this context, the test is more useful in ruling out disease than in detecting it.

The difference between posterior and prior probabilities is strongly dependent upon the true-positive and false-positive ratios for the diagnostic test. A nomogram relating both prior and posterior probabilities to these ratios has been constructed for a wide range of test sensitivities.[3] For tests that are "perfectly sensitive" (TP ratio = 1.0), a family of curves relating prior to posterior probability can be constructed for varying false-positive ratios.[4]

The elementary principles enumerated in this primer have been discussed in more detail by Lusted[5] and by Barnoon and Wolfe.[6] These authors point out that these principles can be applied to a wide range of clinical problems from clinical decisions involving individual patients to matters of public-health policy. Several articles in this issue are devoted to examples of these types of analyses.

References

1. Drum DE, Christacapoulos JS: Hepatic scintigraphy in clinical decision making. J Nucl Med 13:908-915, 1972
2. Metz CE, Goodenough DJ, Rossmann K: Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. Radiology 109:297-304, 1973
3. Fagan TJ: Nomogram for Bayes's Theorem. N Engl J Med 293:257, 1975
4. Katz MA: A probability graph describing the predictive value of a highly sensitive diagnostic test. N Engl J Med 291:1115-1116, 1974
5. Lusted L: Introduction to Medical Decision Making. Springfield, Illinois, Charles C Thomas, 1968
6. Barnoon S, Wolfe H: Measuring the Effectiveness of Medical Decisions. Springfield, Illinois, Charles C Thomas, 1972