Basic Principles of ROC Analysis

Charles E. Metz

The limitations of diagnostic "accuracy" as a measure of decision performance require introduction of the concepts of the "sensitivity" and "specificity" of a diagnostic test. These measures and the related indices, "true positive fraction" and "false positive fraction," are more meaningful than "accuracy," yet do not provide a unique description of diagnostic performance because they depend on the arbitrary selection of a decision threshold. The receiver operating characteristic (ROC) curve is shown to be a simple yet complete empirical description of this decision threshold effect, indicating all possible combinations of the relative frequencies of the various kinds of correct and incorrect decisions. Practical experimental techniques for measuring ROC curves are described, and the issues of case selection and curve-fitting are discussed briefly. Possible generalizations of conventional ROC analysis to account for decision performance in complex diagnostic tasks are indicated. ROC analysis is shown to be related in a direct and natural way to cost/benefit analysis of diagnostic decision making. The concepts of "average diagnostic cost" and "average net benefit" are developed and used to identify the optimal compromise among various kinds of diagnostic error. Finally, the way in which ROC analysis can be employed to optimize diagnostic strategies is suggested.

HOW CAN WE measure the quality of diagnostic information and of diagnostic decisions in a meaningful way? That basic question has become increasingly important in recent years as an abundance of new diagnostic tests have been introduced, and as government and the public grow ever more insistent that the medical community must justify the costs and possible risks of diagnostic procedures.

The question must be addressed, for it will not go away. Any meaningful approach to the evaluation of diagnostic tests must inevitably involve many complex technical and social issues, and one cannot reasonably expect that the typical practicing physician can or should master all of the subtleties involved in evaluation analysis. Still, the basic concepts upon which evaluation analysis rests are quite straightforward and need not be regarded as mysterious. Although these concepts are (unfortunately) often clothed in seemingly occult jargon (because of the need for concise and precise terminology) the principles themselves are mostly formalized common sense, or at least can be recognized as reasonable when explained in plain language.

This monograph will attempt to guide the reader through the basic principles of an approach that provides a structure for the meaningful evaluation of diagnostic techniques. Although this approach is essentially quantitative, its merit does not depend solely on the use of numbers. The approach focuses attention on the *issues* involved in diagnostic evaluation and diagnostic decision making, and the reader will likely find that he has informally considered some or all of these issues already.

DILEMMAS IN EVALUATING DIAGNOSTIC TESTS

What Does "Accuracy" Mean?

Any assessment of diagnostic performance seems to require some comparison of diagnostic decisions with "truth." Perhaps the simplest measure of diagnostic decision quality is the fraction of cases for which the physician is correct, often called "accuracy." Although we are all willing to accept that high accuracy is good (all other things being equal—and that's the catch), the number can be very misleading. In screening for a relatively rare disease, for example, one can be very accurate simply by ignoring all evidence and calling all cases negative. If only 5% of patients have the disease in question, a physician who always blindly states that the disease is absent will be right 95% of the time!

Accuracy is of limited usefulness as an index of diagnostic performance because disease prevalence affects the resulting number strongly, and no mathematical correction for disease prevalence can redeem this index in a meaningful way. One might be tempted to suppose that though this is true, accuracy should be meaning-

From the Department of Radiology, University of Chicago, and the Franklin McLean Memorial Research Institute, Chicago, Ill.

The Franklin McLean Memorial Research Institute is operated by the University of Chicago for the U.S. Department of Energy under Contract EY-76-C-02-0069.

Reprint requests should be addressed to Charles E. Metz, Ph.D., Box 429, 950 East 59th Street, Chicago, Ill. 60637.

0001-2998/78/0804-0003$02.00/0

ful at least as an index for comparison of diagnostic techniques applied to a given population in which disease prevalence is known and fixed. However, here too the index is limited. Two diagnostic modalities can yield equal accuracies but perform differently with respect to the types of correct and incorrect decisions they provide; the incorrect diagnoses from one might be almost all false negative decisions (misses), while those from the other might be nearly all false positive decisions (false alarms), and clearly, the usefulness of these two tests for patient management could be quite different in various situations.

Though accuracy provides a single simple number for diagnostic performance, it is often too simple and must be interpreted with considerable caution. The limitations of this index force us to introduce some complexity into our evaluation scheme: We must sort out the effect of disease prevalence, and we must score separately the various kinds of right and wrong diagnostic decisions.

Sorting things out. Both of the obvious limitations of the accuracy index can be overcome by defining decision performance in terms of the pair of indices:

$$\text{Sensitivity} = \frac{[\text{Number of True Positive (TP) decisions}]}{[\text{Number of actually positive cases}]}$$

and

$$\text{Specificity} = \frac{[\text{Number of True Negative (TN) decisions}]}{[\text{Number of actually negative cases}]}$$

In effect, sensitivity and specificity represent two kinds of accuracy: the first for actually positive cases and the second for actually negative cases. One must note carefully that the terms "positive" and "negative" in these definitions concern some particular disease state, which must be specified clearly in calculating and quoting sensitivity and specificity values. For simplicity, these indices require that all possible states of health and disease be classified into two categories. These categories can be defined in any way that is convenient and meaningful for the problem at hand, but they must be made explicit. For example, patients could be classified as having one or more tumors (malignant or benign) or no tumor; as having malignant tumors or no malignant tumor (which could be benign tumor or no tumor), etc.

Accuracy, or the fraction of the study population that is decided correctly, is related to sensitivity and specificity by the simple formula:

$$\begin{aligned}\text{Accuracy} &\\ = [\text{Sensitivity}] \times &\left[\begin{array}{l}\text{Fraction of the study}\\ \text{population that is actually}\\ \text{positive}\end{array}\right]\\ + [\text{Specificity}] \times &\left[\begin{array}{l}\text{Fraction of the study}\\ \text{population that is actually}\\ \text{negative}\end{array}\right]\end{aligned}$$

The reader should think through the proof of this relationship as a simple exercise in the sort of manipulation that is used repeatedly in our approach. Notice that accuracy is defined as:

$$\text{Accuracy} = \frac{[\text{No. correct decisions}]}{[\text{No. cases}]}$$

so

$$\begin{aligned}\text{Accuracy} &= \frac{\left[\begin{array}{c}\text{No. True}\\ \text{Positive}\\ \text{decisions}\end{array}\right]}{[\text{No. cases}]} + \frac{\left[\begin{array}{c}\text{No. True}\\ \text{Negative}\\ \text{decisions}\end{array}\right]}{[\text{No. cases}]}\\ &= \frac{\left[\begin{array}{c}\text{No. True}\\ \text{Positive}\\ \text{decisions}\end{array}\right]}{\left[\begin{array}{c}\text{No. actually}\\ \text{positive}\\ \text{cases}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{No. actually}\\ \text{positive}\\ \text{cases}\end{array}\right]}{[\text{No. cases}]}\\ &+ \frac{\left[\begin{array}{c}\text{No. True}\\ \text{Negative}\\ \text{decisions}\end{array}\right]}{\left[\begin{array}{c}\text{No.}\\ \text{actually}\\ \text{negative}\\ \text{cases}\end{array}\right]} \times \frac{\left[\begin{array}{c}\text{No. actually}\\ \text{negative}\\ \text{cases}\end{array}\right]}{[\text{No. cases}]}\end{aligned}$$

Hence the relationship is proven. A little arithmetic and a little common sense go a long way in this field!

At this point we must introduce some additional terminology that is commonly used in the approach we are taking. True positive fraction (TPF) is simply the same thing as sensitivity, and true negative fraction (TNF) is simply the same as specificity. As one can see from the definitions of sensitivity and specificity, the terms TPF and TNF are more directly descriptive of the concepts involved and are a lot easier to remember. These new terms suggest two other definitions:

$$\text{False Positive fraction (FPF)} = \frac{[\text{No. False Positive decisions}]}{[\text{No. actually negative cases}]}$$

and

$$\text{False Negative fraction (FNF)} = \frac{[\text{No. False Negative decisions}]}{[\text{No. actually positive cases}]}$$

Note that FPF and FNF represent, respectively, the fractions of actually negative cases and of actually positive cases that are decided incorrectly.

If we presume that all cases are diagnosed as either positive or negative (with respect to a specified disease), then for either actual state, the number of correct decisions plus the number of incorrect decisions must equal the number of cases with that actual state. Thus it is easy to show that the various fractions defined above must be related by

$$\text{TPF} + \text{FNF} = 1$$

and

$$\text{TNF} + \text{FPF} = 1$$

(The reader should prove these relationships as an exercise.) Because of these constraints, one can always compute FNF from knowledge of TPF, for example, so it is necessary only to specify one fraction from each of the above relationships in order to fix all four types of decision fractions.

One additional set of notations must be defined before we proceed. It is common to denote the four decision fractions defined above by using the symbols of conditional probabilities, because each decision fraction represents an estimate of the probability (or relative frequency) of a particular decision, given that (or conditional on the fact that) an individual case actually has a particular health or disease state. Let D represent the disease in question, and let T represent the result of a diagnostic test, i.e., a particular decision. Then FPF, for example, is equivalent to the conditional probability $P(T+ \mid D-)$, which is read as "the probability of a positive test, given the absence of disease." Similarly, TPF is often denoted by $P(T+ \mid D+)$; FNF by $P(T- \mid D+)$; and TNF by $P(T- \mid D-)$. Note that the use of conditional probability notation makes explicit the kinds of test results (decisions), T, and actual disease states, D, that are in the numerators and denominators of the definitions of the four kinds of decision fractions. Also, this notation emphasizes that all four decision fractions are conditional on (i.e., are normalized with respect to) actual disease states.

Finally, the prevalence of disease in the population subjected to the diagnostic test (or for which diagnoses are to be made) can be represented by $P(D+)$, the prior probability of the actual presence of the disease in a case from the population studied. Similarly, $P(D-) = 1 - P(D+)$ represents the prior probability that disease is actually absent in a case from the studied population.

The relationships among the various quantities defined so far are summarized in Table 1. Note, in particular, the sense in which thinking of the conditional probabilities as fractions helps one to remember the definitions and the relationships.

Apples and oranges. The concepts defined in the previous section allow us to sort out the effects of disease prevalence and to score separately the performance of a diagnostic test

Table 1. Definitions of, and Relationships Among, the Various Decision Performance Indices Described in the Text

Definitions	Relationships
TPF = Sensitivity $= P(T+ \mid D+)$	TPF + FNF $= P(T+ \mid D+) + P(T- \mid D+) = 1$
FPF = 1 − (Specificity) $= P(T+ \mid D-)$	TNF + FPF $= P(T- \mid D-) + P(T+ \mid D-) = 1$
TNF = Specificity $= P(T- \mid D-)$	Accuracy = Sensitivity × $P(D+)$
FNF = 1 − Sensitivity $= P(T- \mid D+)$	+ Specificity × $P(D-)$
Disease Prevalence $= P(D+)$	= TPF · $P(D+)$ + TNF × $P(D-)$
	$= P(T+ \mid D+) \times P(D+) + P(T- \mid D-) \times P(D-)$

Table 2. Decision Data and Calculated Indices for Hypothetical Test A*

	Test Result (Diagnosis)		
	Positive (T+)	Negative (T−)	Total Actual States
Actual State			
Positive (D+)	140 (TP)	60 (FN)	200 actually positive
Negative (D−)	100 (FP)	900 (TN)	1000 actually negative
Total test results (diagnosis)	240 positive decisions	960 negative decisions	

*Total cases, 1200.

Calculated indices:

TPF = 140/200 = 0.70; FNF = 1 − TPF = 0.30
FPF = 100/1000 = 0.10; TNF = 1 − FPF = 0.90
P(D+) = 200/1200 = 0.17; P(D−) = 1 − P(D+) = 0.83
Accuracy = TPF × P(D+) + TNF × P(D−) = 0.87

or a diagnostic decision maker with respect to actually positive and actually negative cases.

In order to see how these concepts can be applied to a collection of diagnostic decisions, consider the following hypothetical situation. Suppose that 1200 cases from a defined population have been subjected to some diagnostic test "A," and that the actual health or disease state for each case has been determined by biopsy, follow-up, or some other means. Suppose that 200 actually positive cases were ultimately found in the population studied, and that the diagnostic test to be evaluated yielded 140 TP decisions, 60 FN decisions, 900 TN decisions, and 100 FP decisions. These data can be summarized by the "decision matrix" shown in Table 2. Note that summing across rows yields the number of cases with an actual health or disease state, while summing in a column yields the total number of times that the corresponding decision was made. Note also that the values for TNF, FNF, and accuracy obtained using the relationships summarized in Table 1 are the same as those that would be obtained directly using the definitions of these quantities.

We see from the calculated indices that this test, as used here, is more accurate for actually negative cases than for actually positive cases, since TNF is greater than TPF—even though more actually negative than actually positive cases were decided incorrectly. The latter observation is not paradoxical, but merely reflects the preponderance of actually negative cases in the population studied; recall that TPF, TNF, etc. represent rates and not numbers of cases.

The decision fractions allow us to predict how the accuracy index would change if this same test were applied (in the same way) to a population with a different prevalence of disease, $P(D+)$. We see that if the various decision fractions are kept constant but $P(D+)$ is increased to 0.6, for example, then accuracy would be $(0.7) \times (0.6) + (0.9) \times (0.4) = 0.78$. This value is lower because the test is less accurate for actually positive cases, and these have become more frequent.

Often we wish to compare diagnostic tests. Suppose that the same population of cases used to evaluate test A was studied using a different

Table 3. Decision Data and Calculated Indices for Hypothetical Case B*

	Test Result (Diagnosis)		
	Positive (T+)	Negative (T−)	Total Actual States
Actual State			
Positive (D+)	80 (TP)	120 (FN)	200 actually positive
Negative (D−)	40 (FP)	960 (TN)	1000 actually negative
Total test results (diagnosis)	120 positive decisions	1080 negative decisions	

*Total cases, 1200.

Calculated indices:

TPF = 80/200 = 0.40; FNF = 1 − TPF = 0.60
FPF = 40/1000 = 0.04; TNF = 1 − FPF = 0.96
P(D+) = 200/1200 = 0.17; P(D−) = 1 − P(D+) = 0.83
Accuracy = TPF × P(D+) + TNF × P(D−) = 0.87

test, test B, with the results shown in Table 3. Comparison of Tables 2 and 3 clearly shows that these two tests are performing very differently—though the accuracy indices are the same! Test B is performing worse than test A for actually positive cases since TPF is lower and FNF is higher, but it is performing better for actually negative cases since TNF is higher and FPF is lower. The accuracy indices are equal because this trade-off in performance is balanced by the disease prevalence, $P(D+)$, that we have used in our example. It should be clear that in many applied situations, tests A and B (as used here) are not of equal value; if the implications of a false positive decision for subsequent patient management are bad and overriding, then test A is worse, while if the implications of a false negative decision are bad (and overriding), then test B is worse.

What to do? How can we balance the apples and oranges of TPF and FPF? We could at this point attempt to incorporate into our analysis "weights" for the good and bad of the various types of correct and incorrect decisions. However, first let us consider a further complication, which will suggest a solution to the present dilemma.

The implicit variable. In the use of almost any diagnostic test, test data do not necessarily fall into one of two obviously defined categories that can be uniquely ascribed to the presence or absence of the disease in question.

For diagnostic tests that yield a single number as a result (such as 24-hr thyroid uptake, various blood serum assays, etc.) the distributions of result values in actually positive and in actually negative patients overlap, and no single threshold or decision criterion can be found that separates the populations cleanly; otherwise the test would be perfect! Usually a threshold value must be chosen arbitrarily, and different choices will yield different frequencies for the various kinds of correct and incorrect decisions. For example, if high results tend to indicate the presence of disease but the distributions of test result values in actually negative and in actually positive patients overlap (as shown in Fig. 1) then increasing the threshold value will make both false positive and true positive decisions less frequent, but will make both true negative and false negative decisions more frequent. A threshold value must be selected that is believed to yield an appropriate compromise among these gains and losses.

Similarly, diagnostic tests that yield results that must be judged subjectively, such as imaging studies, usually require that some confidence threshold be established in the mind of the decision maker. If an image suggests the possibility of disease, how strong must that suspicion be in order for the image to be called positive? The confidence threshold that an observer adopts undoubtedly depends on many things: his "style," his estimate of prior odds or probability, and his assessment of the consequences of the various possible correct and incorrect decisions. The concept of confidence threshold may be hard to quantify, but in most situations a confidence threshold can be varied, and the various decision fractions will vary with it.

Recognizing the arbitrary nature of decision threshold selection might seem to complicate our problem even more. Aside from the "apples and oranges" of TPF and FPF, how can we compare tests A and B if the data in Tables 2 and 3 could be changed simply by arbitrarily selecting different decision thresholds, or by using

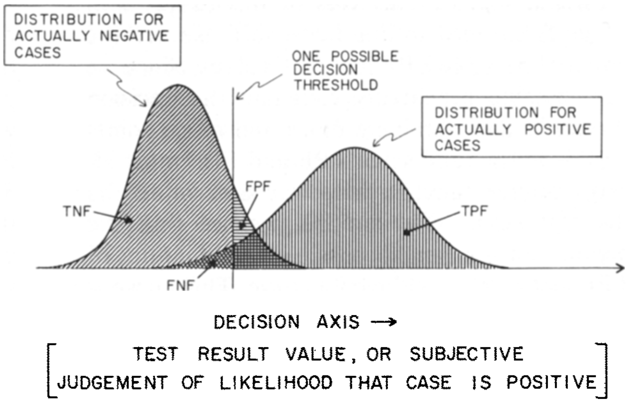
DISTRIBUTION FOR ACTUALLY NEGATIVE CASES
ONE POSSIBLE DECISION THRESHOLD
DISTRIBUTION FOR ACTUALLY POSITIVE CASES
FPF
TNF
TPF
FNF
DECISION AXIS →
TEST RESULT VALUE, OR SUBJECTIVE JUDGEMENT OF LIKELIHOOD THAT CASE IS POSITIVE

Fig. 1. Two hypothetical distributions of a quantity on which decisions are based, showing one possible decision threshold. The conditional probability of each kind of decision is equal to the area under a distribution on one side of the threshold.

a different set of considerations in making a subjective decision?

We resolve this dilemma by intentionally forcing the decision threshold to vary and by observing the resulting changes in the various decision fractions.

INSIGHT PROVIDED BY ROC ANALYSIS

Varying the Variable

If we explicitly change the decision threshold by reinterpreting the results of a quantitative test using a new threshold of abnormality or by having the observer reread a set of images requiring that he be more, or less, certain that a case is positive before calling that case positive, then we will obtain a different set of decision fractions. If we change the decision threshold again to a new level, we will obtain yet again a different set of decision fractions. Since TPF and FPF together determine all four decision fractions (Table 1), we need only keep track of how these two fractions change as the decision threshold is varied.

If we imagine that the distributions of test results (or, for subjective tests, the distributions of some quantity like "estimate of the likelihood of disease, given the test information") are of the form shown in Fig. 1, then we see that lowering the decision threshold, for example, must increase both the TPF and FPF. After some thought, one should realize that whatever the form of the distributions, TPF and FPF must increase or decrease together as the decision threshold is changed.

If we explicitly change the decision threshold several times as described above, we will obtain several different pairs of TPF and FPF. These pairs can be plotted as the "y" and "x" coordinate values of points on a graph such as that shown in Fig. 2. The axes of this graph both range from zero to one because these are the limits of possible TPF and FPF values. Since we can imagine repeatedly changing the decision threshold and obtaining more and more points on this graph, and since TPF and FPF must always change together in a way determined by the test result distributions, we see that the points representing all possible combinations of TPF and FPF must lie on a curve. This curve is called the receiver operating characteristic (ROC) curve for the diagnostic test, since it describes the inherent detection *characteristics* of the test (or, for subjective studies, the observer–test combination) and since the *receiver* of the test information can *operate* at any point on the curve by using an appropriate decision threshold. Fig. 2 shows three possible operating points that might correspond to use of a strict threshold (case called positive only if judged almost definitely positive); of a moderate threshold; or of a lax threshold (case called positive if any suspicion of disease).

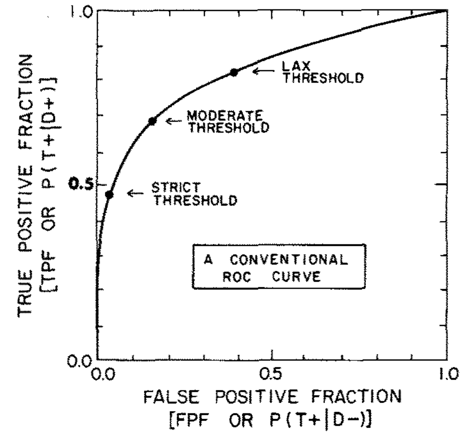

Fig. 2. A typical conventional ROC curve, showing three possible operating points.

Conventional ROC curves of the kind described here (in which two actual states are possible and in which two decision alternatives are available) inevitably must pass through the lower left corner (FPF = 0, TPF = 0) of the graph because all tests can be called negative, and through the upper right corner (FPF = 1, TPF = 1) of the graph because all tests can be called positive. Also, if the test provides information to the decision maker, the intermediate points on a conventional ROC curve must be above the major (lower left to upper right) diagonal of the ROC space, because in that situation a positive decision should be more probable when a case is actually positive than when a case is actually negative, i.e., $P(T+ \mid D+)$ should be greater than $P(T+ \mid D-)$. Finally, one can show on theoretical grounds that if the decision maker uses available information in a proper way, the slope of the ROC curve must steadily decrease (i.e., it must become less steep) as one moves up and to the right on the curve.

What the Curve Means

Essentially, a conventional ROC curve describes the compromises that can be made between TPF and FPF—and hence among the relative frequencies of true positive, false positive, true negative, and false negative decisions—as a decision threshold is varied. By appropriate choice of the decision threshold, a decision maker or observer can operate at (or near) any desired compromise that lies on the curve. Since the ROC curve is a graph of TPF versus FPF, both of which are independent of disease prevalence, it does not depend on the prevalence of disease in the actual population to which the test may be applied.* Thus, ROC analysis provides a description of disease detectability that is independent from both disease prevalence and decision threshold effects.

We will discuss later the issue of optimal choice of an operating point on an ROC curve, but a few comments seem appropriate here. If disease prevalence is very low, then false positive fraction (FPF) must be kept small; otherwise almost all positive decisions will be false positive decisions, and these diagnoses will burden the health care system and patients with many unnecessary follow-up examinations and/or treatments. Also, if consequences of a false positive decision are overridingly bad, perhaps because high-risk surgery would then be done unnecessarily, FPF must again be kept small. In either or both situations, the decision maker should operate on the lower left part of the ROC curve to keep FPF small, even at the expense of a low TPF and correspondingly high FNF. Conversely, if the same test with the same ROC curve is applied to a population in which disease prevalence is high and/or in which the need for finding actually positive cases is of overriding importance, then the decision maker should adjust his decision threshold to operate higher on the curve, accepting a higher FPF in order to keep TPF high and FNF low. The ROC curve shows the extent to which FPF must be increased, for example, in order to increase TPF to any required level.

For diagnostic tests in which the test result must be judged subjectively, an ROC curve describes the decision performance of an observer–test combination. Clearly, disease detectability can be poor if the test provides little information, or if the observer is not skilled in interpreting the information provided, or both. Because it gives an empirical description of decision performance, ROC analysis of subjective diagnostic tests cannot reveal whether the technology or the individual human is performing badly. However, ROC analysis of the decision performance of several individuals using a single diagnostic test can indicate the extent to which usefulness of the test depends on individual skill and/or experience.[1] A more subtle issue related to performance of the decision maker, as opposed to the test, concerns his ability to hold his decision threshold fixed. Variations in use of the decision threshold from decision to decision cause decision performance to be degraded, with a consequent effect on the measured ROC curve.[2] This effect of threshold inconsistency on the measured ROC curve is appropriate and desirable, because any aspect of decision-making behavior that degrades decision performance should be included in an empirical analysis of the observer–test combination.

Dilemmas Resolved

We can now resolve the dilemmas that we faced in attempting to compare the hypothetical tests A and B on the basis of the decision performance data shown in Tables 2 and 3. From the perspective of ROC analysis, the combination of TPF and FPF obtained there for each test merely represents one point on the ROC curve for each test. By varying the decision threshold for one test, we could change the combination of TPF and FPF in such a way that the TPFs for both tests are made equal, allowing comparison of the two resulting FPFs, or we could make the FPFs for both tests equal, permitting comparison of the two TPFs. More directly, we could measure the two curves and compare the curves themselves.

*The curve may depend on the spectrum of disease states classified as actually positive, however. If early disease is harder to detect than advanced disease, for example, then the ROC curve will depend on the mixture of early and advanced actually positive cases studied. Thus, cases in the actually positive component of a study population must be chosen so as to represent the population at large to which the conclusions of the study will be applied. Similarly, the actually negative component should appropriately reflect the relative frequency of normal variants.

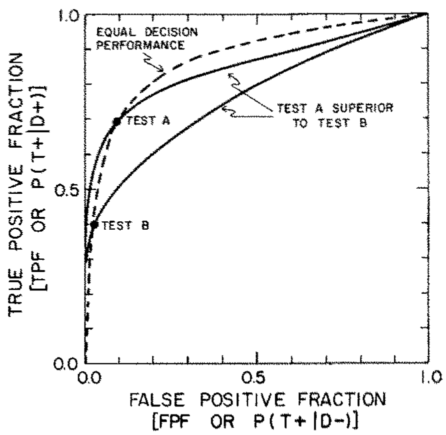

Fig. 3. The decision fractions resulting from the data of Figs. 1 and 2 plotted as points in an ROC space, with possible ROC curves on which these points could lie.

Figure 3 shows an ROC space which plots two points corresponding to the two combinations of TPF and FPF found for tests A and B on the basis of the data given in Tables 2 and 3. If we were to measure ROC curves for the two tests by (consistently) changing the two decision thresholds, the ROC curves might turn out to be those shown by the solid lines. If these curves were found, we could conclude that test A offers greater detectability of the disease in question than does test B, because for any given FPF, the TPF provided by test A is greater, while for any given TPF, the FPF provided by test A is less.

Alternatively, we might find that the two ROC curves are (essentially) the same, such as the dotted curve shown in Fig. 3. In that case we would conclude that the two tests provide equal detectability of the disease in question, because the tests can be made to perform identically by choosing the two decision thresholds appropriately.

In general, we may conclude that better decision or detection performance is indicated by an ROC curve that is higher and to the left in the ROC space. It is conceivable (though not common) that two ROC curves may cross (and possibly recross). In such a case, the relative quality of decision performance provided by the two tests in question must be judged in the context of the diagnostic situation to which they will be applied, taking into account disease prevalence and the costs and benefits of the consequences of the various types of decisions, as we describe later.

PRACTICAL CONSIDERATIONS

The Rating Method Trick

As we have seen, an ROC curve can be generated by varying the decision threshold that defines the cut point between results ascribed to (though not necessarily due to) actually positive and actually negative cases.

Data from a diagnostic test that yields a single quantitative value for each case can easily be rescored as positive or negative by using various decision thresholds. A number of points on the corresponding ROC curve can be plotted in this way, and a smooth curve can be drawn through or fitted statistically to the points.

However, this approach is often impractical for diagnostic tests that must be interpreted subjectively because human observers may not find it possible to associate a continuum of numerical values with their subjective impressions of certainty. The simplest way of expressing a diagnostic decision is in terms of positive or negative, even though that decision may have been reached by comparison of a subjective impression with a decision threshold. These binary (two-valued: yes or no) decisions cannot be reanalyzed to determine what the decision maker would have said if he had used a different confidence threshold, however. Thus, an ROC curve can be generated from subjective yes–no response data only by requiring the decision maker to reread the entire set of cases several times, using a different decision threshold each time. This repeated yes–no approach is clearly burdensome and usually impractical.

A practical technique for generating response data that can be used to plot an ROC curve in such a subjective judgment situation is called the "rating method" and was developed in experimental psychology.[3] Essentially, the method represents a compromise between accepting a yes–no response and requiring that the decision maker select a value from a continuous scale to represent his confidence that the case in question is positive. Instead, the observer or decision maker is required to select one of several ratings or categories of confidence to represent his judgment, based on the information provided by the diagnostic test (and perhaps on other supplementary information available to him). These categories can be given qualitative labels such as: (1) definitely or almost definitely negative; (2) probably negative; (3) possibly positive; (4) probably positive; and (5) definitely or almost definitely positive. The use of five categories seems to represent a reasonable compromise between the needs of ROC analysis and the precision with which an observer can be expected to reproduce his ratings. We show below that use of N categories will yield $(N - 1)$ nontrivial points on the ROC curve.

The rating data obtained in this way are used to compute points on the ROC curve as follows: First, only those responses in the category corresponding to highest certainty that a case is positive are scored as positive decisions, and the rest are scored as negative decisions. Thus, for the category labels listed above, only responses in category 5 would be scored as positive "decisions" at this stage of data analysis. These decisions are then compared with the actual presence or absence of disease for each case, and TPF and FPF are calculated. This combination of TPF and FPF is plotted as a point in the ROC space and can be interpreted as the ROC curve operating point corresponding to use of a strict decision threshold, with which a case is called positive if and only if the decision maker is certain, or almost certain, that the case in question is actually positive.

Next, the rating scale response data are rescored, this time interpreting as a positive decision a response in either of the two categories corresponding to greatest certainty that a case is actually positive. Thus, for the labels listed above, a response in either category 5 or category 4 is scored as a positive decision. The resulting values for TPF and FPF are then calculated and plotted in the ROC space. This point represents an ROC curve operating point corresponding to the use of a less strict decision threshold, that is, corresponding to the situation in which the decision maker would call a case positive if he judges that the case is at least probably positive.

This procedure is then repeated, successively interpreting as a positive decision a rating in any of the three categories of highest certainty that a case is positive (here, categories 5 or 4 or 3 are considered positive); then a rating in any of the highest four categories, etc. When finally any response is scored as a positive decision, both TPF and FPF become equal to 1.0, so the last plotted operating point is always in the upper right corner of the ROC graph. A smooth curve is then drawn through or fitted statistically to the plotted points to yield the measured ROC curve.

Curve fitting. The rating method yields several points in the ROC space that represent experimental estimates of operating points on a single ROC curve. Because the number of cases that can be included in any ROC experiment is limited by practical considerations, and because human decisions are not always reproducible, each plotted point is subject to statistical error.

Standard deviations of the variations that can be expected in any one plotted operating point (if the experiment were repeated using a different set of the same number of cases) can be estimated by the expressions:[4]

Standard Deviation of TPF*

$$= \sqrt{\frac{\text{TPF} \times (1 - \text{TPF})}{(\text{No. actually positive cases}) - 1}}$$

and

Standard Deviation of FPF

$$= \sqrt{\frac{\text{FPF} \times (1 - \text{FPF})}{(\text{No. actually negative cases}) - 1}}$$

These expressions can be used to plot plus and minus one or two standard deviation error bars vertically and horizontally around the experimental points in the ROC space in order to provide a visual impression of the reliability of the points.[4] Note that (1) the standard deviations depend on the position of a point in the ROC space, being largest when TPF or FPF is close to 0.5; that (2) the standard deviation of TPF is inversely related to the number of actually positive cases used in the experiment; and that (3) the standard deviation of FPF is inversely related to the number of actually negative cases used. Since precision of TPF and FPF are usually equally important, it is customary to attempt to use roughly equal numbers of actually positive and actually negative cases in an ROC experiment. These estimates of ROC

*The denominators inside the square roots are of the form $(N - 1)$ here rather than N to yield unbiased estimates of variance. In practice, this is usually a minor issue.

point reliability can be used as a guide in drawing a smooth curve that passes appropriately through or near the plotted points. Often a smooth curve fitted subjectively by eye provides an adequate estimate of the full ROC curve.

If a more objective curve-fitting procedure is desired, some assumption must be made regarding the functional form of the curve to be fit to the data. A commonly used assumption in experimental psychology is that the ROC curve is of the same functional form as would be generated from two Gaussian or "normal" probability distributions centered at different positions on the decision axis, and with possibly different standard deviations, as shown in Fig. 1. Each decision is assumed to be made by comparing the decision variable outcome (position on the horizontal axis) with some decision threshold and deciding positive if the threshold is exceeded. Although the applicability of this underlying theoretical model cannot be proven even for idealized experimental situations, various theoretical arguments can be made in its behalf. The literature of experimental psychology contains much empirical evidence that curves of the functional form predicted by this model provide good fits to ROC data from experiments in which decisions are based on subjective judgments.

The ROC curves predicted by this theoretical model depend on two parameters: the distance between the centers of the two normal distributions on the decision axis (expressed in units of the standard deviation of one of the distributions) and the ratio of the standard deviations of the two distributions. Various combinations of these two parameters yield different ROC curves, and one combination can usually be found that fits experimental ROC data quite well. Conveniently, the ROC curves are predicted by this theoretical model graph as straight lines if they are plotted on a pair of transformed coordinate axes that are linear not with respect to TPF and FPF, but instead with respect to the standard deviates corresponding to the TPF and FPF values.* Graph paper with these transformed double probability coordinate scales is available,† and can be used to plot the ROC data points in such a way that a straight line can be fit to the points. The slope and one axis intercept of this fitted straight line then correspond to the two parameters of the underlying theoretical model, and these can be used to summarize the detectability of disease described by the ROC data.[5]

If an objective statistical curve-fitting procedure is desired, conventional least-squares fitting of a straight line on a double-probability graph is not appropriate because the assumptions implicit to conventional least-squares methods (equal variance vertically, no variance horizontally) are not valid for ROC data. Instead, a special maximum likelihood curve-fitting computer program should be used, which finds the pair of model parameters that make the observed ROC data most likely (i.e., least unlikely). Different programs are available for ROC data generated in yes–no experiments[6] or in rating-method experiments.[7]

No fully satisfactory statistical technique has been developed as yet to test the significance of apparent differences between measured ROC curves. Perhaps the best method available at the present time is the testing of differences between areas under the curves fitted by the maximum-likelihood procedure described above. A promising approach not yet fully explored involves testing the apparent differences between pairs of curve parameters by means of the multivariate analysis of variance.

Cases, Truth, and Common Sense

A fundamental aspect of almost any objective approach to the evaluation of diagnostic decision making (whether in terms of accuracy, sensitivity and specificity, or ROC analysis) is the need for a sufficient number of cases in which the actual state of health or disease has been determined. Diagnostic truth must be known in order to score the quality of each decision, and enough cases must be used to ensure acceptable statistical precision in the measured performance indices. Although these require-

*Consider a normal distribution with standard deviation equal to 1.0, centered on Z = 0. The transformed coordinates mentioned above represent the values of Z such that the areas under this distribution to the left of Z correspond to TPF and FPF, respectively.

†"Double Integrated Normal Chart," available as item Y4 231 from the Codex Book Co., P. O. Box 366, Norwood, Mass. 02062.

ments are sometimes tedious to satisfy in clinical situations, ROC analysis is no more demanding in this regard than other objective methods of evaluation analysis. In short, the quality of diagnostic decisions cannot be determined if the correct answers are not known.

The problem of establishing "truth" is straightforward in evaluation studies that use artificial test samples or phantom images, but this problem can be exceedingly tedious and frustrating in studies employing actual clinical cases. The definition of truth is ultimately a philosophical issue, of course, and operational standards for diagnostic truth must be established for the purposes of evaluation analysis; these must take into account the goals of the evaluation study, potential sources of bias, and common sense. In short, standards of truth need not be perfect, but they must be considerably more reliable than the tests to be evaluated; judgments of truth should be independent from information provided by the tests to be evaluated, and one must balance thoughtful reflection on the potential errors and difficulties of such evaluation studies against the useful (even if limited) information that they can provide.

In the selection of cases to be included in an evaluation study, due consideration must be given to include an appropriate spectrum of disease characteristics in the sample case population, because the conclusions drawn from the study are applicable only to, and cannot be defined more specifically than, the sample population.[8]

The various issues that should be considered in designing a study for the evaluation of diagnostic medical imaging procedures are discussed in a general protocol currently in the final stages of preparation.*

No simple answer exists to the question of how many cases are necessary for meaningful conclusions to be drawn from an ROC analysis of decision performance, but several issues should be considered. First, no matter what means may be used to infer the significance of apparent differences between ROC curves, the required precision of measured ROC points will depend on the magnitude of the differences that actually exist: More cases are needed to demonstrate subtle differences in diagnostic performance than gross differences.

Second, statistical variations in ROC data and fitted ROC curves are due to at least two factors: the extent to which the limited number of cases used in an ROC experiment represents the total population of such cases at large, and the extent to which diagnostic test results and subjective diagnostic judgments are reproducible. Although the cumulative effects of these two sources of variation can be expressed in terms of binomial or multinomial statistics and can be estimated by the expressions for standard deviations quoted above, the relative magnitude of the individual effects has not been studied and their interaction is not understood. However, the fact that both of these two effects do occur unquestionably complicates the issue of interpreting apparent differences between measured ROC curves. Because of these two sources of statistical variation, an observed difference between the decision performance of two diagnostic tests acting on the same sample population may in fact be more significant than an assumption of sample independence would suggest, because if the limited case sample is atypically difficult for one test, it may be atypically difficult for the other also; in this situation the ROC curves for the two tests should vary up and down together if they are applied to different sample populations of the same limited size. Thus, error bars computed on the basis of the independent sample assumption may be unduly pessimistic concerning the significance of differences between curves in this situation.

Because no generally accepted statistical test yet exists for demonstrating the quantitative statistical significance of apparent differences between ROC curves, the number of cases required to achieve significance cannot be predicted. This state of affairs is certainly unsatisfactory, and current theoretical efforts hold promise for better statistical techniques in the future. Meanwhile, common sense and experience suggest that meaningful qualitative conclusions can be drawn from ROC experiments performed with as few as about 100 clinical cases[1] or experimental images.[9]

*This document is currently in preparation by Bolt Beranek and Newman, Inc., Cambridge, Mass. under National Cancer Institute Contract NO1-CB-64010 ("Standard Protocol for Evaluation of Imaging Techniques in Cancer Diagnosis"; John A. Swets, Principal Investigator).

GENERALIZED ROC METHODS

The conventional ROC methods that we have described apply to situations in which actual states of health and disease are grouped into two categories and in which two decision alternatives are available to the decision maker. In this section we sketch how these methods can be generalized to apply to more complicated decision-making situations.

The most fundamental property of the ROC approach is that it describes the trade-offs that are available among the conditional frequencies of various types of correct and incorrect decisions. By viewing the approach in this broad way, we can see that a generalized ROC approach would account for the ways in which the frequencies of certain types of decisions must vary with the frequencies of other types of decisions as one or more decision threshold is changed.

Consider first the situation in which the decision maker must not only call an actually positive case positive, but must also state where the case is positive in order to receive credit for a fully true positive decision. If localization of disease to within the proper image quadrant is required, for example, and if disease can be present in at most one quadrant, then five actual states and decision alternatives are available: "no disease," "disease in upper left quadrant," etc. We have shown theoretically and experimentally[10,11] that decision performance in this more complex task can be predicted from knowledge of the conventional ROC curve measured for the two-alternative detection-only task, and that the resulting generalized ROC curve is a curved line in three-dimensinal space, which can be plotted as two curves on a two-dimensional graph.

Another situation of interest is that in which more than one lesion may be present, and in which the observer must, in effect, count the lesions. We have shown that if the possible lesions are similar, decision performance in this multiple-signal task can again be predicted from knowledge of the conventional ROC curve (measured when zero or one lesion may be present), and that the generalized ROC curve is a curved line in multi-dimensional space, which can be plotted as a set of two-dimensional graphs.[12]

These two studies have shown that decision performance in some multialterative tasks employing medical images can be related uniquely and predictably to decision performance in a simple two-alternative task that is measured by a conventional ROC curve. Thus, in these situations the conventional ROC curve provides a sufficient conceptual and experimental description of decision performance.

A common aspect of the tasks used in these two studies is that the decision maker can be assumed to base his selection of one of several decision alternatives on the repeated comparison of a single kind of judgment against a single decision threshold. In the multiple-signal detection task, for example, he is assumed to try to detect lesions in various parts of an image by repeating a similar judgmental process and then adding up the number of lesions that he believes he has found.

An appropriate theoretical model for what we might call a simultaneous detection and differential diagnosis task is less clear.[9] For example, suppose that the decision maker is confronted with a population of cases, each one of which may be actually negative, positive with disease A, or positive with disease B. No fully general multialternative ROC approach is yet available to measure and describe decision performance in this task. An approach that may suffice at present is the measurement of three conventional ROC curves, either by grouping the actual states into two alternatives in the three possible ways, or by deleting cases with one actual state in each of three decision experiments.

Theoretical and experimental efforts to deal with this important situation within the context of ROC analysis are continuing.

IMPLICATIONS FOR MEDICAL DECISION MAKING

In performing a diagnostic study, one pays a price (in terms of money and the risk of possible complications) to gain information that should be of benefit in subsequent patient management. In this section we use the perspective of ROC analysis to address three questions: "How can one balance the benefits of correct diagnostic decisions against the costs of incorrect decisions?" "How can one judge whether the information purchased is worth the price paid?"

and "How can one evaluate combinations of diagnostic tests?"

We admit at the outset that the conceptual approach we sketch here is overly simplistic in the sense that some complicated issues are ignored or dismissed casually, and in the sense that the quantitative data needed to implement calculations suggested by the approach may be hard to obtain. Still, real diagnostic situations do exist that can be effectively modeled in the terms outlined here, and meaningful estimates of the required data can be made. The fundamental contribution of this approach, we believe, lies in the extent to which it forces us to focus our attention on the relevant issues in a systematic way. The approach can provide numbers, but more important, it can provide insight.

Cost/Benefit Analysis

Consider how we might formulate the average cost of the consequences of performing a diagnostic test. The term cost can be interpreted in a narrow sense of money or healthy patient days lost, or it can be thought of as combinations of various components.[13]

The average cost of the consequences of performing a diagnostic test must include, first, the (average) price we must pay to do the test; we can call this the "overhead cost" and denote it by C_0. To this we must add the costs of the medical consequences of each type of diagnostic decision, and since we want an average cost, we must weigh each of these decision consequence costs by the probability that the type of decision in question occurs. Thus, for the two-alternative decision situation we can express the average cost ($\bar{C}$) resulting from the use of the diagnostic test as:

$$\begin{aligned}\bar{C} = C_0 &+ C_{TP} \times P(\mathrm{TP}) + C_{TN} \times P(\mathrm{TN}) \\ &+ C_{FP} \times P(\mathrm{FP}) + C_{FN} \times P(\mathrm{FN})\end{aligned}$$

where, for example, P(TP) is the probability that a true positive decision is made, and C_{TP} represents the average cost of the medical consequences of a true positive decision.* This kind of expression can be extended easily to include multialternative decision situations, but we consider the two-alternative case here for simplicity.

The probability of a true positive decision, P(TP), is equal to the probability that a case from the population studied is actually positive, $P(D+)$, multiplied by the probability that an actually positive case will be diagnosed as positive using the test in question, $P(T+ \mid D+)$. Thus we can replace P(TP) by $P(D+)P(T+ \mid D+)$. Similarly, $P(\mathrm{TN}) = P(D-)P(T- \mid D-)$; $P(\mathrm{FP}) = P(D-)P(T+ \mid D-)$; and $P(\mathrm{FN}) = P(D+)P(T- \mid D+)$. Further, from the relationships listed in Table 1, we know that $P(T- \mid D+) = 1 - P(T+ \mid D+)$ and that $P(\mathrm{T}- \mid D-) = 1 - P(T+ \mid D-)$. Thus, the expression above can be rewritten as:

$$\begin{aligned}\bar{C} = C_0 &+ C_{TP} \times P(D+) \times P(T+ \mid D+) \\ &+ C_{TN} \times P(D-) \times [1 - P(T+ \mid D-)] \\ &+ C_{FP} \times P(D-) \times P(T+ \mid D-) \\ &+ C_{FN} \times P(D+) \times [1 - P(T+ \mid D+)]\end{aligned}$$

or after some rearrangement of terms:

$$\begin{aligned}\bar{C} = &-\{[C_{FN} - C_{TP}] \times P(D+)\} \times P(T+ \mid D+) \\ &+ \{[C_{FP} - C_{TN}] \times P(D-)\} \times P(T+ \mid D-) \\ &+ \{C_0 + C_{TN} \times P(D-) + C_{FN} \times P(D+)\}\end{aligned}$$

Inspection of this expression reveals several fundamental issues. First, whatever the average decision consequence costs may be, the average diagnostic cost ($\bar{C}$) increases or decreases with the overhead cost (C_0). Thus, a new test that provides better decisions, and hence reduces decision consequence costs, may in fact increase diagnostic cost if its overhead cost is too high.

Second, the average diagnostic cost (C) depends on both $P(T+ \mid D+)$ and $P(T+ \mid D-)$, which are the same as TPF and FPF and are the coordinates of an ROC curve. Since a decision maker can change these quantities together—that is, since he can move his operating point along the ROC curve—by using a different decision threshold, average cost depends on the decision threshold used and usually can be made larger or smaller. Thus, in terms of cost-benefit analysis, the best operating point on a given ROC curve is the operating point that minimizes average cost (and hence maximizes average benefit) in a particular applied diagnostic situation. Note that the above expression is of the form $\bar{C} = k_1 P(T+ \mid D+) +$

*Benefits can be expressed as negative costs, if desired. At this point, however, it seems clearest to express all decision outcomes in terms of costs. Note that the consequences of a TP decision, for example, usually represent a liability, though almost always a smaller liability than those of a FN decision.

$k_2P(T+ \mid D-) + k_3$. By differentiating the expression, and setting $d\bar{C} = 0$, one can show that this optimal operating point must occur where the curve slope is given by:

$$\begin{bmatrix}\text{ROC Curve}\\ \text{Slope at Best}\\ \text{Operating Point}\end{bmatrix} = \frac{P(D-)}{P(D+)} \times \frac{[C_{FP} - C_{TN}]}{[C_{FN} - C_{TP}]}$$

Note that the best operating point on the ROC curve does not depend on the test overhead cost (C_o) although the minimum average cost ($\bar{C}_{\min}$) attainable at that point does depend on C_o.

In order to understand the effects of the decision-consequence costs and disease prevalence on the optimal ROC operating point—and hence on the optimal compromise between TPF and FPF—recall that the slope of a conventional ROC curve is largest on the lower left portion of the curve and smallest on the upper right portion.

To consider the effects of disease prevalance, note that if the disease in question is rare in the population studies, $P(D+)$ will be much less than 1.0 and $P(D-) = 1 - P(D+)$ will be nearly 1.0. Thus, the ratio $P(D-)/P(D+)$ will be large, and the decision maker should operate toward the lower left portion of his ROC curve (where TPF is small but FPF is much smaller) by using a relatively strict decision threshold. This is necessary when disease is rare (in screening situations, for example) because otherwise almost all positive decisions will be false positive decisions. Conversely, when the disease in question is common, so that $P(D-)$ is small, the best operating point is toward the upper right part of the ROC curve—where TPF is high but FPF is high also. Otherwise, in this situation almost all negative decisions would be false negative decisions.*

To consider the effects of the various decision consequence costs, note that if the difference between the costs of a false positive and of a true negative decision, $C_{FP} - C_{TN}$, is much greater than the difference between the costs of a false negative and of a true positive decision, $C_{FN} - C_{TP}$ (which might be the situation if treatment for the disease or further diagnostic tests were harmful to actually healthy patients but of little benefit to actually diseased patients), then the optimal curve slope is large. Thus, both TPF and FPF are best kept small, because otherwise more harm would be done by the FP decisions than good would be done by the TP decisions. Conversely, if $C_{FP} - C_{TN}$ is much less than $C_{FN} - C_{TP}$, as would be true if treatment or further testing were relatively harmless for actually healthy patients but very beneficial to diseased patients, then the decision maker should be operating high and to the right on his ROC curve, and TPF should be kept large, even at the expense of a large FPF.

If one uses the expression above to find the operating point on an ROC curve that is optimal in a particular applied situation and then substitutes the resulting optimal curve coordinates, $P(T+ \mid D+)$ and $P(T+ \mid D-)$, into the previous expression, one obtains the minimum possible average diagnostic cost ($\bar{C}_{\min}$) that is attainable on the ROC curve and hence attainable with the test it describes. $\bar{C}_{\min}$ can then be used as an index to describe the utility of the test in an applied situation. Since the $\bar{C}_{\min}$ index takes into account both the cost of performing the test and also the costs of the decision consequences realized from the test, it provides a conceptually meaningful way of comparing tests so that the costs of the tests themselves (in terms of money and/or risk) are balanced against the consequences of the decisions they allow.

A slightly different perspective on this ROC approach to cost/benefit analysis can be obtained by thinking in terms of the average net benefit ($\overline{\text{NB}}$) of a diagnostic test, which we define[9] as the amount by which using the test can reduce minimum average diagnostic cost:

$$\overline{NB} = \bar{C}_{\min}\,(\text{not using test}) - \bar{C}_{\min}\,(\text{using test})$$

In the most general situation, the two $\bar{C}_{\min}$ can be obtained by the procedure outlined above. Two ROC curves are required: one measured using all diagnostic information available for each case up to the point in the diagnostic sequence at which the test in question is to be employed, and the other measured using this information and also the results of the test in question. Average net benefit of the additional

*It is of some interest to note that the accuracy index discussed earlier is meaningful when it is appropriate to assume that $C_{FP} = C_{FN}$ and $C_{TN} = C_{TP}$, i.e., when the two kinds of wrong decisions are equally bad and the two kinds of correct answers are equally good. Then we see from the above expression that accuracy is maximized by operating on the curve where the slope equals $P(D-)/P(D+)$.

test will be positive (greater than zero) if the average cost of the decision consequences is reduced by more than the additional diagnostic overhead cost, that is, if the diagnostic information is worth the price.

At least one situation exists for which only a single ROC curve is required. If the decision to be made in the absence of the diagnostic test is negative (as in screening situations, for example) then the average cost of not using the test is

$$\overline{C}\,(\text{not using test}) = C_{\text{TN}} \times P(D-) + C_{\text{FN}} \times P(D+)$$

and so average net benefit is given by[14]

$$\begin{aligned}\overline{NB} = &\{[C_{\text{FN}} - C_{\text{TP}}] \times P(D+)\} \times P(T+ \mid D+) \\ &-\{[C_{\text{FP}} - C_{\text{TN}}] \times P(D-)\} \\ &\times P(T+ \mid D-) - C_0\end{aligned}$$

One can show that in this case average net benefit ($\overline{NB}$) is maximized at the same point on the ROC curve for which average cost ($\overline{C}$) is minimized. A graphical approach that can be used to find the optimal operating point has been published elsewhere.[14]

Decision Analysis and Sequences of Tests

Other papers in this seminar discuss decision analysis in some detail. Essentially, decision analysis provides a means for choosing the course through a decision tree that maximizes average utility—i.e., that maximizes $\overline{NB}$ or minimizes $\overline{C}$. In the previous section, we discussed how one can choose the best compromise between TPF and FPF in order to maximize the average utility of decisions based on a single diagnostic test or a fixed combination of tests.

Diagnostic tests are rarely used alone. Instead, the results of several diagnostic tests are usually combined with clinical background information to decide the disease state of the patient or to decide that additional diagnostic tests should be performed. In order to choose the best sequence of diagnostic tests, that is, to optimize diagnostic strategy, one must recognize that TPF and FPF for each diagnostic test usually can be changed together by changing the decision threshold for the test. Thus, the test result branching probabilities in a decision tree are not fixed, but can be varied by selecting different operating points on the corresponding ROC curves. Full optimization of diagnostic strategy involves choosing not only the best sequence of tests, but also the best operating point on the ROC curve for each test.

A simple example illustrating this kind of full optimization has been published elsewhere.[14] Suppose that we have available two diagnostic tests for the same disease, one with low overhead cost but providing only moderate detectability, and another with much higher overhead cost but providing greater detectability. In terms of $\overline{C}$ or $\overline{NB}$, should one of the tests be used alone, or should the cheaper, less definitive test be used as a screening procedure, with the more expensive, more reliable test used only on patients called positive by the first test? The answer depends, of course, on disease prevalance, on the two overhead costs, on the set of decision consequence costs, and on the ROC curves for the two tests. The published example illustrates combinations of parameters that can yield different conclusions and shows that the optimal operating points on the two ROC curves depend on whether the tests are used alone or in combination.

SUGGESTIONS FOR FURTHER READING

Introductory discussions of ROC analysis for diagnostic evaluation have been published by Turner[15] and by McNeil and colleagues,[16,17] and these articles are recommended for the additional perspective that they provide. Another introductory article by Swets[18] traces the development of ROC analysis in experimental psychology and indicates applications in other fields. We have published elsewhere a partially technical discussion of the ROC approach to diagnostic evaluation that includes examples of the various techniques,[14] and also a concise summary with an extensive bibliography.[19]

A recent introductory book by Egan[20] clearly illustrates the mathematical relationships among various decision strategies, decision variable distributions, and the corresponding ROC curves. *Signal Detection Theory and Psychophysics* by Green and Swets[3] continues as the standard comprehensive reference work on ROC techniques. Finally, although it does not consider the implications of ROC analysis for optimizing diagnostic strategies, a classic book by Raiffa[21] provides an excellent introduction to the principles of decision analysis.

REFERENCES

1. Turner DA, Fordham EW, Pagano JV, et al: Brain scanning with the Anger multiplane tomographic scanner as a secondary examination. Radiology 121:115-124, 1976

2. Goodenough DJ, Metz CE: Implications of a "noisy" observer to data processing techniques, in Raynaud C, Todd-Pokropek AE (eds): Information Processing in Scintigraphy. Orsay, France, CEA, 1975, pp 400-419

3. Green DM, Swets JA: Signal Detection Theory and Psychophysics (rev ed). Huntington NY, Krieger, 1974, pp 99-106

4. Green DM, Swets JA: Signal Detection Theory and Psychophysics (rev ed). Huntington NY, Krieger, 1974, 401-404

5. Green DM, Swets JA: Signal Detection Theory and Psychophysics (rev ed). Huntington, NY, Krieger, 1974, 61-64

6. Dorfman DD, Alf E: Maximum likelihood estimation of parameters of signal detection theory—a direct solution. Psychometrika 33:117-124, 1968

7. Dorfman DD, Alf E: Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. J Math Psych 6:487-496, 1969

8. Metz CE, Starr SJ, Lusted LB: Quantitative evaluation of visual detection performance in medicine: ROC analysis and determination of diagnostic benefit, in Hay GA (ed): Medical Images: Formation, Perception and Measurement. London, Wiley, 1977, pp 220-241

9. Goodenough DJ: Radiographic Applications of Signal Detection Theory (Ph.D. thesis). Chicago, U Chicago, 1972

10. Starr SJ, Metz CE, Lusted LB, et al: Visual detection and localization of radiographic images. Radiology 116:533-538, 1975

11. Starr SJ, Metz CE, Lusted LB: Comments on generalization of receiver operating characteristic analysis to detection and localization tasks. Phys Med Biol 22:376-379, 1977

12. Metz CE, Starr SJ, Lusted LB: Observer performance in detecting multiple radiographic signals: prediction and analysis using a generalized ROC approach. Radiology 121:337-347, 1976

13. Edwards W, Guttentag M, Snapper K: A decision-theoretic approach to evaluation research, in Struening EL, Guttentag M (eds): Handbook of Evaluation Research, Vol. 1. Beverly Hills Calif, Sage, 1975, pp 139-181

14. Metz CE, Starr SJ, Lusted LB, et al: Progress in evaluation of human observer visual detection performance using the ROC curve approach, in Raynaud C, Todd-Pokropek AE (eds): Information Processing in Scintigraphy. Orsay, France, CEA, 1975, pp 420-439

15. Turner DA: An intuitive approach to receiver operating characteristic curve analysis. J Nucl Med 19:213-220, 1978

16. McNeil BJ, Keeler E, Adelstein SJ: Primer on certain elements of medical decision making. N Engl J Med 293:211-215, 1975

17. McNeil BJ, Adelstein SJ: Determining the value of diagnostic and screening tests. J Nucl Med 17:439-448, 1976

18. Swets JA: The relative operating characteristic in psychology. Science 182:990-1000, 1973

19. Metz CE, Starr SJ, Lusted LB: Quantitative evaluation of medical imaging, in Medical Radionuclide Imaging, Vol. 1. Vienna, IAEA, 1977, pp 491-504

20. Egan JP: Signal Detection Theory and ROC Analysis. New York, Academic Press, 1975

21. Raiffa H: Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Reading Mass, Addison-Wesley, 1968