Presented at the 1979 AAPM Summer School, "Recording System Measurements and Techniques," Chapel Hill, NC, July 22-28, 1979. To be published in the proceedings by The American Institute of Physics (in press).

APPLICATIONS OF ROC ANALYSIS IN DIAGNOSTIC IMAGE EVALUATION*

Charles E. Metz, Ph.D.
The University of Chicago and
The Franklin McLean Memorial Research Institute†
Chicago, Illinois 60637

ABSTRACT

The need for Receiver Operating Characteristic (ROC) analysis is indicated by a discussion of the limitations of "accuracy" and of "sensitivity" and "specificity" as indices of diagnostic detection or discrimination performance. The concept of a variable decision threshold is shown to lead in a natural way to the ROC curve as a means for specifying diagnostic performance. Practical techniques for measuring ROC curves are described, and directions for possible generalizations of conventional ROC analysis are indicated.

INTRODUCTION

How can we measure the quality of diagnostic information and diagnostic decisions in a meaningful way? That basic question has become increasingly important in recent years as an abundance of new diagnostic tests has 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; it will not go away.

The fundamental relationships between the physical properties of a diagnostic medical image (such as resolution, contrast, and statistical fluctuations) and the ability of a human observer to properly detect and interpret relevant image features are poorly understood. In real diagnostic tasks, these relationships are undoubtedly complicated by problems of complex background structure, normal anatomical variations, and observer training. Thus, at present, one cannot confidently predict the diagnostic performance of a medical imaging procedure from knowledge of its physical characteristics. Instead, one must objectively measure the diagnostic detection performance that can be achieved by human observers who view images made with real medical imaging systems. Hopefully, the data obtained in this way ultimately

* Much of the text of this paper is taken from: Metz, C.E., Basic Principles of ROC Analysis, Seminars in Nuclear Medicine 8: 283-298, 1978.

† Operated by The University of Chicago for the U.S. Department of Energy under Contract No. EY-76-C-02-0069.

will contribute to an improved understanding of the visual process in diagnostic medicine. For now, these data -- if properly collected -- can provide an empirical quantitative description of diagnostic imaging system performance.

Any meaningful approach to the evaluation of diagnostic performance must inevitably involve many complex technical and social issues, and one cannot reasonably expect that the typical physicist or physician should master all of the subtleties involved. Still, the basic concepts upon which diagnostic performance 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 paper 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 only upon 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. To the extent that this is true, the reader may find himself in the position of Moliere's gentleman who was pleased to learn that he had been speaking prose for years.

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, which is often called "Accuracy." Although we are 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 so strongly, and no mathematical "correction"or"normalization" for disease prevalence can redeem this index in a meaningful way. One might be tempted to suppose that, though this be true, "Accuracy" should be meaningful at least as an index for comparison of

diagnostic techniques applied to a given population in which disease prevalence is known and fixed. Here, too, the index is limited, however. 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 negatives (misses), while those from the other might be almost all false positives (false alarms). Clearly, the relative 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 indeces:

$$\text{SENSITIVITY} = \frac{\text{Number of true positive (TP) decisions}}{\text{Number of actually positive cases}}$$

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; this disease state 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, etc.

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

$$\text{ACCURACY} = [\text{SENSITIVITY}] \times [\text{Fraction of all cases that is actually positive}] + [\text{SPECIFICITY}] \times [\text{Fraction of all cases that is actually negative.}]$$

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:

$$\text{ACCURACY} = \frac{\#\text{ correct decisions}}{\#\text{ cases}}$$

so

$$\text{ACCURACY} = \left[\frac{\#\text{ True Positive decisions}}{\#\text{ cases}}\right] + \left[\frac{\#\text{ True Negative decisions}}{\#\text{ cases}}\right]$$

$$= \left[\frac{\#\text{ True Positive decisions}}{\#\text{ actually positive cases}}\right] \times \left[\frac{\#\text{ actually positive cases}}{\#\text{ cases}}\right]$$

$$+ \left[\frac{\#\text{ True Negative decisions}}{\#\text{ actually negative cases}}\right] \times \left[\frac{\#\text{ actually negative cases}}{\#\text{ cases}}\right]$$

and 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 (abbreviated "TPF" is simply the same thing as "Sensitivity," and True Negative Fraction)(abbreviated "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, for this writer at least, are a lot easier to remember. These new terms suggest two other definitions:

$$\text{FALSE POSITIVE FRACTION (FPF)} = \frac{\#\text{ False Positive decisions}}{\#\text{ actually negative cases}}$$

and

$$\text{FALSE NEGATIVE FRACTION (FNF)} = \frac{\#\text{ False Negative decisions}}{\#\text{ actually positive cases}}$$

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

$$TPF + FNF = 1$$

and

$$TNF + FPF = 1$$

(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 kind of 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 $P(T+|D-)$, which is read as "the probability of a positive test, given the absence of disease." Similarly, TPF is often denoted by $P(T+|D+)$; FNF by $P(T-|D+)$; and TNF by $P(T-|D-)$. 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 that we've defined so far are summarized in Table I. 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 separate out the effect of disease prevalence and to score separately the performance of a diagnostic test or a diagnostic decision maker with respect to actually positive and actually negative cases.

TABLE I. Definitions of, and relationships among, the various decision performance indices described in the text. (Metz, Ref. 3)

Definitions:

$$TPF = SENSITIVITY = P(T+|D+)$$

$$FPF = 1-(SPECIFICITY) = P(T+|D-)$$

$$TNF = SPECIFICITY = P(T-|D-)$$

$$FNF = 1-(SENSITIVITY) = P(T-|D+)$$

$$\text{Disease Prevalence} = P(D+) = 1 - P(D-)$$

Relationships

$$TPF + FNF = P(T+|D+) + P(T-|D+) = 1$$

$$TNF + FPF = P(T-|D-) + P(T+|D-) = 1$$

$$\begin{aligned} ACCURACY &= SENSITIVITY \times P(D+) \\ &\quad + SPECIFICITY \times P(D-) \\ &= TPF \times P(D+) + TNF \times P(D-) \\ &= P(T+|D+) \times P(D+) + P(T-|D-) \times P(D-) \end{aligned}$$

In order to see how these concepts can be applied to a collection of diagnostic decisions, consider the following hypothetical situation. Suppose that 1,200 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 later 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 true positive (TP) decisions, 60 false negative (FN) decisions, 900 true negative (TN) decisions, and 100 false positive (FP) decisions. These data can be summarized by the "decision matrix" shown in Table II. 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 I are the same as those that would be obtained using the definitions of these quantities directly.

We see from the calculated indices that this test, used as it has been 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 then actually positive cases were decided incorrectly. The latter observation is not paradoxical, but merely reflects the preponderance of actually

TABLE 2. Decision data and calculated indices for hypothetical Test "A". (Metz, Ref. 3).

Actual State	Test Result (Diagnosis): Positive (T+)	Test Result (Diagnosis): Negative (T-)	
Positive (D+)	140 (TP)	60 (FN)	200 actually + cases
Negative (D-)	100 (FP)	900 (TN)	1000 actually - cases
	240 + decisions	960 - decisions	1200 total cases

Calculated Indices

$$TPF = \frac{140}{200} = 0.70; \quad FNF = 1\text{-}TPF = 0.30$$

$$FPF = \frac{100}{1000} = 0.10; \quad TNF = 1\text{-}FPF = 0.90$$

$$P(D+) = \frac{200}{1200} = 0.17; \quad P(D-) = 1\text{-}P(D+) = 0.83$$

$$ACCURACY = TPF \times P(D+) + TNF \times P(D-) = 0.87$$

negative cases in the population studied; recall that TPF, TNF, etc., represent "rates" and not "numbers of cases."

The decision fractions allows 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+). If the various decision fractions are kept constant but P(D+) is increased to 0.6, for example, then "Accuracy" would be (0.7) x (0.6) + (0.9) x (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" were studied using a different test, Test "B", with the results shown in Table III. Comparison of Tables II and III 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--TPF is lower and FNF is higher--but it is performing better for actually negative cases--

TABLE III. Decision data and calculated indices for hypothetical test "B". (Metz, Ref. 3).

Actual State	Test Result (Diagnosis): Positive (T+)	Test Result (Diagnosis): Negative (T-)	
Positive (D+)	80 (TP)	120 (FN)	200 actually + cases
Negative (D-)	40 (FP)	960 (TN)	1000 actually - cases
	120 + decisions	1080 - decisions	1200 total cases

Calculated Indices

$$TPF = \frac{80}{200} = 0.40; \quad FNF = 1\text{-}TPF = 0.60$$

$$FPF = \frac{40}{1000} = 0.04; \quad TNF = 1\text{-}FPF = 0.96$$

$$P(D+) = \frac{200}{1200} = 0.17; \quad P(D-) = 1\text{-}P(D+) = 0.83$$

$$ACCURACY = TPF \times P(D+) + TNF \times P(D-) = 0.87$$

TNF is higher and FPF is lower. The "Accuracy" indices are equal because this "trade-off" in performance is just 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, the Test A is worse, and 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 (or, equivalently, of TPF and TNF)? 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. First, however, let us consider a further complication, which will suggest a solution to the present dilemma.

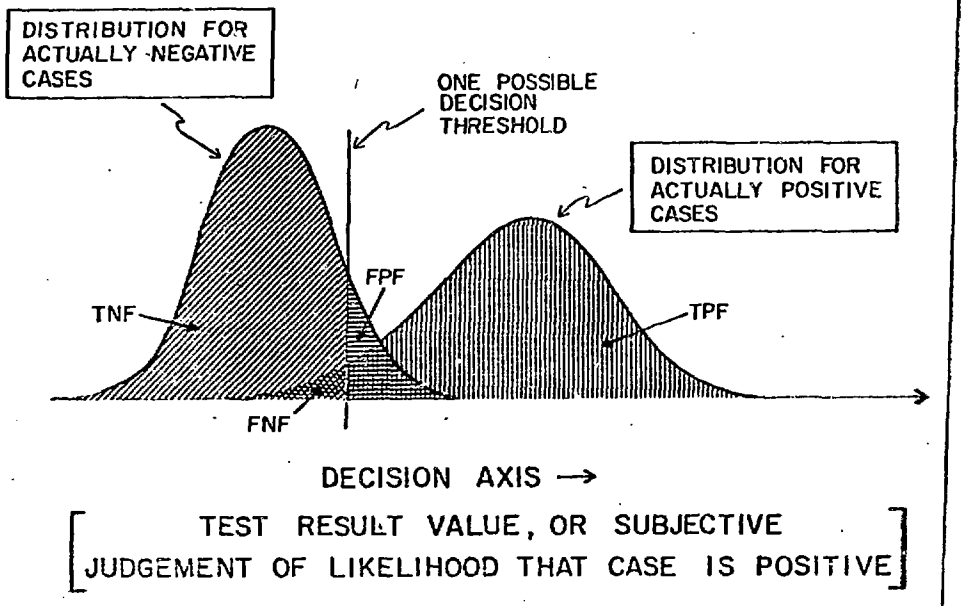

Fig. 1. Two hypothetical distributions of a quantity on which decisions are based, showing one possible decision threshold. The conditional probability of each kind of decision is equal to the area under a distribution on one side of the threshold. (Metz, Ref. 3)

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 hour 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 which 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 Figure 1, then increasing the threshold value will make

both false positive and true positive decisions less frequent, but will also make both true negative and false negative decisions more frequent. A threshold value must be selected that is believed to yield an appropriate compromise between these gains and losses.

Similarly, diagnostic tests which 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 than an observer adopts undoubtedly depends upon 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--and the concept of a confidence threshold may be hard to quantify. Still, 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 II and III could be changed simply by arbitrarily selecting different thresholds or by using 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.

THE INSIGHT PROVIDED BY RECEIVER OPERATING CHARACTERISTIC 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 re-read 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 another set of decision fractions. Since TPF and FPF together determine all four decision fractions, 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 Figure 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.

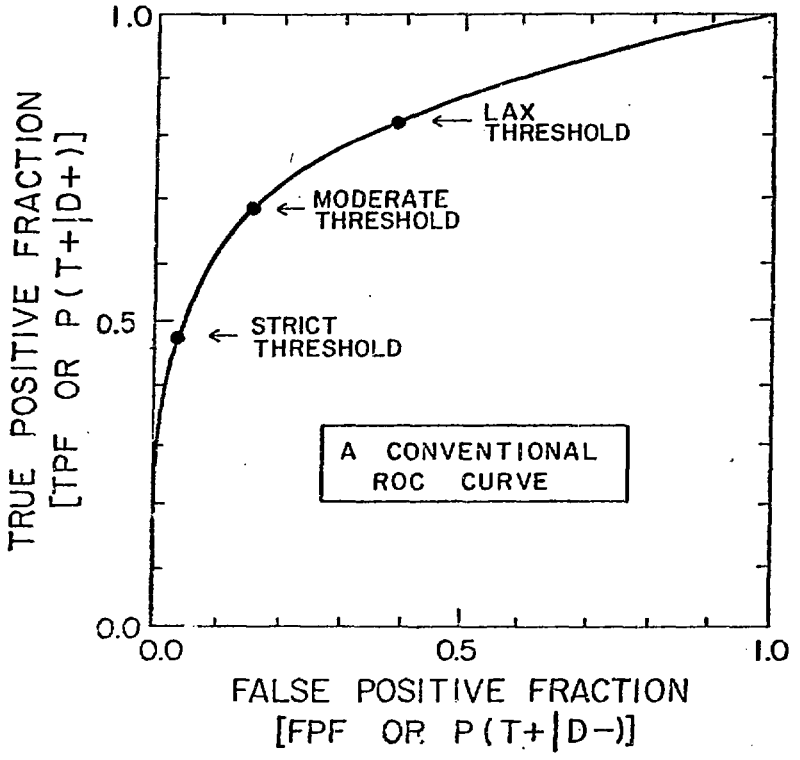

Fig. 2. A typical conventional ROC curve, showing three possible operating points. (Metz, Ref. 3)

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 like that shown in Figure 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" or "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 using an appropriate decision threshold. Figure 2 shows three possible operating points that might correspond to use of "strict" threshold (case called "positive" only if

judged almost definitely positive), a "moderate" threshold, or a "relaxed" threshold (case called "positive" if any suspicion of disease).

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 (FPF = 0, TPF = 0) corner of the graph because one can adopt a threshold so strict that almost all tests are called negative, and the curve must pass through the upper right (FPF = 1, TPF = 1) corner of the graph because one can adopt a threshold so relaxed that almost all tests are 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 diagonal (i.e., 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., $P(T+|D+)$ should be greater than $P(T+|D-)$. Finally, one can show theoretically that, if the decision maker knows the underlying probability density functions and uses test 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 of 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 for a given test. 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 which are independent from disease prevalence when a fixed decision threshold is used, the ROC curve does not depend upon the prevalence of disease in a population to which the corresponding test may be applied.* Thus ROC analysis provides a description of disease detectability that is independent from both disease prevalence and decision threshold effects.

---

* 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.

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 upon 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 fixed his decision threshold. 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 II and III. 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

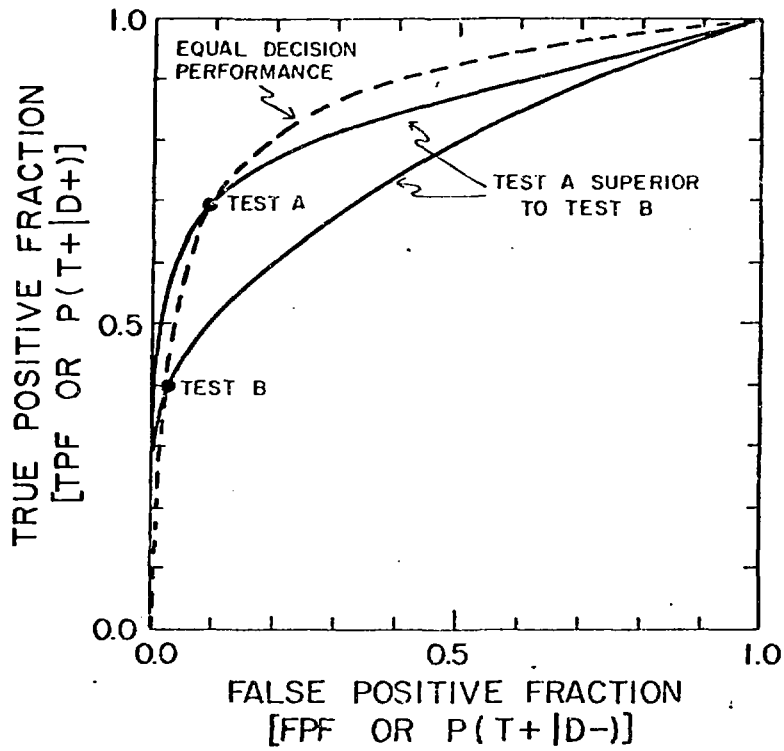

Fig. 3. The decision fractions resulting from the data on Tbls. 1 and 2 plotted as points in an ROC space, with possible ROC curves on which these points could lie. (Metz, ref. 3)

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.

Figure 3 shows an "ROC" space" in which are plotted 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 II and III. If we were to measure ROC curves for the two tests by changing (consistently) 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, and 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 in Figure 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, because disease prevalence and the costs and benefits of the consequences of the various types of decisions determine the part of an ROC curve on which a decision-maker should operate.[3]

Figure 4 displays ROC curves obtained in an experiment designed to evaluate the relative visual detectability of small, low contrast objects that is provided by four different radiographic screen-film systems. Each graph shows the ROC curves obtained by a single observer. These results are of particular interest in that the (RP, TF-2) and (RP/R, PS) systems provide very different detectability but have essentially the same speed--and hence require the same patient exposure. The (RP, PS) and (RP/R, TF-2) systems require approximately twice and one-half the exposure of the other systems, respectively. Thus these ROC curves show that the (RP, TF-2) system is clearly superior to the (RP/R, PS) system for detection of such objects, and they indicate the gain or loss in detectability that can be achieved by increasing or decreasing patient exposure by a factor or two.

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.

This approach is often impractical for diagnostic tests that must be interpreted subjectively, however, 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 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

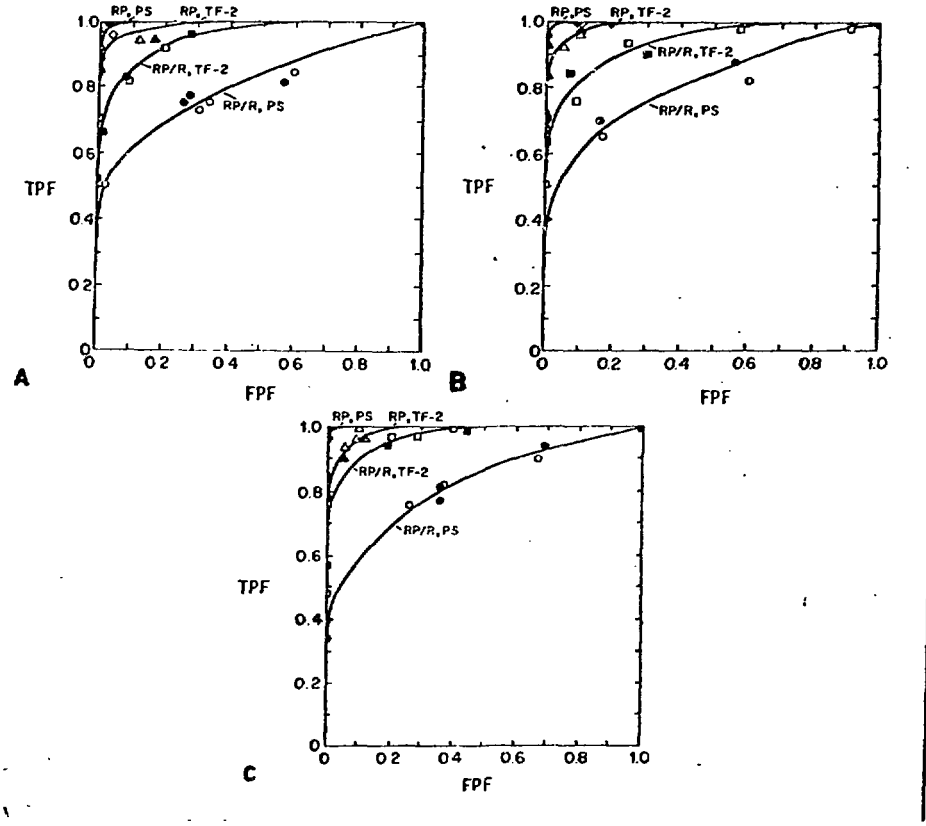

Fig. 4. ROC curves generated by (A) observer 3, a senior radiologist; (B) observer 4, a physicist; and (C) observer 5, a physicist. These curves were obtained in a radiographic signal detection experiment described elsewhere.[11] The signal was the radiographic image of a 2-mm diameter Lucite bead, and noise resulted from the radiographic mottle of the following diagnostic screen-film combinations: RP-Kodak RP X-omat medical x-ray film (normal speed); RP/R-Kodak RP Royal X-omat medical x-ray film (fast speed); PS-DuPont Cronex Par Speed Screen (medium speed); and TF-2-Radelin TF-2 Screen (fast speed). Open and solid symbols of a given shape indicate independent trial runs with the same observer and the same set of images. Each trial run consisted of approximately 100 observations. Note the reproducibility of the curves from observer to observer for this simple detection task. (Metz et.al. Ref. 16).

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 "re-read" 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 judgement situation is called the "Rating Method" and was developed in experimental psychology.[4] 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 judgement 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) non-trivial 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, responses in category "5" only 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 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 intrepreting 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

Table IV. Simulated rating scale data and calculation of ROC points.

RATING SCALE DATA

Confidence Rating:

	1	2	3	4	5	
Actually (+) cases	5	6	5	12	22	$\Sigma = 50$
Actually (-) cases	30	19	8	2	1	$\Sigma = 60$

Entries show number of cases for which indicated rating was used.

CALCULATION OF ROC POINTS

A. (5) = "+" decision

TPF = 22/50 = 0.44
FPF = 1/60 = 0.02

B. (5 or 4) = "+" decision

TPF = (22+12)/50 = 0.68
FPF = (1+2)/60 = 0.05

C. (5, 4, or 3) = "+" decision

TPF = (22+12+5)/50 = 0.78
FPF = (1+2+8)/60 = 0.18

D. (5, 4, 3, or 2) = "+" decision

TPF = (22+12+5+6)/50 = 0.90
FPF = (1+2+8+19)/60 = 0.50

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, "5" or "4" or "3" = "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.

Table IV shows an example of rating scale data (generated by computer simulation) and the calculation of ROC operating points from those data. Figure 5 displays the calculated operating points on an ROC graph, together with the $\pm$ 1 standard deviation error bars estimated from the data (by the method explained in the next section) and the maximum likelihood ROC curve estimated from

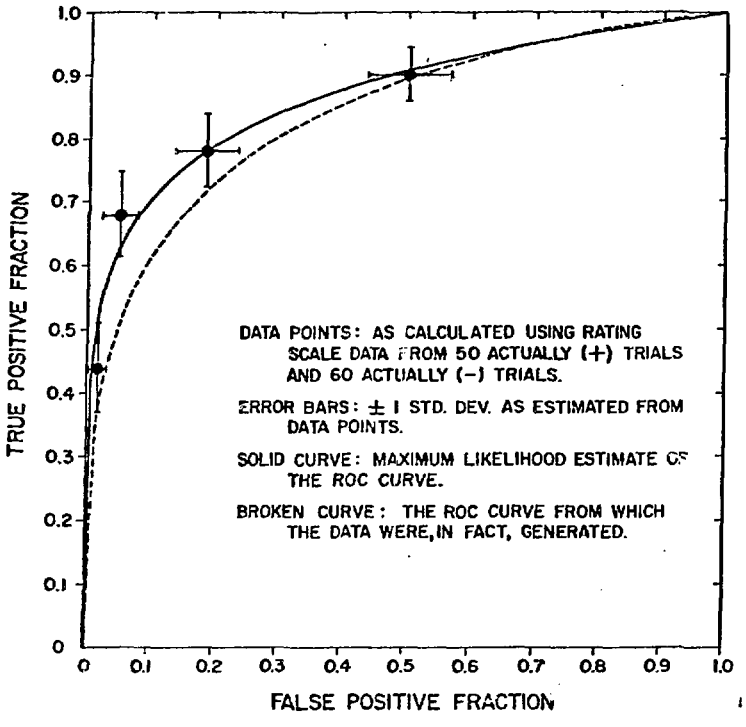

Fig. 5. Simulated rating scale data with the actual (broken) and fitted (solid) ROC curves

the data (using a procedure referenced in the next section). Also shown, by a broken line, is the actual ROC curve from which the rating scale data were generated by computer simulation. The discrepancy between the actual and estimated ROC curves is typical of that which can be expected if about 50 trials of each kind are used to measure an 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, 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 was repeated using a different set of the same number of cases--can be estimated by the expressions[5]:*

---

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

$$\text{STD. DEV. (TPF)} = \sqrt{\frac{\text{TPF} \times (1-\text{TPF})}{(\text{\# actually positive cases}) - 1}}$$

and

$$\text{STD. DEV. (FPF)} = \sqrt{\frac{\text{FPF} \times (1-\text{FPF})}{(\text{\# actually negative cases}) - 1}}$$

These expressions can be used to plot $\pm$ 1 or 2 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.[5] 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; (2) the standard deviation of TPF is inversely related to the number of actually positive cases used in the experiment; and (3) the standard deviation of FPF is 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 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. An assumption commonly used 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 Figure 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 judgements.

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 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.[6]

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[7] or in rating-method experiments.[8]

The maximum-likelihood programs mentioned above provide, as a by-product, estimates of the variances and covariance of the two ROC curve parameters. These can be used to construct a test of the statistical significance of apparent differences between a measured ROC curve and an assumed curve or between two ROC curves measured from statistically independent data. Statistical testing can be done either in terms of a single index of detectability derived from the two curve parameters, or in terms of the two parameters simultaneously using an appropriate Chi-square statistic with two degrees of freedom.

Truth, Cases, 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 requirements are

---

* 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, Massachusetts 02062

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 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;[9] 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.[9,10]

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 states 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 upon 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 judgements are

---

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

reproducible. Although the cumulative effects of these two sources of variation can be expressed in terms of binomial and 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. The fact that both of these two effects do occur unquestionably complicates the issue of interpreting apparent differences between measured ROC curves, however. 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: 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 population samples 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.[11]

GENERALIZED RECEIVER OPERATING CHARACTERISTIC METHODS

The conventional ROC methods that we have described up to this point 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 thresholds 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, then five actual

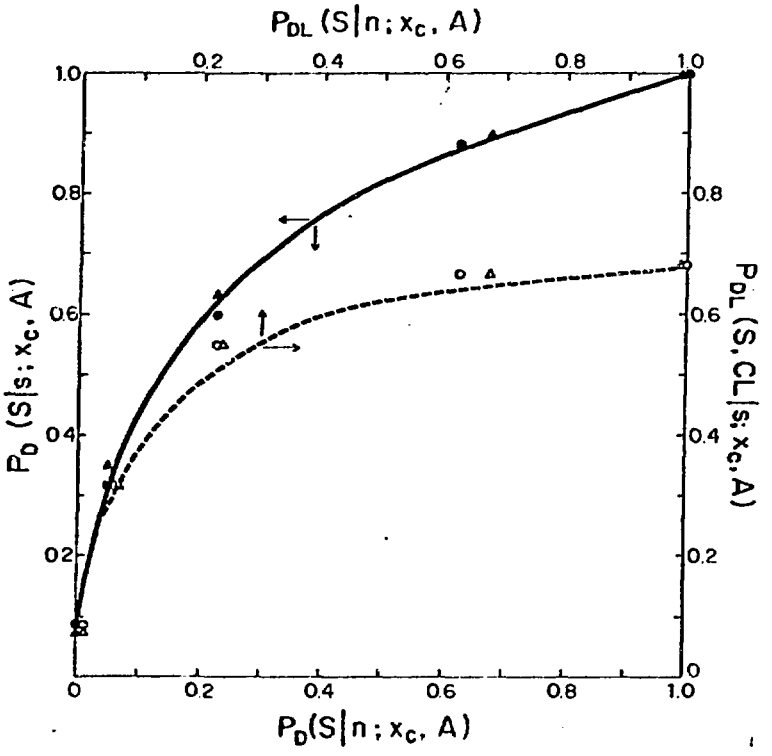

Fig. 6. Conventional ROC curve and generalized curve for detection and localization task. (Starr et al., ref. 12)

states and decision alternatives are available: "no disease," "disease in upper left quadrant," etc. We have shown theoretically and experimentally[12,13] 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-dimensional space, which can be plotted as two curves on a two-dimensional graph.

Typical results obtained using this generalized ROC approach[12] are shown in Figure 6. The solid symbols of different shape represent conventional ROC data points obtained from separate viewing sessions by the same observer (viewing the same image set, which consisted of 50 signal-plus-noise and 50 noise-only images). The solid curve was fit to these data points and was used to predict the lower, broken curve, which should represent observer performance when the image quadrant containing the signal must be specified. The open symbols show data obtained in an experiment in which both detection and localization were required and agree with the predicted broken curve.

Another situation of interest is that for which more than one lesion, for example, may be actually present and for which the observer must, in effect, count the lesions present. 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 multidimensional space, which can be plotted as a set of two-dimensional graphs.[14]

These two studies have shown that decision performance in some multi-alternative tasks employing medical images can be related uniquely and predictably to decision performance in a simple two-alternative task, which 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 judgement 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 judgment process and then "adding up" the number of lesions that he has "found."

An appropriate theoretical model for what we might call a "simultaneous detection and differential diagnosis" task is less clear.[10] 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 multi-alternative 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, risk of complications, and/or radiation exposure) to gain information that should be of benefit in subsequent patient management. ROC analysis provides a means of measuring and describing diagnostic detectability in terms of the combinations that can be achieved among the relative frequencies of true positive, false positive, true negative, and false negative decisions. Thus, through ROC analysis one can determine the information that a diagnostic test can provide. The term "information" here can be

interpreted either in the loose sense of "detectability" or in the technical sense developed by Shannon.[15,16]

With disease detection performance specified by ROC analysis, several important questions remain, however. In a particular diagnostic task, which is the best of the possible combinations among the various decision frequencies, that is, what is the best operating point on the ROC curve? How can one judge whether the diagnostic information purchased by the use of a diagnostic test is (expected to be) worth the price paid? And how can a diagnostic test best be used within the context of a diagnostic strategy? These questions can be addressed, at least conceptually, by combining ROC analysis with the techniques of cost/benefit analysis and decision analysis. Discussions of this approach can be found elsewhere.[3,17]

SUGGESTIONS FOR FURTHER READING

Introductory discussions of ROC analysis for diagnostic evaluation have been published by Swets[18], Turner[19], and by McNeil and colleagues[20,21], and these papers are recommended for the additional perspective that they provide. Other introductory papers by Swets[22] and by Swets and Green[23] trace the development of ROC analysis in experimental psychology and indicate 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[17] and also a concise summary with an extensive bibliography.[24]

A recent introductory book by Egan[25] 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[4] 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[26] provides an excellent introduction to the principles of decision analysis.

REFERENCES

1. D. A. Turner, E. W. Fordham, J. V. Pagano, et al, Radiology 121, 115 (1969).
2. D. J. Goodenough, C. E. Metz, in C. Raynaud, A. E. Todd-Pokropek, Information Processing Scintigraphy (CEA, Orsay, France, 1975).
3. C. E. Metz, Sem. Nucl. Med. 8, 283 (1978).
4. D. M. Green, J. A. Swets, Signal Detection Theory and Psychophysics (Krieger, Huntington, NY) p. 99.
5. Ibid., p. 401.
6. Ibid., p. 61
7. D. D. Dorfman, E. Alf, Psychometrika 33, 117 (1968).

8. D. D. Dorfman, E. Alf, J. Math. Psych. 6, 487 (1969).
9. D. F. Ransohoff, A. R. Feinstein, N. Engl. J. Med. 299, 926 (1978).
10. C. E. Metz, S. J. Starr, L. B. Lusted, in G. A. Hay, Medical Images: Formation, Perception and Measurement (Wiley, London, 1977) p. 220.
11. D. J. Goodenough, Radiographic Application of Signal Detection Theory (PhD Thesis) (U. Chicago, Chicago, 1972).
12. S. J. Starr, C. E. Metz, L. B. Lusted, et al, Radiology 116, 553 (1975).
13. S. J. Starr, C. E. Metz, L. B. Lusted, Phys. Med. Biol. 22, 376 (1977).
14. C. E. Metz, S. J. Starr, L. B. Lusted, Radiology 121, 337 (1976).
15. C. E. Shannon, W. Weaver, The Mathematical Theory of Communication (Univ. of Illinois Press, 1949).
16. C. E. Metz, D.J. Goodenough, K. Rossmann, Radiology 109, 297 (1973).
17. C. E. Metz, S. J. Starr, L. B. Lusted, et al, in C. Raynaud, A. E. Todd-Pokropek, Information Processing in Scintigraphy (CEA, Orsay, France, 1975) p. 420.
18. J. A. Swets, Invest. Radiol. 14, 109 (1979).
19. D. A. Turner, J. Nucl. Med. 19, 213 (1978).
20. B. J. McNeil, E. Keeler, S. J. Adelstein, N. Engl. J. Med. 17, 439 (1976).
21. B. J. McNeil, S. J. Adelstein, N. Engl. J. Med. 293, 211 (1975).
22. J. A. Swets, Science 182, 990 (1973).
23. J. A. Swets, D. M. Green, in H. L. Pick et al, Psychology, from Research to Practice (Plenum Press, New York, 1978) p. 311.
24. C. E. Metz, S. J. Starr, L. B. Lusted,in Medical Radionuclide Imaging, Vol. 1. (IAEA, Vienna, 1971), p. 491.
25. J. P. Egan, Signal Detection Theory and ROC Analysis (Academic Press, New York, 1975).
26. H. Raiffa, Decision Analysis: Introductory Lectures on Choices under Uncertainty. (Addison-Wesley, Reading MA, 1968).