An Intuitive Approach to Receiver Operating Characteristic Curve Analysis

David A. Turner

Rush-Presbyterian-St. Luke's Medical Center and Rush Medical College, Chicago, Illinois

Meaningful comparison of diagnostic imaging tests inevitably requires expression of observer performance. The most commonly used measures of observer performance—sensitivity, specificity, and percentage accuracy—fail to take account of variability of the diagnostic-criterion level employed by observers. The diagnostic-criterion level can be thought of as the dividing line between diagnostic image findings called positive for disease and findings that are called negative. A "strict" criterion level leads to high specificity and low sensitivity, whereas a "lax" criterion level results in low specificity and high sensitivity. Since sensitivity and specificity, as well as accuracy, change with the criterion level employed, comparison of imaging tests in these terms may yield ambiguous or misleading results. Receiver operating characteristic (ROC) curves avoid this problem by comparing sensitivity and specificity over a wide and continuous range of criterion levels. Hence, ROC curves provide an unambiguous representation of the relative inherent detectability of disease by the diagnostic imaging tests being compared. ROC curves can also be used to determine optimum criterion levels that maximize accuracy, average net benefit, or other measures of clinical efficacy. Thus, in a variety of applications, the ROC method can increase the precision with which medical imaging tests are evaluated.

J Nucl Med 19: 213–220, 1978

New diagnostic imaging techniques have been introduced into medical practice at an accelerating pace in recent years. The evaluation of the relative usefulness of these new procedures has been made increasingly important by the rapidly rising economic cost of medical care. Inevitably, any meaningful comparison of the efficacy of diagnostic imaging techniques will involve some expression of *observer performance*—that is to say, some estimate of the "correctness" of the diagnoses that result from the use of each technique. Unless observer performance is expressed in a way that is appropriate, however, the results of experiments comparing diagnostic imaging techniques may be ambiguous or misleading.

It will be argued below that certain measures of observer performance commonly used to compare diagnostic imaging tests—for example "sensitivity" (*1*), "specificity" (*1*), and "overall percentage accuracy" (*2*)—are not always optimal for this purpose because they fail to account for the variability of the so-called diagnostic "criterion level" (*3*). The purposes of this discussion will be to illustrate the need to resort to a particular mode of expressing observer performance that does take this variability into account, namely the receiver operating characteristic (ROC) curve (*3,4*), and to leave the reader with a basic understanding of the ROC method.

Definition of terms. It is the common experience in medicine that no matter how one tests patients, the appearance of healthy and diseased patient populations will overlap to some extent, making it impossible to separate the groups completely. As an example, let us consider the familiar "cardiothoracic

Received Aug. 1, 1977; revision accepted Oct. 10, 1977.

For reprints contact: David A. Turner, Dept. of Nuclear Medicine, Rush-Presbyterian-St. Luke's Medical Ctr., 1753 West Congress Parkway, Chicago, IL 60612.

ratio" (CTR) (*5*), which may be used as a "test" for the presence of heart disease. (The CTR is the maximum transverse diameter of the cardiac silhouette, as measured on a chest radiograph, divided by the maximum transverse diameter of the thorax. It is usually expressed as a percentage.) Figure 1A illustrates hypothetical distributions of CTRs for two populations, one with and the other without heart disease. As can be seen, there is considerable overlap.

The upper limit of normal of the CTR is generally taken as 50%. Patients with a CTR of 50% or less may be said to have a "negative" test for heart disease, and those with a CTR above 50% may be said to have a "positive" test. Since 50% is the "dividing line" between positive and negative tests, it is the diagnostic-criterion level.

When we use the CTR as a test for heart disease and choose a diagnostic criterion level of 50%, Figure 1A shows us that there will be four kinds of "test outcome." Most of the abnormal patients will have CTRs above 50%, and, when the test is applied to them, they will be correctly identified as having heart disease—that is to say, the test outcome will be true positive (Fig. 2). Some of the diseased patients, however, will have CTRs less than 50%. For these patients, the test outcome will be false negative. The fraction of diseased patients correctly identified by the test (the number of abnormal patients with positive tests divided by the total number of abnormal patients tested) is referred to as the true-positive fraction. Those normal patients with cardiothoracic ratios above 50% will have a false-positive CTR test outcome, while those with ratios of 50% or less will have a true-negative test outcome. The fraction of normal patients who correctly have negative tests (the number of normal patients with negative tests divided by the total number of normal patients tested) is the true-negative fraction. The false-positive fraction is the number of normal patients with positive tests divided by the total number of normal patients. The sum of the true-negative fraction and the false-positive fraction is 1, so that both values are known if one of them is given. The same holds true for the true-positive plus false-negative fractions (Fig. 2). The true-positive fraction expressed as a percentage is frequently referred to as the "sensitivity" of a test, whereas the true-negative fraction as a percentage is called the "specificity" (*1*).

The dependence of sensitivity and specificity on the diagnostic-criterion level. The "accuracy" of diagnostic medical tests is frequently expressed in terms of specificity and sensitivity, with the implication that every test has only one true-negative fraction and one true-positive fraction. In fact, the

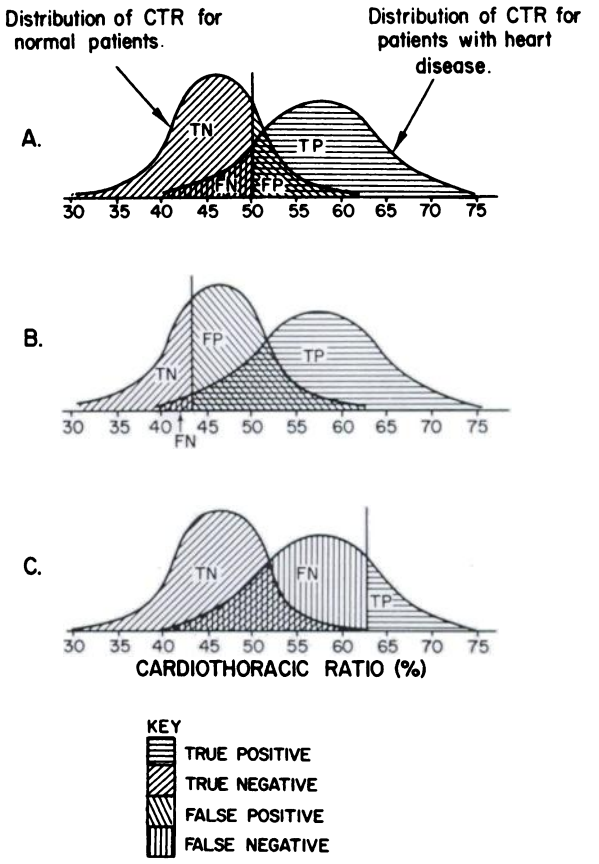

FIG. 1. Hypothetical distribution of cardiothoracic ratios for patients with and without heart disease. TN = true negative; TP = true positive; FN = false negative; FP = false positive. (A) Criterion level = 50%. (B) "Low" criterion level (43%). (C) "High" criterion level (63%).

		DISEASE PRESENT	DISEASE ABSENT
DIAGNOSTIC TEST	POSITIVE	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	NEGATIVE	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

TRUE POSITIVE FRACTION	=	$\frac{TP}{TP + FN}$	=	"SENSITIVITY"
TRUE NEGATIVE FRACTION	=	$\frac{TN}{TN + FP}$	=	"SPECIFICITY"
FALSE POSITIVE FRACTION	=	$\frac{FP}{TN + FP}$	=	1 - TRUE NEGATIVE FRACTION
FALSE NEGATIVE FRACTION	=	$\frac{FN}{TP + FN}$	=	1 - TRUE POSITIVE FRACTION

FIG. 2. Two × two decision matrix at top. Note (below) that TRUE-POSITIVE FRACTION + FALSE-NEGATIVE FRACTION = 1, so that if one is known, both are known. As noted in text, same is true of TRUE-NEGATIVE and FALSE-POSITIVE FRACTIONS.

sensitivity and specificity of a test will change if the diagnostic-criterion level is altered. For example, we can easily increase the sensitivity of the CTR test by lowering the diagnostic criterion level to 43% (Fig. 1B). Nearly all of the patients with heart disease will be detected (i.e., they will have true-positive tests). However, Fig. 1B also shows us that when the criterion level is lowered, the specificity of the test drops dramatically, i.e. there is a fall in the true-negative fraction (and a rise in the false-positive fraction). On the other hand, the specificity can be increased by raising the criterion level to 63% (Fig. 1C), but only at a great loss in sensitivity. We can see by these examples that the specificity and sensitivity of the CTR test are not fixed values, but vary *continuously* (see Fig. 3) with changes in the criterion level.

The result of a CTR test is a specific number. The distribution of the results of this test for patients with and without heart disease, therefore, can easily be represented in graphic terms. In principle, the results of subjectively interpreted medical imaging tests also can be thought of as having distinct distributions for normal and abnormal patients, analogous to the distributions for the CTR test, in spite of the fact that these results are not easily expressed in terms of specific numbers. For example, a radiologist diagnoses pneumonitis on a chest radiograph by noting that the radiodensity in the involved region of lung is "greater than normal." The radiodensity of lung in regions of pneumonitis varies greatly, however, from very dense lesions, which are easily detected on chest radiography, to lesions that result in only minimal increase in radiodensity and are not readily detectable. Similarly, the *apparent* radiodensity of the lungs of normal patients varies considerably because of variation in chest-wall thickness, technical exposure factors, and so on. In some cases, the radiodensity of the lungs of normal patients will *appear* to exceed that of some patients with pneumonitis. Hence, we can think of the apparent radiographic density of the lungs of patients with and without pneumonitis as having two overlapping distributions analogous to the normal and abnormal distributions of cardiothoracic ratios.

As in the case of the CTR test, the sensitivity and specificity of chest radiography in the detection of pneumonia will vary with the criterion level used by the radiologist to decide whether or not pneumonitis is present. If he concludes that pneumonitis is present given only a very slight "increase" in radiodensity above the mean density expected for a normal population of patients, he will "miss" relatively few cases of pneumonitis. On the other hand, his true-negative fraction (specificity) will be relatively low (and his false-positive fraction will be high). He can increase his specificity (i.e. decrease his false-positive fraction) by raising his criterion level so that he calls the radiograph "positive" for pneumonitis only if there is a relatively great increase in the radiodensity of the lung. He will do so, however, only at the cost of a decreasing true-positive fraction (sensitivity), that is to say, he will "miss" more cases of pneumonitis.

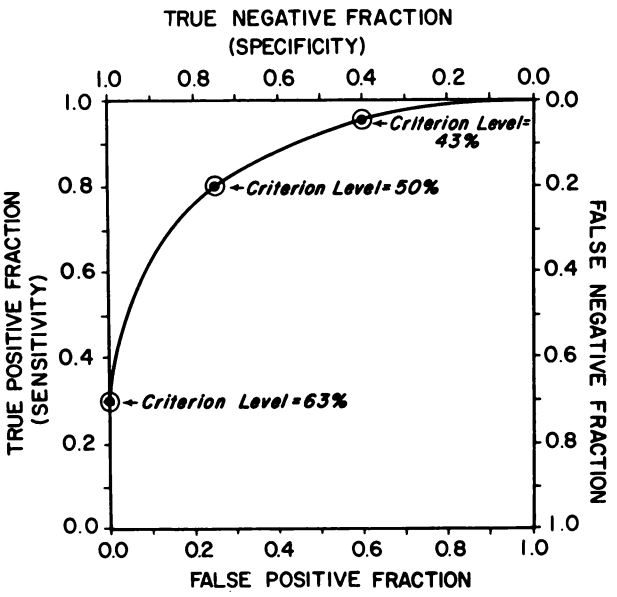

FIG. 3. From text and Fig. 1, it is apparent that both true-positive and false-positive fractions of CTR test increase as criterion level is lowered. In this graph, a hypothetical true-positive fraction has been plotted against a hypothetical false-positive fraction for each of the three criterion levels indicated in Fig. 1. All axes have been labeled to indicate relationships between true-positive fraction, false-negative fraction, false-positive fraction, and true-negative fraction. Smooth curve connecting points indicates that these fractions vary *continuously* as criterion level is changed. As pointed out later in text, this curve is a receiver operating characteristic (ROC) curve.

Whereas the criterion level is directly controlled in tests that have numeric results, it is very difficult to control precisely in subjectively interpreted medical imaging tests. It is the common experience of those who make subjective interpretations of imaging tests that their criterion levels vary somewhat from day to day (intra-observer variation). Furthermore, there may be great variation in the criterion levels used by different observers to decide the "positivity" or "negativity" of a medical imaging test, a major cause of inter-observer variation.

Since the sensivity and specificity of a test change as the criterion level varies, and since the criterion level is difficult to control precisely in subjectively interpreted medical imaging tests, use of these measures to compare medical imaging tests may yield results that are ambiguous or misleading. For example,

suppose two kinds of imaging tests for detecting brain disease were being compared, and the true-positive and false-positive fractions were as listed in Table 1. We would conclude that Test B is more sensitive than Test A. However, this conclusion may be unwarranted. Since more false-positive errors were made by the observer with Test B than with Test A, the "greater sensitivity" of Test B may have been the result of the observer using a relatively lax diagnostic criterion level. If he had been able to adjust the criterion level so that the false-positive fractions for each test were the same, the observed difference in the "sensitivities" of the tests might have disappeared.

The preceding example illustrates the point that no statement can be made about the relative inherent detectability of disease with two diagnostic tests if one appears "more sensitive" but "less specific" than the other. The inherent detectability of disease is clearly greater with one test than another only if it is more specific and more sensitive, equally specific and more sensitive, or equally sensitive and more specific. However, even if one of these situations obtain, it may be difficult to tell how much better the former test is, since a change in the criterion level may cause a large change in sensitivity while causing only a small change in specificity.

It should be apparent from the foregoing discussion that it is especially inappropriate to compare the "sensitivities" of medical imaging tests without regard to their false-positive fractions (or true-negative fractions). For example, this practice could lead one to conclude that diagnosing pneumonia in all patients with fever, cough, or chest pain is better than testing for pneumonia with chest radiography, since the sensitivity of the former practice would be close to 100%! Although this obvious example may seem ludicrous, one doesn't have to search the medical literature at great length to find medical tests advocated on the basis of their "sensitivity" alone.

It should be pointed out that sensitivity and specificity are not the only measures that fail to account for variation in the criterion level. For example, the "overall percentage accuracy" (OPA) (*2*) (i.e. percentage of "correct answers") (overall percentage accuracy = True positives + True negatives/Total no. of tests × 100) is similarly deficient, and should be used only with great care to compare diagnostic imaging tests. (OPA has other problems as well. It is sensitive to the prevalence of disease in the population examined. Furthermore, it may be relatively *insensitive* to changes in lesion detectability, especially when the prevalence of disease is very high or very low.)

In summary, many commonly used measures of observer performance, such as sensitivity and specificity, may be misleading or ambiguous when used for comparing diagnostic imaging tests because they do not account for variability of the criterion level. There is, however, a method that overcomes this problem.

TABLE 1. TRUE- AND FALSE-POSITIVE FRACTIONS FOR TWO HYPOTHETICAL TESTS FOR BRAIN DISEASE

	True-postive fraction	False-positive fraction
Test A	.85	.05
Test B	.95	.10

The receiver operating characteristic curve. We come now to the focal point of our discussion, namely the so called receiver operating characteristic curve. An ROC curve is a way of representing observer performance which—unlike sensitivity, specificity, and overall percentage accuracy—takes variability of the criterion level into account.

Let us suppose that we could continuously vary the criterion level for a diagnostic test. We would then observe continuous, monotonic variation (variation in the same direction) in the true-positive fraction and the false-positive fraction as the criterion level was varied. If we were to plot the changing true- and false-positive fractions against each other in a coordinate space we would obtain a smooth curve (Figs. 3–7), which is referred to as a receiver operating characteristic curve (*3,4*).

In the jargon of signal/detection theory, an ROC curve depicts continuous, monotonic variation of the conditional probability of a true-positive response, p(S|s), and the conditional probability of a false-positive response, p(S|n) as the decision-criterion level is varied. The notation p(S|s) can be read "probability of the observer responding 'signal', given the presence of an actual signal"; p(S|n) is read "the probability of the observer responding 'signal', given noise." p(S|s) is numerically equal to the expected true-positive fraction, and p(S|n) is numerically equal to the expected false-positive fraction.

The inherent detectability of lesions (signals) in an image is related to the position of the ROC curve in the coordinate space; that is to say, the further up and to the left the curve is located, the greater is the inherent detectability of the signal (*4*). For example, Curve A in Fig. 4 indicates greater detectability than Curve B, because no matter what p(S|n) is selected, the corresponding p(S|s) always will be greater with the test represented by Curve A.

ROC curves can be generated experimentally in several ways. For clinical experiments, the most

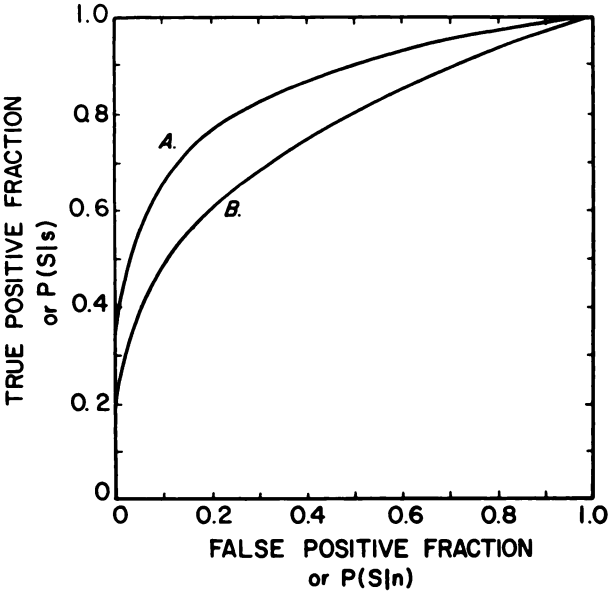

FIG. 4. Curve A indicates greater detectability than curve B.

convenient is the so-called rating method (*3,6,7*), in which the true-positive fractions and false-positvie fractions corresponding to various criterion levels are obtained by having the observer indicate his level of confidence that lesions are present in an image by means of a multi-category rating scale. For example, let us consider an actual experiment (*8*) in which two devices for brain scintigraphy, a scintillation camera and a tomographic scanner, were compared. Brain scintigraphy of a population of patients with and without focal brain disease was performed with both devices. The scintigrams from each device were interpreted independently by three observers, who classified each study as definitely abnormal, probably abnormal, possibly abnormal, probably normal, or definitely normal. In this rating experiment, the five classifications used by the observers imply four criterion levels for calling a scintigram "positive" which distinguish the five classifications. The most strict of these criterion levels is implied by the first category; that is to say, if we call "positive" only those scintigrams classified as "definitely abnormal," both the true-positive and false-positive fractions should be relatively low. The next criterion level is implied by the second category; in other words, the true-positive fraction (sensitivity) will increase if we call "positive" those studies classified as either definitely or probably abnormal, but only at the cost of a higher false-positive fraction. By successively adding the "possibly abnormal" and "probably normal" categories to the "positive" scintigram group, true-positive and false-positive fractions are generated at two progressively more lax criterion levels.

Table 2 lists the true- and false-positive fractions actually generated in the experiment by "Observer 1". Note the confusion that might result from an attempt to compare the tomographic scanner and scintillation camera in terms of single sets of "sensitivity" and "specificity." If we chose to call a scintigraphic study "positive" only if it had been classified as definitely or probably abnormal, we would conclude that the tomographic brain scan is a much more sensitive (62% against 53%) but slightly less specific test for focal brain disease than the camera scintigram. If we chose to include as well those studies classified as "possibly abnormal," however, the sensitivities of the two tests would appear the same, but the tomographic scan would seem more specific! The confusion vanishes as soon as the data are expressed in the form of ROC curves, as shown in Fig. 5. When we take the data from Table 2 and plot each true positive fraction against the corresponding false-positive fraction (Fig. 5) we obtain a series of four "operating points" each for the experiments performed with the tomographic scans and the camera scintigrams. A smooth ROC curve is then drawn through each set of points. We now have

TABLE 2. TRUE- AND FALSE-POSITIVE FRACTIONS AT FOUR CRITERION LEVELS. ONE OBSERVER INTERPRETING CAMERA SCINTIGRAMS OR TOMOGRAPHIC SCANS OF THE BRAIN (*8*)

		Categories considered "positive"			
		Def. abnl.	Def. abnl. Prob. abnl.	Def. abnl. Prob. abnl. Poss. abnl.	Def. abnl. Prob. abnl. Poss. abnl. Prob. nl.
Camera scintigrams	True-positive fraction	.49	.53	.68	.83
	False-positive fraction	.00	.00	.05	.16
Tomographic scans	True-positive fraction	.60	.62	.68	.74
	False-positive fraction	.00	.02	.02	.06

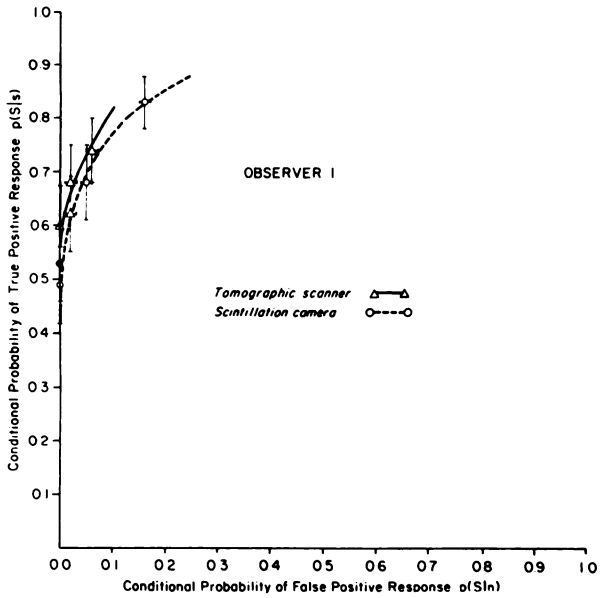

FIG. 5. See description in text. (From reference 8, with permission)

complete descriptions of the relative inherent detectability of lesions with these two systems, because we are comparing the systems over a wide and continuous range of criterion levels. It is clear that the observer performed slightly better reading the tomographic scans than the camera scintigrams, because the expected true-positive fraction is slightly greater for the tomographic scans than for the camera scintigrams over the entire range of false-positive fractions represented.

ROC curves can be conveniently fitted to experimentally determined operating points by plotting the points on double-probability paper and drawing a straight line through the points (*9*). The curve then is transferred to the ROC plot in which the axes are linear with respect to p(S|s) and p(S|n). This procedure is theoretically valid for situations in which the signal and noise distributions are known to be normal or easily transformed into normal distributions. Although the nature of the distributions of signals and noise are unknown in clinical detection situations, it is an empirical fact that most experimentally determined ROC curves can be plotted as nearly straight lines on double-probability paper (*10*).

If the detection task is very simple, requiring no training or experience, the position of an ROC curve will be related mainly to the inherent detectability of the signal in the image. For example, Goodenough and associates conducted an experiment testing the detectability of small lucite beads in radiographs made with four combinations of films and intensifying screens (*3*). Figures 6A and B show ROC curves generated by a physicist and a senior radiologist, respectively. Note that the curves for each film-screen combination occupy approximately the same position in each coordinate space. On the other hand, for more complex tasks, such as the detection of lesions on a brain scintigram, the training and experience of the observer may affect the absolute position of the ROC curves in the coordinate space. Turner and associates conducted an experiment in which patients undergoing brain scintigraphy with a scintillation camera were re-examined with a tomographic scanner if the camera scintigrams were dif-

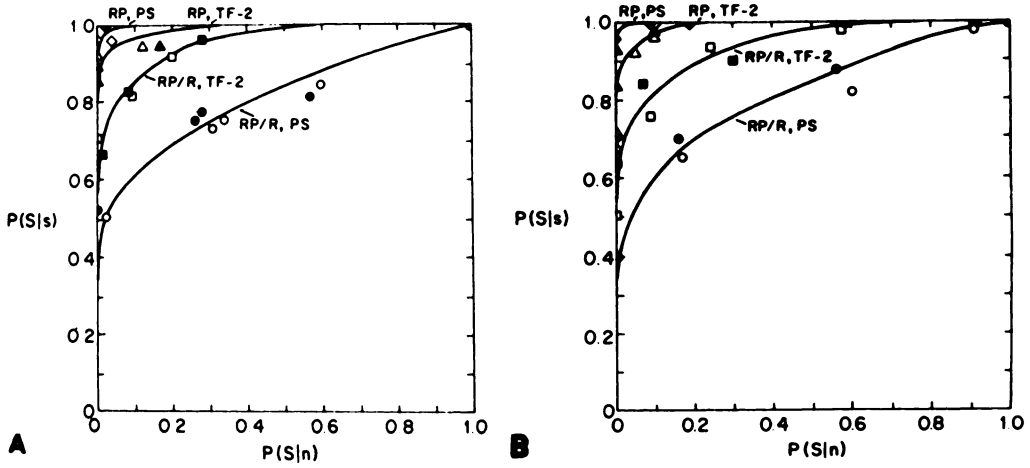

FIG. 6. ROC curves generated by (A) a senior radiologist and (B) a physicist, describing detectability of a 2 mm plexiglass bead in images made with four combinations of radiographic films and screens: RP-Kodak RP X-omat medical X-ray film (normal speed); RP/R-Kodak RP-Royal X-omat medical X-ray film (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 given observer and the given screen-film combination. Each independent trial run consisted of approximately 100 observations. Note reproducibility of curves from observer to observer for this simple detection task. (From C. E. Metz, D. J. Goodenough and K. Rossman, *Radiology* 109: 297–303, 1973, with permission)

ficult to interpret (*11*). The performance of four observers of varying experience reading the camera scintigrams alone, the tomographic scans alone, and the tomographic scans and camera scintigrams together were then compared (Fig. 7). The *relative* positions of the curves for each type of scan read separately and both types read together were the same for all observers. However, the *absolute* positions of the curves show better performance by Observer 1, who had 11 yr of experience in nuclear medicine, than by Observer 4, who had only 3 mo of experience.

In the examples cited above, diagnostic observers generated conventional ROC curves by choosing between two alternatives, namely, disease was either present or absent; they were not required to indicate the number of lesions present or their locations, as is frequently important in clinical diagnostic imaging. Metz and associates (*12*) and Starr and associates (*13*) have shown that the ROC method can be generalized to include these more complex tasks. Furthermore, observer performance in the latter tasks can be predicted from conventional ROC curves.

At the time of this writing, no entirely satisfactory statistical method has been described for quantitatively testing the significance of the separation of ROC curves. It has been customary to fit the experimentally determined operating points with error bars which represent the square root of binomial variance (*14*) (Figs. 5 and 7). A qualitative impression of the significance of curve separation can then be gained by visual inspection of the curves and the associated error bars. The subjective nature of this method leaves something to be desired. Furthermore, the method fails to account for certain variables in experimental design. For example, when an experiment compares two diagnostic tests by applying them to the same patient population, the ROC curves are generated from statistically dependent sets of observations. Since binomial variance fails to account for this, it would seem that the significance of the separation of curves in such an experiment tends to be underestimated. (The underlying assumption is that binomial variance is at least in part due to random errors in the sampling of patient populations.) On the other hand, binomial variance does not appear to account fully for biological or psychological variability of observers, or variability of the imaging systems being tested.

The obviously unsettled state of the art of statistical testing of the separation of ROC curves has been an impediment to general acceptance of ROC analysis. However, it should be remembered that expression of observer performance in terms of sensitivity and specificity for the sake of statistical analysis may yield misleading results: since sensitivity and specificity vary with the criterion level, the cause of statistically significant differences in these parameters may be conscious or unconscious variation in the criterion levels used by observers rather than differences in the inherent detectability of lesions by the two diagnostic tests being compared. (See Table 2 and Fig. 5.)

In spite of the problems discussed above, the ROC method is a powerful tool. It can be used to compare

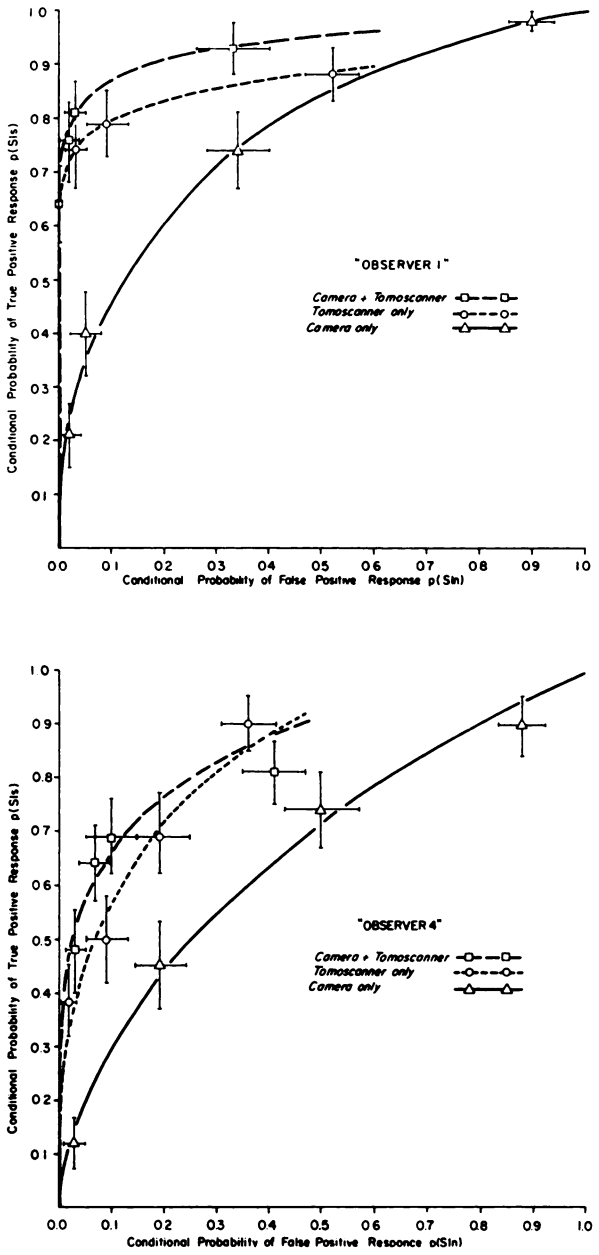

FIG. 7. "Observer 1" had 11 yr of experience in nuclear medicine and "Observer 4" had 3 mo. Note that *relative* position of curves is same for both observers, but absolute positions indicate better performance by more experienced observer. (From reference *11*, with permission.)

diagnostic imaging tests, in terms of observer performance, in a way that avoids confusion due to variation in the diagnostic-criterion level. It can also be used to choose "optimum" diagnostic-criterion levels so as to maximize overall percentage accuracy, average net benefit, or other measures of clinical efficacy (*10*). The ROC method can be employed to demonstrate the clinical efficacy of a diagnostic test by comparing diagnostic performance with and without the test (*15*). Thus, in a broad range of applications, the ROC method can increase the precision with which medical imaging tests are evaluated.

ACKNOWLEDGMENTS

I am indebted to Charles E. Metz, of the Center for Radiologic Image Research (USPHS Grant GM 18940), Department of Radiology, The University of Chicago, whose counsel has added greatly to my understanding of ROC analysis and has been invaluable in the preparation of the manuscript. I am also grateful to Ms. Brenda J. Shead for secretarial assistance.

REFERENCES

1. VECCHIO TJ: Predictive value of a single diagnostic test in unselected populations. *N Engl J Med* 274: 1171–1173, 1966

2. NISHIYAMA H, LEWIS JT, ASHARE AB, et al: Interpretation of radionuclide liver images: Do training and experience make a difference? *J Nucl Med* 16: 11–16, 1975

3. GOODENOUGH DJ, ROSSMANN K, LUSTED LB: Radiographic applications of receiver operating characteristic (ROC) curves. *Radiology* 110: 89–95, 1974

4. GREEN DM, SWETS JA: *Signal Detection Theory and Psychophysics (reprint with corrections).* Huntington, N.Y., Krieger, 1974, pp 30–52

5. MEZAROS WT: *Cardiac Roentgenology.* Springfield, Charles C. Thomas, 1969, p 54

6. GREEN DM, SWETS JA: *Op cit.*, pp 99–107

7. METZ CE, GOODENOUGH DJ: Quantitative evaluation of human visual detection performance using empirical receiver operating characteristic curves. In *Information Processing in Scintigraphy.* Metz CE, Pizer SM, Brownell GL, eds. Oak Ridge, USERDA Technical Information Center, 1975, Publication CONF-730687, pp 140–152

8. TURNER DA, RAMACHANDRAN PC, ALI AA, et al: Brain scanning with the Anger multiplane tomographic scanner as a primary examination. Evaluation by the ROC method. *Radiology* 121: 125–129, 1976

9. GREEN DM, SWETS JA: *Op cit.*, pp 61–62

10. METZ CE, STARR SJ, LUSTED LB: Quantitative evaluation of visual detection performance in medicine: ROC analysis and determination of diagnostic benefit. In *Medical Images.* Hay GA, ed. New York, John Wiley and Sons, 1977, pp 220–241

11. TURNER DA, FORDHAM EW, PAGANO JV, et al: Brain scanning with the Anger multiplane tomographic scanner as a second examination. Evaluation by the ROC method. *Radiology* 121: 115–124, 1976

12. METZ CE, STARR SJ, LUSTED LB, et al: Observer performance in detecting multiple radiographic signals. *Radiology* 121: 337–347, 1976

13. STARR SJ, METZ CE, LUSTED LB, et al: Visual detection and localization of radiographic images. Prediction and analysis using a generalized ROC approach. *Radiology* 116: 533–538, 1975

14. GREEN DM, SWETS JA: *Op cit.*, pp 401–404

15. MCNEIL BJ, HESSEL SJ, BRANCH WT: Measures of clinical efficacy. III. The value of the lung scan in the evaluation of young patients with pleuritic chest pain. *J Nucl Med* 17: 163–169, 1976