Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems

Charles E. Metz, PhD

Receiver operating characteristic (ROC) analysis provides the most comprehensive description of diagnostic accuracy available to date, because it estimates and reports all of the combinations of sensitivity and specificity that a diagnostic test is able to provide. After sketching the 6 levels at which diagnostic efficacy can be assessed, this paper explains the conceptual foundations of conventional ROC analysis, describes a variety of indices that can be used to summarize ROC curves, and describes several forms of generalized ROC analysis that address situations in which more than 2 decision alternatives are available. Key issues that arise in ROC curve fitting and statistical testing are addressed in an intuitive way to provide a basis for judging the validity of ROC studies reported in the literature. A list of sources for free ROC software is provided. Receiver operating characteristic methodology has reached a level of maturity at which it can be recommended broadly as the approach of choice for radiologic imaging system comparisons.

Key Words: Efficacy, evaluation, medical imaging, receiver operating characteristic analysis, ROC

J Am Coll Radiol 2006;3:413-422. Copyright © 2006 American College of Radiology

INTRODUCTION[1]

A hierarchical model for diagnostic efficacy[2] developed by a scientific committee of the National Council on Radiation Protection and Measurements [2] provides a concise conceptual overview of the issues involved in evaluating diagnostic systems. This model's 6 levels are as follows:

(1) Technical efficacy: at the model's lowest level, a diagnostic test is considered effective if its result is accurate and precise in a physical sense, for example, if the test measures 1 or more physical properties of the human body in a way that agrees with a "gold standard," and its results are reproducible. Aspects of technical efficacy in medical imaging include spatial or temporal resolution, noise magnitude and texture, and contrast sensitivity.

(2) Diagnostic accuracy: the second level of efficacy concerns the extent to which the results of a diagnostic test agree, in some statistical sense, with patients' actual states of health or disease. Virtually all practical measures of diagnostic accuracy quantify the ability of a test to distinguish between 2 (usually composite) states of truth, such as "normal" vs "abnormal" or "positive" vs "negative," with respect to a specified disease. Examples of diagnostic-accuracy measures include percentage correct, sensitivity and specificity, and receiver operating characteristic (ROC) curves. Of these, ROC curves provide the most comprehensive description, because they indicate all of the combinations of sensitivity and specificity that a diagnostic test is able to provide as the test's "decision criterion" is varied.

(3) Diagnostic-thinking efficacy: given the prevalence of a particular disease and given the sensitivity and specificity (or more generally the ROC curve) of a diagnostic test for the presence of that disease, one can easily compute the factor by which the prior odds of disease change after the test's result is obtained. However, the extent to which a diagnostic test affects physicians' subjective estimates of disease likelihood must be answered empirically. This level of efficacy is sometimes difficult to quantify, but it

Department of Radiology, University of Chicago, Chicago, Ill.

Corresponding author and reprints: Charles E. Metz, PhD, University of Chicago, Department of Radiology, MC 2026, 5841 South Maryland Avenue, Chicago, IL 60637; e-mail: c-metz@uchicago.edu.

[1] Parts of this section are adapted from an earlier tutorial review [1].

[2] According to the words' strict definitions, *efficacy* pertains to ideal conditions, whereas *effectiveness* refers to routine practice. I use the terms synonymously here for simplicity.

0091-2182/06/$32.00 • DOI 10.1016/j.jacr.2006.02.021

provides a conceptual link between the more easily interpreted tiers above and below.

(4) Therapeutic efficacy: this is the lowest level at which the effects of a diagnostic test on patient management are assessed directly. The basic question is how and by how much a particular diagnostic test changes the way in which patients are treated; for example, how does therapy differ when it is chosen without or with knowledge of a test's result?

(5) Patient-outcome efficacy: here, the goal of diagnostic medicine is confronted directly: a diagnostic test is considered effective at this level only if patient health (as measured, eg, in "quality-adjusted life years") is demonstrably improved by use of the test. This is the kind of efficacy that is of greatest interest to most patients and physicians, and it is an indispensable component of any meaningful "cost-benefit" or "cost-effectiveness" analysis. However, a definitive assessment of efficacy at this level requires prospective randomized and controlled clinical trials, in which practical, statistical, and ethical problems can be formidable [2].

(6) Societal efficacy: any cost-benefit or cost-effectiveness analysis of a diagnostic test at level 5 focuses on the benefits and personal risks that accrue to the patients who are candidates for the test. However, the fact that medical costs are borne increasingly by society as a whole implies that social utilities should somehow be taken into account when benefits and costs are evaluated. This is the domain of "societal efficacy," which in principle merges private and public considerations to assess diagnostic tests within the context of the social endeavor.

Efficacy at this hierarchical model's upper levels usually is of greatest direct interest, but lower-level efficacy is almost always easier to quantify reliably. Fortunately, efficacy at the higher levels sometimes can be estimated from measurements at lower levels by the use of collateral data and appropriate assumptions. Most studies of diagnostic efficacy in medical imaging focus on the measurement of diagnostic accuracy (level 2), because this is the lowest level at which human observers are included and often the highest level at which scientifically rigorous methods can be used.

For many years, diagnostic accuracy was measured and reported in terms of a kind of "batting average": the percentage of diagnostic decisions that proved to be correct. This "percentage-correct" measure has the fairly obvious limitation that it can depend strongly on disease prevalence [3]: if only 1% of the patients in a screening population have a particular disease, for example, then a system can be "99% accurate" simply by blindly calling all patients negative with respect to that disease. Moreover, the percentage-correct measure does not reveal the relative frequencies of false-positive and false-negative errors, which usually have substantially different clinical consequences.

Both of these disadvantages are overcome if diagnostic performance is reported in terms of a pair of indices: "sensitivity" (the fraction of patients actually having the disease in question who are correctly diagnosed as positive) and "specificity" (the fraction of patients actually without the disease who are correctly diagnosed as negative). In effect, these indices quantify separately the "accuracies" of the system for actually positive and actually negative patients, respectively. False-negative and false-positive diagnoses are accounted for implicitly by these indices, and a change in disease prevalence does not affect their numerical values if constant decision criteria are used. The terms *true-positive fraction* (TPF) and *true-negative fraction* are synonymous with *sensitivity* and *specificity*, respectively. In a complementary way, the "false-negative fraction" and the "false-positive fraction" (FPF) represent the conditional probabilities or frequencies with which actually positive and actually negative patients are diagnosed incorrectly [3,4]; thus, false-negative fraction = 1 − TPF = 1 − sensitivity, and FPF = 1 − true-negative fraction = 1 − specificity. Because of the interrelationships among these measures, it is necessary only to indicate a single pair; conventionally, either sensitivity and specificity or TPF and FPF are used. The use of sensitivity or TPF alone is inadequate, because the performance of the diagnostic system with regard to actually negative patients is then unknown.

The sensitivity-specificity pair, or one of its equivalents, describes diagnostic accuracy more meaningfully than the single index of percentage correct, and it has been used widely in the medical literature. A single pair of numbers representing sensitivity and specificity is not entirely adequate, however, because it confounds 2 aspects of diagnostic accuracy that can vary independently: (a) the inherent capacity of a diagnostic system to distinguish between actual states of health and disease, and (2) the balance between the frequencies of false-positive and false-negative errors that a decision maker chooses to adopt in a clinical task when a particular discrimination capacity is available [5].

The limitations of reporting diagnostic accuracy in terms of a single sensitivity-specificity or TPF-FPF pair are most evident in studies that attempt to compare diagnostic tests, because often, one test is found to have higher sensitivity (higher TPF) but lower specificity (higher FPF) than the other; in other words, one test is more accurate for actually positive patients, whereas the other is more accurate for actually negative patients. This is clearly problematic in deciding which of the 2 tests to use, because diagnostic testing would not be needed if the presence or absence of the disease were known.

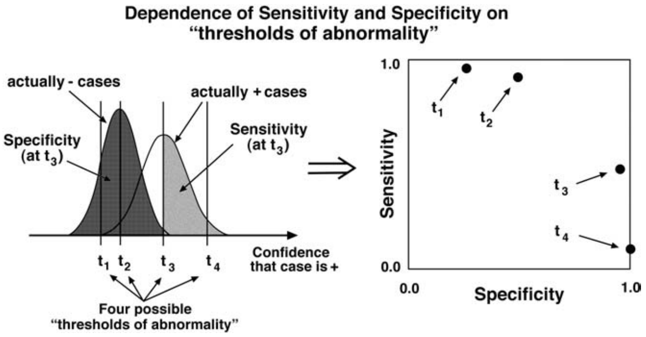

Fig 1. The effect of decision-threshold setting on sensitivity and specificity. With a decision threshold setting at t_3 on the decision axis in the left panel, the pair of distributions shown there produces a combination of sensitivity (light gray shading) and specificity (dark gray shading) that plots as the point labeled “t_3” on the graph of sensitivity vs specificity in the right panel of the figure. Similarly, the decision thresholds labeled “t_1,” “t_2,” and “t_4” in the left panel yield the combinations of sensitivity and specificity that are labeled correspondingly in the right panel.

THE BASIC CONCEPTS OF ROC ANALYSIS

The dilemma that arises when one diagnostic test has higher sensitivity but lower specificity than another can be resolved by noting that the sensitivity and specificity of virtually all diagnostic tests can be changed by modifying the “threshold of abnormality” or “decision threshold” that is used for the test. For example, consider 2 radiologists with equal skill who read mammograms to detect breast cancer. If one of these radiologists reads the images more aggressively than the other, recommending biopsy if a mammogram indicates even the slightest possibility of abnormality, that radiologist will detect malignancy more often than the other but will generate more negative biopsy results as well. This inverse relationship between sensitivity and specificity is demonstrated in Figure 1, in which the 2 bell-shaped curves in the left panel represent schematically the distributions of a radiologist’s confidence that an imaging study indicates the disease in question to be present for actually negative and actually positive patients, respectively. If the radiologist were to recommend patients for biopsy if and only if his or her confidence in a positive interpretation exceeded the decision threshold denoted by the vertical line at t_3 in this panel, the radiologist’s sensitivity would be the fraction of the area under the “actually +” distribution that lies to the right of t_3 (approximately 50% here), whereas his or her specificity would be the fraction of the area under the “actually −” distribution that lies to the left of t_3 (approximately 95% in Figure 1). This combination of sensitivity and specificity can be plotted as a point on a graph that indicates specificity on its horizontal axis and sensitivity on its vertical axis, as shown by the point labeled “t_3” in the right panel of Figure 1. Similarly, the decision thresholds labeled t_1, t_2, and t_4 in the left panel of Figure 1 would yield the combinations of sensitivity and specificity that are labeled correspondingly in the right panel of the figure.

By considering a much larger number of possible settings of the decision threshold t, one would generate a much larger number of points in the right panel of Figure 1, in the limit producing a continuous curve such as that shown on the left side of Figure 2, which indicates all of the combinations of sensitivity and specificity that this radiologist can obtain. For historical reasons, the relationship considered here is usually presented by plotting sensitivity (ie, TPF) as a function of 1 − specificity (ie, FPF) rather than as a function of specificity itself, thereby producing an ROC curve such as that shown on the right side of Figure 2. Perhaps confusingly, “operating points” (ie, combinations of FPF and TPF) near the lower-left end of an ROC curve are obtained by setting the decision threshold at a position toward the right end of the decision axis shown in Figure 1, thereby producing few false-positive decisions but relatively few true-positive decisions as well, whereas operating points near the upper-right end of the ROC curve are obtained by setting the decision threshold toward the left end of the decision axis, thereby producing more true-positive decisions but also more false-positive decisions. Thoughtful consideration of these figures reveals that ROC curves higher or lower than the curves shown in Figure 2 correspond to pairs of underlying distributions that are less overlapped or more overlapped, respectively, than those shown in Figure 1. It is important to note that the extent to which the pair of underlying distributions overlaps, and hence the height of the ROC curve, depends, in general, not only on

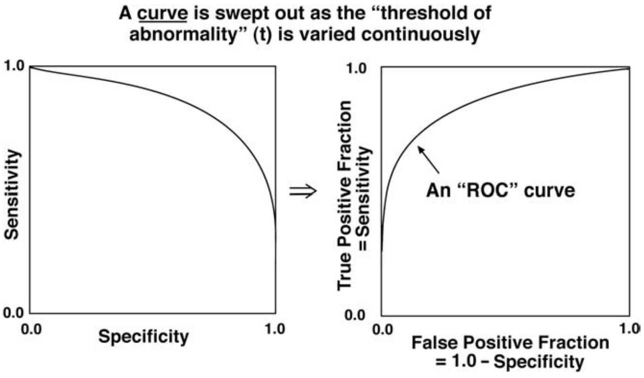

Fig 2. The left panel shows the continuous relationship between sensitivity and specificity that is obtained from Figure 1 by varying the decision threshold continuously. The right panel, which represents a typical receiver operating characteristic (ROC) curve, is a mirror image of the left because it plots sensitivity vs 1 − specificity.

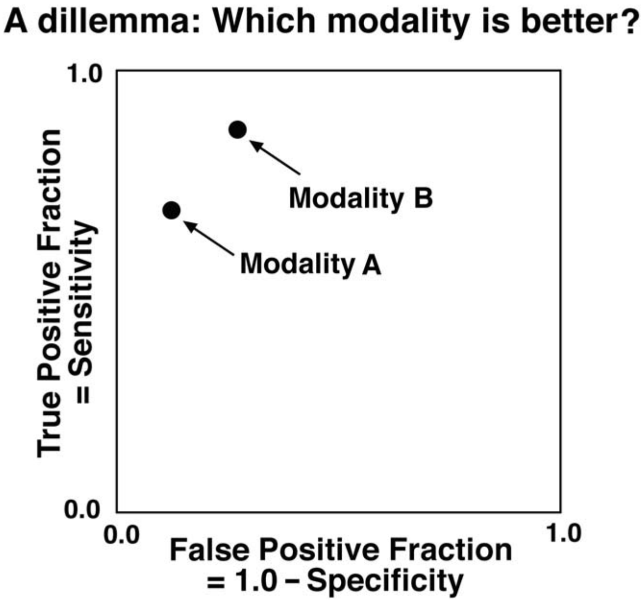

Fig 3. A dilemma in comparing the diagnostic accuracies of 2 imaging systems when only true-positive fraction (or sensitivity) and false-positive fraction (or specificity) are known.

the skill of the radiologist but also on technical aspects of the imaging procedure, such as spatial resolution, noise and contrast, and characteristics of the case population, such as its balance between early and advanced disease among actually positive patients.

With these concepts in mind, one can resolve the dilemma mentioned in "Introduction" and shown in Figure 3, in which diagnostic imaging modality B has a higher TPF (higher sensitivity) but also a higher FPF (ie, lower specificity) than modality A.

Suppose that the ROC curves for modality A and modality B have been determined and are found to look like those shown in Figure 4. In this instance, one can conclude that modality B provides greater diagnostic accuracy, because it can be used to achieve a higher TPF than modality A at the same FPF if the decision threshold for modality B is moved toward the right end of the decision axis in Figure 1 until the FPF for modality B equals that of modality A (ie, operating point B in Figure 4 moves down and to the left along its ROC curve until it lies over point A) or if the decision threshold for modality A is moved toward the left end of the decision axis in Figure 1 until the FPF for modality A equals that of modality B (ie, operating point A in Figure 4 moves up and to the right along its ROC curve until it lies under point B). Given the ROC curves shown in Figure 4, modality B also can be used to achieve a lower FPF than modality A at the same TPF if the decision threshold for modality B is moved toward the right end of the decision axis in Figure 1 until the TPF for modality B equals that of modality A (ie, operating point B in Figure 4 moves down and to the left along its ROC curve until it lies to the left of point A) or if the decision threshold for modality A is moved toward the left end of the decision axis in Figure 1 until the TPF for modality A equals that of modality B in Figure 4 (ie, operating point A in Figure 4 moves up and to the right along its ROC curve until it lies to the right of point B). This conclusion is not inevitable given only the 2 operating points shown in Figure 3, however, because determination of the ROC curves for modalities A and B could yield the result shown in Figure 5, which indicates that the 2 modalities provide equal diagnostic accuracy because their operating points can be made to coincide by appropriate adjustment of either modality's decision-threshold setting. Moreover, in a third scenario not depicted here, one would find modality A to be superior if its ROC curve is higher everywhere than the ROC curve for modality B, which can occur if the curve for modality A rises more steeply from the lower-left corner of the unit square than the curve for modality B rises there. In short, higher ROC curves indicate greater diagnostic accuracy, and one diagnostic test or imaging modality can be judged unambiguously superior to another if its ROC curve lies above the other test's ROC everywhere in the unit square. A comparison of 2 diagnostic tests is more complicated when their ROC curves cross, however. The latter situation is addressed next.

ROC-RELATED INDICES OF DIAGNOSTIC ACCURACY

Experience indicates that real-world ROC curves must be described by at least 2 parameters that represent separately each curve's height and the strength or weakness of its symmetry around the negative (−45°) diagonal of the unit

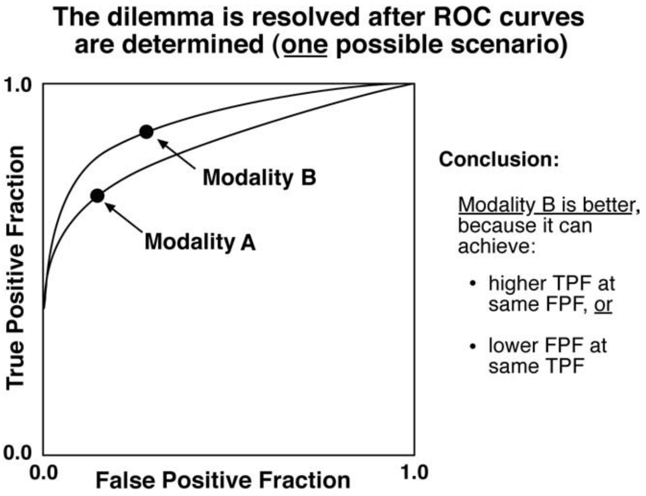

Fig 4. The dilemma is resolved if the receiver operating characteristic (ROC) curve of each imaging system is determined (one scenario). FPF = false-positive fraction; TPF = true-positive fraction.

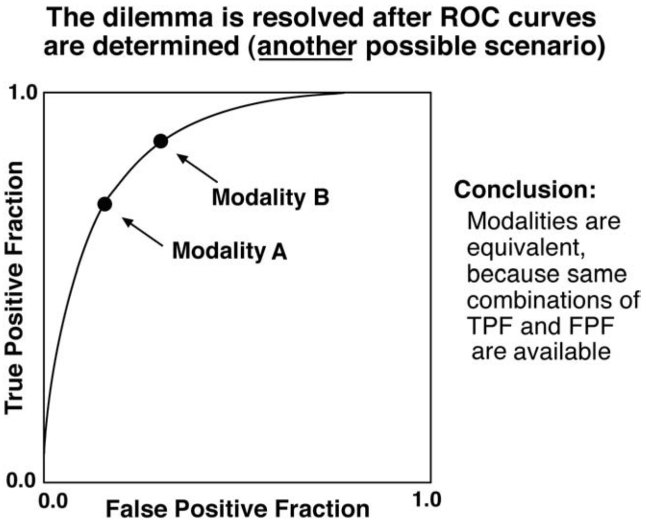

Fig 5. The dilemma is resolved if the receiver operating characteristic (ROC) curve of each imaging system is determined (a second scenario). FPF = false-positive fraction; TPF = true-positive fraction.

square in which ROC curves are plotted. Therefore, ROC curves usually cannot be summarized fully by a single number, and system comparisons that are based entirely on a single-number summary index may lead to erroneous conclusions. The concepts of “better” and “worse” are inherently 1-dimensional, however, so any requirement that systems be ranked in a comparison study dictates the use of some single-number summary index. The key need in such situations is to choose an index that is as meaningful and reliable as possible in a practical sense. Several candidate indices of this kind are described below.

If 2 ROC curves in a comparison do not cross, as in Figure 4, the diagnostic accuracies indicated by those curves can be ranked unambiguously in terms of the total area under each curve within the unit square. This widely used area index, denoted by “AUC” in general or by “A_z” when an ROC curve is described by the standard binormal model discussed later, can be interpreted in 3 ways: as sensitivity averaged over all possible specificities (ie, average ROC curve height); as specificity averaged over all possible sensitivities (ie, average ROC curve “leftness”); or, less obviously, as the probability of identifying correctly the actually positive case in a “2-alternative forced-choice” experiment, in which each trial requires an observer to distinguish between a randomly selected actually positive case and a randomly selected actually negative case. Values of the area index range from .50 for the lowest possible ROC curve—the positive (ie, +45°) diagonal of the unit square, on which the probability of a positive reading is the same for each case, regardless of truth—to 1.00 for the highest possible ROC, which climbs along the left side of the unit square and then becomes horizontal when it reaches the square’s upper-left corner.

The conventional area index described above is “global” in the sense that it summarizes an entire ROC curve. However, this index can be misleading when 2 ROC curves cross, because then it can have the same value for both curves even though one ROC is higher for some range of FPF, or, even more dangerously, its value can be lower for the curve that is higher in the FPF range of primary clinical interest. Two kinds of “partial-area” indices have been proposed to address this limitation: one represents the area under an ROC curve within a selected range of FPF [6], whereas the other represents the area to the right of the ROC within a selected range of TPF [7]. For a given ROC curve, the value of each of these “regional” indices depends on the range of FPF or TPF that is used, which must be indicated explicitly when the index value is reported and should be chosen on the basis of clinical considerations. Both normalized and nonnormalized variants of these partial-area indices have been suggested in the literature, so any study that uses such an index to summarize its ROC curves must indicate precisely how the index was calculated.

Partial-area indices are based on a range of FPF or TPF to focus a comparison on clinically relevant sensitivities and specificities without presuming to judge the precise balance between sensitivity and specificity that is most appropriate in a particular clinical situation. However, other ROC-based indices of diagnostic accuracy can be used to focus the comparison more narrowly if that is desired. The simplest of these “local” indices are TPF at a preselected value of FPF and FPF at a preselected value of TPF, which summarize differences between ROC curves in terms of a single difference in TPF or FPF, respectively. Despite the intuitive and computational advantages that their simplicity provide, these indices are not used widely because they suffer from important practical and theoretical limitations: physicians usually cannot agree on the single value of FPF or TPF at which diagnostic tests should be compared, even in a narrowly defined clinical task, and decision theory indicates that ROC curves should not be compared at a single value of FPF or TPF but instead at a single critical likelihood ratio, which corresponds to a particular value of ROC curve slope. The latter limitation can be overcome by considering the relative benefits and costs of true-positive, true-negative, false-positive, and false-negative decisions; by noting that the balance among the frequencies of these 4 kinds of decisions changes as a diagnostic test’s operating point is moved along its ROC curve; and by then comparing ROC curves in terms of the maximum “expected net benefit” or “expected utility” that can be achieved on each curve [3,8]. However, this theoretically attractive local index is rarely used because of the difficulty of achieving consensus regarding the relative benefits and costs of the different kinds of correct and incorrect decisions.

GENERALIZED ROC ANALYSIS

Conventional ROC analysis fully describes all of the tradeoffs that a particular human or automated decision maker can achieve among the frequencies of true-positive, true-negative, false-positive, and false-negative decisions in any particular 2-group classification task, that is, in any situation in which only 2 states of truth are relevant, in which the decision maker must decide to which of the 2 states each test case belongs, and in which a population of test cases has been defined and sampled. The terms *positive* and *negative* are used abstractly here to denote any state (eg, disease X is present) and its complement (disease X is absent); in practice, these 2 complementary groups can be defined in any desired way as long as the true group membership of each test case can be determined by some gold standard (eg, biopsy) to provide a basis for scoring decision performance. In medical imaging applications of conventional ROC analysis, the 2 groups usually are composite in the sense that each group includes a variety of states; in a task that involves the detection of disease X, for example, the patients who actually have disease X can include both early and late examples of the disease, whereas the patients who in fact are without disease X can include a variety of other diseases that show symptoms similar to those of the disease in question. One should note that the ROC curves obtained in such situations usually depend on the spectrum of states within each defined group of the population, both of which must be represented adequately in a study's test-case sample to obtain a reliable estimate of diagnostic accuracy. Perhaps the most extreme example of composite groups in clinical 2-group classification tasks is that of distinguishing between normal and abnormal patients, which is conceptually valid but usually impractical because of the difficult sampling issues involved. The extent to which a 2-group evaluation of diagnostic accuracy is clinically relevant depends in part on the particular chore at hand (eg, is a radiologist's task considered one of determining dichotomously the presence or absence of a candidate disease state or instead one of differential diagnosis?) and in part on the particular role an imaging procedure is thought of as playing in patient workup (eg, is a radiologist's task in mammography to detect any breast mass for the purpose of deciding which patients to send to needle biopsy or instead to distinguish between the presence and absence of malignant breast masses?).

Although diagnostic decision making, especially in disease detection, often can be modeled as a sequence of 2-group classification tasks and evaluated by conventional ROC analysis on that basis, one aspect of a radiologist's role that cannot be modeled as a 2-group classification task, even in disease detection, is the joint task of reporting the location or locations of a particular disease as well as its presence. For example, conventional ROC analysis cannot be used directly to quantify a diagnostic system's ability to direct surgery to the particular site at which a single tumor is present or to distinguish among the absence of any tumor, the presence of a single tumor, and the presence of multiple tumors in each case. Moreover, conventional ROC analysis cannot be used to quantify classification performance in tasks where patients must be sorted into 3 or more groups simultaneously (eg, no breast mass, benign breast mass, or malignant breast mass) rather than sequentially (eg, a decision of no breast mass or some breast mass followed by a decision of benign breast mass or malignant breast mass, if some breast mass were detected at the first stage). Several generalized forms of ROC analysis that have been proposed to meet these needs are described briefly here to acquaint the reader with the distinctions among them. Although these generalized techniques involve a variety of subtle methodologic issues, some of which have not been fully resolved, their increasing use in evaluation studies demands an understanding of the advantages and limitations of each.

Localization ROC (LROC) analysis, the first generalized form of ROC methodology to appear in the radiologic literature [9], applies to detection tasks in which each image contains either no lesion or 1 lesion, and the radiologist is required to specify the single location, if any, at which a lesion is judged to be present. Data are collected in terms of a single confidence rating for each case that expresses the likelihood of a lesion being present, as in conventional ROC analysis, plus a location report that can be expressed either in terms of a grid (eg, image quadrant) or geometrical position (eg, image coordinates). A curve is then plotted from the confidence-rating data in a way essentially similar to the method used for plotting a conventional ROC curve, except that the decision maker is given credit for a "true-positive" report only when the reported location is sufficiently close to the lesion's true position in an actually positive image. Actually negative images are scored in the same way as in conventional ROC analysis, so the horizontal axis of the generalized curve is the same as that of a conventional ROC curve (Figure 6), indicating the probability of reporting a lesion at any location in an actually negative image.

However, the vertical axis of the generalized "LROC curve" (Figure 7) that results from the stricter scoring rule for actually positive images represents the probability of not only correctly detecting but also correctly locating a lesion when it is actually present. Therefore, the operating point on the LROC curve at each value of FPF $>$ 0 is at least slightly lower than the point at the same FPF on the corresponding conventional ROC curve, which can be computed from the same data simply by ignoring the decision maker's location reports. One should note from Figures 6 and 7 that LROC curves are plotted in a unit square, like conventional ROC curves, but in general do not enter the upper-right corner of that square, because decision makers who are scored by LROC analysis cannot achieve the generalized equivalent of TPF $=$ 1.0 merely by calling all im-

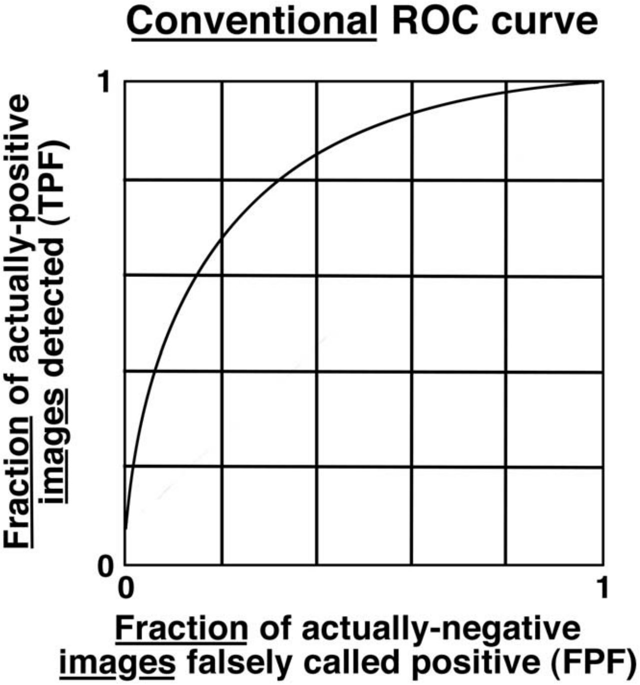

Fig 6. A conventional receiver operating characteristic (ROC) curve. The decision maker's task is to distinguish actually positive images or cases from actually positive images or cases; no localization of lesions is required. FPF = false-positive fraction; TPF = true-positive fraction.

ages positive (ie, by operating at FPF = 1.0). One should note also that the result of any LROC study depends on the amount of location error that is allowed by the data analyst, because this tolerance determines whether a rating above the decision threshold on an actually positive image is scored as a true-positive or a false-negative. A straightforward theoretical relationship between an LROC curve and its corresponding conventional ROC curve [9] for any given location-error tolerance indicates that comparisons based on LROC and conventional ROC analysis can be expected to rank imaging systems in the same way, thereby providing some assurance that conventional ROC methodology will not produce misleading results in comparison studies even though it does not assess location accuracy. However, other considerations indicate that the location data acquired in an LROC study provide greater statistical power than conventional ROC analysis [10].

Free-response operating characteristic (FROC) analysis [11,12] can be thought of as a generalization of LROC methodology that accommodates the possibility of more than 1 lesion per image by allowing multiple reports on each image. Data are collected in terms of a confidence rating regarding the possible presence of a lesion at each location that the decision maker considers worthy of mention. By considering various possible settings of the decision threshold that distinguishes positive from negative reports, a curve that shows the relationship between fraction of lesions detected and number of false-positive reports per image (Figure 8) is then plotted from the confidence-rating and location data. One should note that the horizontal axis of an FROC curve is not normalized to a fraction with a maximum value of 1.0; instead, it must extend to an arbitrarily large number of false-positive reports per image because the number of candidate locations for false-positive reports is unknown. As in LROC analysis, the location data acquired in an FROC study provide potentially greater statistical power than conventional ROC analysis, but the study's result depends on the amount of location error that is allowed by the data analyst. Moreover, in some situations, FROC curves have been found empirically to depend strongly on the "satisfaction-of-search effect" [13], which causes observers to limit their numbers of false-positive reports in a subjective and unpredictable way that is not related to lesion detectability.

Alternative FROC (AFROC) analysis [14] is a variant of FROC analysis that uses a different measure of false-positive rate to reduce the impact of satisfaction-of-search effects on study results. An AFROC curve (Figure 9) uses the same vertical axis as an FROC curve, but its horizontal axis indicates the fraction of images containing 1 or more false-positive reports at each level of lesion detectability, thereby eliminating any dependence of the curve on variations in an observer's subjective tendency to refrain from mentioning

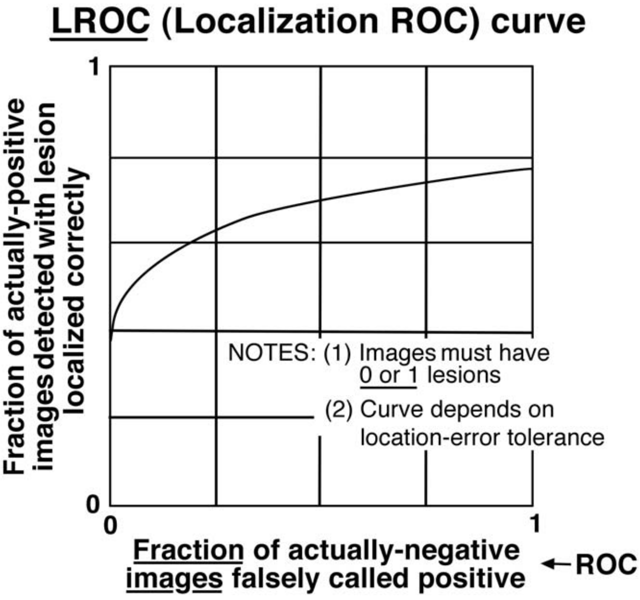

Fig 7. A localization receiver operating characteristic (LROC) curve, which is similar to a conventional receiver operating characteristic (ROC) curve except that each actually positive image or case must contain exactly 1 lesion, and that lesion must be located correctly to be scored as a true-positive detection.

locations that are less salient than several others on a given image. As in LROC and FROC analysis, the location data acquired in an AFROC study provide potentially greater statistical power than conventional ROC analysis [15], but the study's result depends on the amount of location error that is allowed by the data analyst.

Less progress has been made in developing generalized ROC methods that quantify decision performance in classification tasks that involve more than 2 groups. Recent work has shown that generalizing conventional ROC analysis to accommodate multiple-group classification potentially involves an enormous increase in complexity, requiring a 5-dimensional ROC surface in 6 dimensions to provide a universally applicable description of classification performance even for as few as 3 groups [16], and that previously proposed approaches to this problem are severely limited [17].

CURVE FITTING

Curve fitting in any branch of science involves 3 steps: (1) choosing a family of fitting functions with adjustable parameters that is able to summarize the data of interest with adequate fidelity; (2) adopting a measure that quantifies the goodness of any particular fit; and (3) computing the values of the fitting function's parameters that provide the best fit, according to the adopted measure.

Conventional ROC curves are usually fit by using the so-called binormal model, which assumes that the true curve shape is one of those produced by a pair of normal (Gaussian) decision-variable distributions [18].[3] Extensive experience indicates that the standard binormal model provides excellent fits to statistically reliable ROC data in a very broad range of applications [19,20]. However, this model can yield fitted curves with inappropriate shapes when data sets are weak (eg, are based on a small number of cases or contain ratings in only a small number of discrete categories) [21]. Therefore, several "proper" models for fitting conventional ROC curves have been proposed to deal with such situations [10, 22-24].[4] Of these, the proper binormal model [23,24] seems to be the most robust and has the advantage that its fits are essentially identical to those of the widely accepted standard binormal model when the latter produces a fit with appropriate shape. Localization ROC curves can be fit by use of a model proposed by Swennson

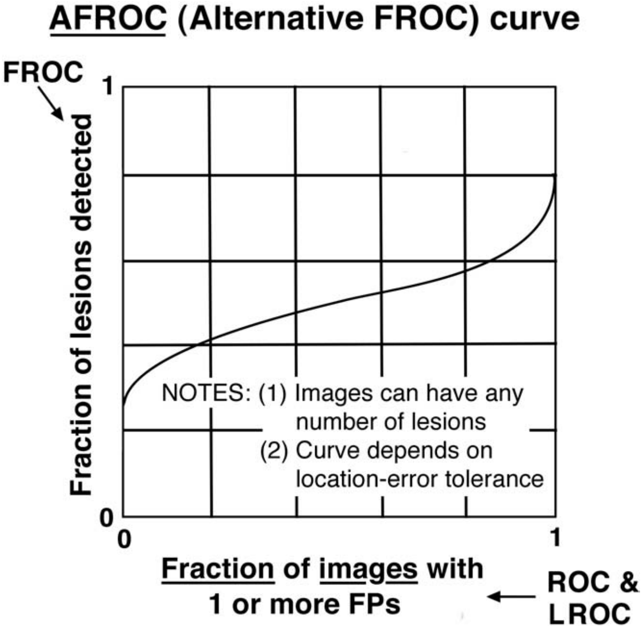

Fig 9. An alternative free-response operating characteristic (AFROC) curve, which is similar to a free-response operating characteristic curve except for its horizontal axis. FP = false-positive; LROC = location receiver operating characteristic; ROC = receiver operating characteristic.

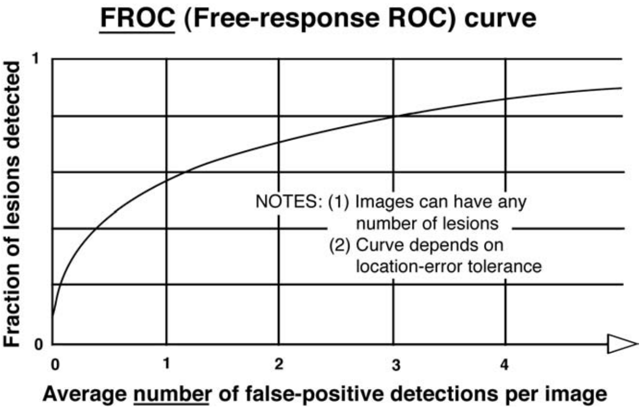

Fig 8. A free-response operating characteristic (FROC) curve. Each image or case can contain any number of lesions. Each correctly located true-positive detection and each false-positive location report is scored independently. The horizontal axis cannot be normalized to range from 0.0 to 1.0, because the maximum possible number of false-positives on each image or case is unknown.

[3] Real-world decision variables often are not normally distributed, so it is important to understand that this assumption is much less restrictive than it might seem. Any monotonic transformation of the decision variable that underlies an ROC curve changes the shapes of the decision-variable distributions but does not affect the ROC, which is determined entirely by the ranks (ie, the relative numerical order) of decision-variable outcomes. Therefore, the assumption concerning ROC curve shape made in adopting the binormal model is equivalent to a rather loose assumption that some unknown monotonic transformation of the real-world decision variable would yield a pair of approximately normal decision-variable distributions.

[4] A proper ROC curve is one on which slope decreases steadily as the operating point moves from FPF = 0.0 to FPF = 1.0.

[10], whereas experience to date indicates that the standard binormal model can be used to fit AFROC curves [14]. As yet, no reliable model for fitting FROC curves has been found, probably because of the variable impact of satisfaction-of-search effects on FROC curves.

Fitting conventional or generalized ROC curves by standard least squares techniques is inappropriate for a variety of reasons; instead, maximum likelihood should be used to account correctly for the statistical properties of confidence-rating data. Free software is available for this purpose, as noted below.

TESTING THE STATISTICAL SIGNIFICANCE OF DIFFERENCES

Differences between estimated ROC curves must be subjected to statistical significance testing to determine whether those differences can be ascribed to random variation or, instead, are likely to be real. A sometimes bewildering variety of statistical tests has been developed to assess the significance of differences between ROC curve estimates. Although the details of those techniques are beyond the scope of this paper, 2 fundamental issues are mentioned briefly here to provide a basis for choosing an appropriate statistical test when the need arises or for judging the validity of ROC studies reported in the literature.

One distinction among currently available statistical tests is the sense in which "difference" is quantified: in terms of the entire ROC curves or in terms of a particular global, regional, or local index. Good statistical practice requires that a clinically relevant definition of difference be chosen before any significance calculation is performed, because calculating the results of several statistical tests and then selecting the most (or least) significant result wrongly inflates (or deflates) the reported significance of any ROC difference that was found.

The other important distinction among statistical tests for differences between ROC estimates concerns the source(s) of statistical variation that the tests take into account. Recall that a "p value" is supposed to represent the probability of finding a difference larger than the difference seen in the experiment if, in fact, no difference exists in the population from which the experiment's data were sampled. This probability can be interpreted as a proportion of results obtained by repeating the experiment many times under identical conditions, so we must ask, under what conditions is the experiment (conceptually) to be repeated? For example, do we imagine repeating an ROC experiment with independently sampled radiologists and independently sampled cases, with independently sampled radiologists but the same cases, with the same radiologists but independently sampled cases, or with the same cases and the same radiologists? The results of an ROC experiment repeated many times under each of these different scenarios can be expected to vary by different amounts: to vary most when both radiologists and cases are resampled independently for each repetition, for example, and to vary least when both radiologists and cases are held constant across repetitions [25]. Hence, a statistical test for differences in ROC estimates must be selected to take into account the source(s) of variation that allow a practically meaningful conclusion to be drawn. For example, some statistical tests for ROC differences take only reader variation into account; therefore, any conclusion drawn from such a test can be applied to the population of radiologists from which the experiment's radiologists were sampled, but only to the particular cases that were used in the experiment at hand. Similarly, some statistical tests for ROC differences take only case-sample variation into account, so their conclusions can be applied to the population of cases from which an experiment's cases were sampled, but only to the particular radiologists who participated in that experiment. Clearly, then, both reader variation and case-sample variation should be taken into account when statistical significance is calculated for most ROC studies, because scientific progress depends on the replication of experimental results in independent laboratories, where different readers and different cases are likely to be used. This is the domain of so-called multireader, multicase (MRMC) ROC analysis, which has been refined greatly during the past 15 years and is now the methodology of choice for most ROC comparison studies [26].

As a final general note on statistical issues in ROC analysis, one should keep in mind that ROC comparison studies can fail to demonstrate a statistically significant difference for either of 2 fundamentally different reasons: because there was in fact little or no difference between imaging systems in the populations of readers and cases studied or because the number of readers or cases used in the experiment was too small to provide adequate statistical power. Therefore, a failure to demonstrate a statistically significant difference must not be considered equivalent to "proving" that no difference exists but instead must be accompanied by some measure of the range in which the real difference is likely to lie, given the experiment's data (eg, a confidence interval or retrospective power analysis). Confidence intervals are especially helpful in overcoming the difficulty of interpreting "p" values in situations in which no significant difference was found [27].

FREE SOFTWARE

Although the calculations associated with most meaningful ROC curve-fitting and significance-testing techniques are extremely complicated, free software for these purposes is available from several investigators:

- A number of programs for fitting conventional ROC curves as well as software for MRMC ROC analysis can be obtained from Kevin Berbaum, PhD, Department of Radiology, University of Iowa (kevin-berbaum@uiowa.edu).
- A version of the Iowa MRMC software that has been modified to test the significance of differences between AFROC curve estimates can be obtained from Dev Chakraborty, PhD, Department of Radiology, University of Pittsburgh, (http://www.devchakraborty.com/index_files/Page396.htm).
- Software for curve fitting and various statistical analyses of conventional ROC data can be downloaded from the University of Chicago's Web site (http://dpc10@imap.pitt.edu).
- Software for a wide variety of statistical analyses of conventional ROC data, including power predictions and an alternative approach to MRMC analysis, can be obtained from Nancy Obuchowski, PhD, Department of Quantitative Health Sciences, Cleveland Clinic Foundation (http://www.bio.ri.ccf.org/html/rocanalysis.html).

REFERENCES

1. Metz CE. Quality of the observed image. ICRU News 1989;2:20-3.
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.
3. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-98.
4. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720-33.
5. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109-21.
6. McLish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989; 9:190-5.
7. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996;201:745-50.
8. Halpern EJ, Alpert M, Krieger AM, Metz CE, Maidment AD. Comparisons of ROC curves on the basis of optimal operating points. Acad Radiol 1996;3:245-53.
9. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology 1975;116:533-8.
10. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys 1996;23:1709-25.
11. Egan JP, Greenberg GZ, Schulman AI. Operating characteristics, signal detection, and the method of free response. J Acoust Soc Am 1961;33:993-1007.
12. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free response approach to the measurement and characterization of radiographic observer performance. J Appl Photog Eng 1978;4:166-72.
13. Berbaum KS, Franken EA, Dorfman DD, et al. Satisfaction of search in diagnostic radiology. Invest Radiol 1990;25:133-40.
14. Chakraborty DP, Winter LHL. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology 1990;174:873-81.
15. Chakraborty DP. Statistical power in observer performance studies: a comparison of the ROC and free-response methods in tasks involving localization. Acad Radiol 2002;9:147-56.
16. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in N-class classification. IEEE Trans Med Imaging 2004;23:891-5.
17. Edwards DC, Metz CE. Review of several proposed three-class classification decision rules and their relation to the ideal observer decision rule. Proc SPIE Med Imag 2005;5749:128-37.
18. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. J Math Psych 1969;6:487-96.
19. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 1986;99:181-98.
20. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. Med Decis Making 1988;8:197-203.
21. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234-45.
22. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Dagga HA. Proper ROC analysis: the bigamma model. Acad Radiol 1997;4:138-49.
23. Pan X, Metz CE. The "proper" binormal model: parametric ROC curve estimation with degenerate data. Acad Radiol 1997;4:380-9.
24. Metz CE, Pan X. "Proper" binormal ROC curves: theory and maximum-likelihood estimation. J Math Psych 1997;43:1-33.
25. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad Radiol 1997;4:587-600.
26. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. Acad Radiol 2002;8:1264-77.
27. Metz CE. Quantification of failure to demonstrate statistical significance: the usefulness of confidence intervals. Invest Radiol 1993;28:59-63.