review

CLIN. CHEM. 39/4, 561–577 (1993)

Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine

Mark H. Zweig[1] and Gregory Campbell[2]

The clinical performance of a laboratory test can be described in terms of diagnostic accuracy, or the ability to correctly classify subjects into clinically relevant subgroups. Diagnostic accuracy refers to the *quality* of the information provided by the classification device and should be distinguished from the *usefulness*, or actual practical value, of the information. Receiver-operating characteristic (ROC) plots provide a pure index of accuracy by demonstrating the limits of a test's ability to discriminate between alternative states of health over the complete spectrum of operating conditions. Furthermore, ROC plots occupy a central or unifying position in the process of assessing and using diagnostic tools. Once the plot is generated, a user can readily go on to many other activities such as performing quantitative ROC analysis and comparisons of tests, using likelihood ratio to revise the probability of disease in individual subjects, selecting decision thresholds, using logistic-regression analysis, using discriminant-function analysis, or incorporating the tool into a clinical strategy by using decision analysis.

Indexing Terms: *receiver-operating characteristic curves · data analysis · diagnostic accuracy · likelihood ratio · diagnostic threshold · test efficiency · predictive value*

Reports evaluating some clinical aspect of laboratory test performance frequently appear in this and many other journals. However, the elements of performance that are addressed vary. What is clinical performance? Terms commonly used include sensitivity and specificity, efficiency, accuracy, utility, value, worth, effectiveness, usefulness, and efficacy. Often the word diagnostic precedes the term, i.e., diagnostic value or diagnostic efficiency. Other terms such as predictive value (positive and negative), likelihood ratio, odds ratio, and likelihood quotient have been used. The meaning of these terms is often vague and variable, particularly utility, worth, value, usefulness, and effectiveness, but even diagnostic accuracy seems to mean different things to different people.

As laboratorians, we are often interested in how well a test performs clinically, because we are considering replacing an existing test with a newer one, adding a new test to our laboratory's menu, eliminating tests where possible, or just because we want to know something about the value of what we are doing. In the first six issues of *Clinical Chemistry* during 1991, at least 18 studies addressed questions about clinical performance. Some of these studies assessed test performance merely by calculating the mean test results for the various sample groups they studied. Others calculated sensitivity, specificity, efficiency, and (or) predictive value. Five of the 18 studies included receiver- (or relative-) operating characteristic (ROC) plots to represent test performance.[3]

It is apparent that both the concepts and the measures of performance varied from study to study. The lack of a standardized approach to performance makes for a confusing situation in which the investigators and readers fail to communicate and understand clearly the information of interest to both groups. Without agreement about the concept of performance and how it should be measured and represented, the struggle to understand the meaning of the data gathered is likely to continue.

Here we offer a definition of accuracy to be used as a measure of performance, and review the principles and application of the ROC plot as an index of diagnostic accuracy (*1*). ROC plots are fundamental; their pivotal position provides a unifying concept in the process of test evaluation (Figure 1). Once the data are collected and the plots generated, numerous other assessments, comparisons, indices, and analyses can follow. Even clinical decision analysis (not shown), a complex but important tool for medical decision making, involves the use of data generated for ROC plotting. It is this central position of ROC plots that we will describe in this review (Figure 1).

ROC methodology is based on statistical decision theory and was developed in the context of electronic signal detection and problems with radar in the early 1950s (*2*). An ROC-type plot was used in the late 1950s to describe the ability of an automated Pap smear analyzer to discriminate between smears with and without malignant cells. Curves of true-negative vs false-negative results were used to select an operating point for the instrument that would provide an optimum trade-off between false-positive and false-negative results (*3*).

By the mid-1960s, ROC plots had been used in experimental psychology and psychophysics (*2*). Following work in psychophysics by Green and Swets (*4*), Leo Lusted, a radiologist, suggested using ROC analysis in medical decision making in 1967 and began using it in studies of medical imaging devices in 1969 (*5, 6*). Others wrote about it (*7*), and eventually ROC analysis made its way into other areas of medicine.

Laboratory tests are ordered to help answer questions about patient management. How much help an individual test result provides varies and, in any case, is a highly

[1] Clinical Pathology Department, Warren G. Magnuson Clinical Center, and [2] Biometry and Field Studies Branch, National Institutes of Neurological Diseases and Stroke, National Institutes of Health, Bethesda, MD 20892.

Received August 20, 1991; accepted November 2, 1992.

[3] Nonstandard abbreviations: ROC, receiver-operating characteristic; CK, creatine kinase; AMI, acute myocardial infarction; CAD, coronary artery disease; HDL, high-density lipoprotein; LR, likelihood ratio; PV(+), predictive value of a positive test result; and PV(−), predictive value of a negative test result.

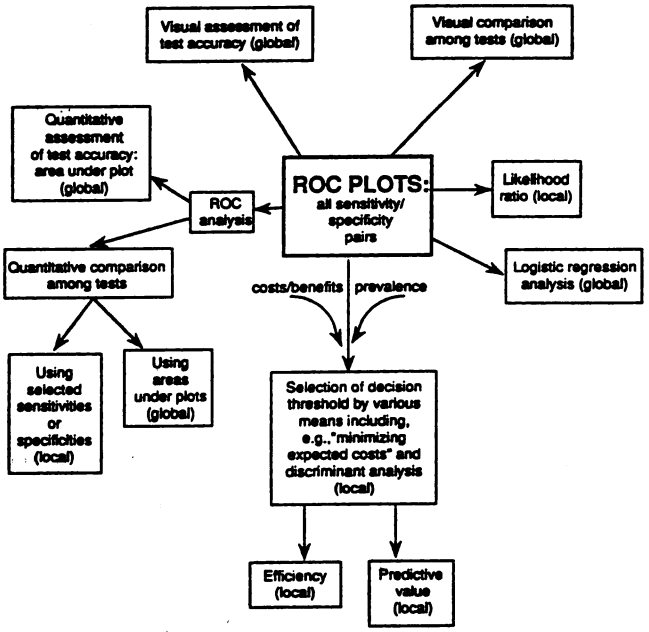

Fig. 1. Diagram showing central position of ROC plots in test performance evaluation

complicated issue. Management decisions and strategies are complex activities requiring the physician to consider probabilities of disease, quality of the data available, effectiveness of various treatment/management alternatives, probability of outcomes, value (and cost) of outcomes to the patient, etc. Many types of clinical data (including laboratory results) are usually integrated into a complex decision process. Most often, a single laboratory test result is not the sole basis for a diagnosis or a patient-management decision. Therefore, some have criticized the practice of evaluating the diagnostic performance of a test as if it were used alone. However, each clinical tool, whether it is a clinical chemistry test, an electroencephalogram, an electrocardiogram, a nuclide scan, a roentgenogram, a biopsy, a view through an orifice, a pulmonary-function test, or a sonogram, is meant to make some definable discrimination. It is important to know just how inherently accurate each diagnostic discriminator (test) is. We emphasize that assessing accuracy, without engaging in comprehensive clinical decision analysis, is a valid and useful activity for clinical laboratories. It is far more feasible and less laborious than decision analysis, the latter being important in devising strategies and policies for patient management but unnecessary for addressing a number of clinical laboratory issues.

Diagnostic Accuracy and Usefulness

Accuracy: Quality of the Information

Diagnostic accuracy is the most fundamental characteristic of the test itself as a classification device; it measures a test's ability to discriminate among alternative states of health. In its simplest form, it is the ability to distinguish between just two states of health or circumstances. It involves distinguishing between health and disease, benign and malignant disease, responders and nonresponders to therapy, and predicting who will and who will not get sick.

Indeed, the ability of the test to distinguish between the relevant alternative states or conditions of the subject (i.e., diagnostic accuracy) is the most basic property of the test as an aid in decision making. This property is the place to start when assessing what contribution a test can make to the patient-management process. If the test cannot provide the relevant distinction, it will not be valuable for patient care. On the other hand, once we establish that a test does discriminate well, we can explore its role in the process of patient management to determine the practical usefulness of the information in a management strategy. This exploration is clinical decision analysis, and measures of test accuracy are part of the important input used to carry out such analysis.

Usefulness: Practical Clinical Value of the Information

Usefulness refers to the practical value of the information in managing patients [Swets and Pickett term this efficacy (*8*)]. A test may have considerable ability to discriminate, yet be of little practical value for patient care. This could happen for several reasons: the cost or undesirability of false results may be so high that there is no decision threshold for the test for which the trade-off between sensitivity and specificity is acceptable; there may be a less invasive or less expensive means to obtain comparable information; the test may be so expensive or technically demanding that its availability is limited; or the test could be so uncomfortable or invasive that subjects will not submit to it.

We note that exploration of the usefulness of medical information, such as test data, involves many factors or considerations that are not properties of the test system or device, but rather properties of the circumstances of the clinical application. These include the prior probability of disease (prevalence), the possible outcomes and relative values of these outcomes, the costs to the patient (and others) of incorrect information (false-positive and false-negative classifications), and the costs and benefits of various treatment options. These factors may affect the usefulness of the test; therefore, it is helpful to separate, conceptually, the characteristic that is fundamental and inherent in tests themselves, discrimination ability, from the interaction that results when this discrimination ability is affected by external factors in the course of patient management.

Accuracy vs Usefulness

Therefore, we define diagnostic accuracy as the ability to discriminate between two subclasses of subjects, when there is some clinically relevant reason to do such. This concept of accuracy refers to the *quality* of the information (classification) provided by the test and should be distinguished from the practical *usefulness* of the information (*2*). Both are aspects of test performance. Second, we suggest that assessment of accuracy is the place to start in evaluating test performance. If a test cannot discriminate between clinically relevant subclasses of subjects, there is little reason to explore a possible clinical role. If, on the other hand, a test exhibits substantial ability to discriminate, then by examining the degree of accuracy of the test or by comparing its accuracy with that of other tests, we can decide whether to continue with a more complex assessment of its role in management of patient care (decision analysis).

Assessing Test Performance in the Clinical Laboratory

How does this concept of accuracy relate to laboratory medicine? Consider some questions about test performance raised in the 18 reports in *Clinical Chemistry* mentioned earlier:

1. How well do serum amylase or lipase results discrim-

inate between acute pancreatitis and other causes of abdominal pain?

2. How well can assay of human papillomavirus DNA in exfoliated cervical cells by nucleic acid hybridization detect papillomavirus infection?

3. Can urinary neopterin concentration discriminate between active and inactive systemic lupus erythematosus?

4. Can parathyrin-related protein concentration discriminate hypercalcemia of malignancy from other causes of hypercalcemia?

5. How do total human chorionic gonadotropin and its free β-subunit compare in screening maternal sera for Down syndrome?

6. How does serum troponin T compare with creatine kinase (CK), CK-MB, and myoglobin in the diagnosis of acute myocardial infarction (AMI)?

7. How does carbonic anhydrase III in serum compare with CK in the detection of muscular dystrophy?

In all of these examples, a question of discrimination is being posed; this is obvious in items 1, 3, and 4, but is also true in the others. In patients presenting to an emergency room with chest pain, we try to discriminate between AMI and other causes of chest pain. When screening pregnant women for Down syndrome, we are distinguishing between the presence and absence of the condition. Most clinical questions involve distinction or discrimination between two or more alternatives. Although very often there are more than two alternative states of health at issue, the clinical question can still be framed in terms of a dichotomy: the presence or absence of some state. For example, among elderly persons with anemia, we may want to discriminate between iron deficiency and all other causes. In assessing the performance of a test, the question is: Do the test result distributions from the two (or more) subgroups differ? If they do not differ, obviously the test results cannot discriminate between the two subgroups; accuracy is nil. If the test results do not overlap at all, then there is perfect discrimination, i.e., perfect accuracy for those subjects. Most often the distribution of results partially overlap.

As clinical chemists operating a clinical laboratory, we are required to make practical decisions: Which of the new automated CK-MB assays should we use in our laboratory? Should we replace our electrophoretic assay with a faster, easier immunometric assay? Should we offer plasma apolipoprotein determinations? Should we offer an immunoassay for prostatic acid phosphatase instead of a conventional enzymatic assay? Should we replace prostatic acid phosphatase with prostate-specific antigen?

Adding tests, replacing tests, or updating methodology may require some judgment about diagnostic performance. These are some of the reasons why such studies appear frequently in this journal. Assessing accuracy helps address these practical questions. If a study indicates that an automated CK-MB assay discriminates between patients with and without AMI as well as or better than an existing electrophoretic method, we would consider adopting the new method. If an enzymatic acid phosphatase assay is shown to be as accurate as an RIA of prostatic acid phosphatase in identifying prostatic cancer in elderly men with suspicious signs, we would retain the simpler, less expensive technology.

Laboratorians considering replacing one methodology with another will often compare test results from the two assays by using linear-regression analysis of the split sample data. Take CK-MB, for example. A laboratory considering using an automated enzyme immunometric assay instead of electrophoresis may run a comparison by assaying patients' specimens by both systems to determine the degree of agreement between the existing method and the putative replacement. Underlying this approach is the assumption that the existing method gives good results and is the standard against which the new candidate is assessed. However, suppose the candidate method is actually a more accurate discriminator. Where the newer test is more accurate, it will disagree with the older. Unless the cause for the disagreement is discovered, the new test may be judged less accurate and undesirable when, in fact, it might be more accurate. A more valid approach to determine which is superior is to assess the accuracy of both methods against the truth. (At times, expressing truth in terms of outcome rather than diagnosis may be more feasible or relevant. This is particularly appealing when a good gold standard for diagnosis is lacking, such as for AMI.) This assessment against the truth is more difficult to perform, but yields far more relevant and valid information. The value of the study justifies the greater time and effort required, and the study, if published and available to all, need not be performed in every laboratory every time this conversion is considered.

Diagnostic Sensitivity/Specificity

We have defined diagnostic accuracy and tried to indicate its usefulness to laboratorians, but how is it measured and expressed? It is measured as diagnostic sensitivity and specificity, concepts well known for years in laboratory medicine (*9*). (Sensitivity and specificity in this review always refer to the diagnostic, rather than analytical, type.) Currently, laboratorians regularly think about performance of tests in these terms; they are commonly calculated, reported in the scientific literature, and may appear in manufacturer's literature as well. Thus, the importance of the underlying concept of test accuracy is already recognized. However, reporting only one value for sensitivity and specificity provides a possibly misleading and even hazardous oversimplification of accuracy. Tests do not have only one sensitivity or specificity, but many. Therefore, calculating one or just a few sensitivity/specificity pairs provides only a brief glimpse at a test's performance, a glimpse that may be far from revealing a test's real diagnostic abilities.

Indeed, tests can and do exhibit the complete spectrum of sensitivities or of specificities; it is the pairs that are limited and describe the accuracy of a test in discriminating between states of health. One can always identify a decision threshold (decision level, decision criterion, "cutoff" value) that corresponds to a diagnostic sensitivity of 100%, or one that yields sensitivities of ~95% or 90%, etc. For one test, however, the specificity corresponding to 95% sensitivity might be 97%, whereas for another test applied to the same clinical question the corresponding specificity might be only 70%. For any test in which the distributions of results from the two categories of subjects overlap, there are inevitable trade-offs between sensitivity and specificity. As the decision threshold, used to classify the subjects as positive or negative based on test results, is varied over the spectrum of possible results, the sensitivity and specificity will move in opposite directions. As one increases, the other decreases. For each decision threshold, there is a combination of sensitivity and specificity. Which one(s) describe(s) the tests' accuracy? Only the entire spectrum of sensitivity/specificity pairs provides a complete picture of test accuracy.

Graphical Displays of Diagnostic Accuracy

ROC plots provide a view of this whole spectrum of sensitivities and specificities because all possible sensitiv-

ity/specificity pairs for a particular test are graphed (*2, 7, 10–15*). A decision threshold must be chosen for a test to be used in patient care, but there is no need to choose any particular decision threshold for assessing accuracy. In fact, it is undesirable to do so, because assessing performance at a single point may result in misleading impressions about test performance, or in erroneous comparisons between tests (*11, 13*). The ROC plot provides a comprehensive picture of the ability of a test to make the distinction being examined *over all decision thresholds.*

Two other commonly used ways to represent the clinical results of a test are the dot diagram and the frequency histogram. Like ROC, both report all of the data; we strongly recommend that one of these three approaches be used. Figure 2 shows a dot diagram of serum CK-BB concentrations in two groups of subjects (*16*). All subjects had presented to an emergency room with typical chest pain suggestive of AMI. The obvious overlap in the distribution of results from those who had an AMI and those who did not results in trade-offs between sensitivity and specificity. If a decision threshold of 6 μg/L is chosen, the test exhibits a sensitivity (correct identification of AMI) of 100%, but the specificity (identification of non-AMI) is only ~50%. Specificity can be raised by increasing the decision threshold, but at CK-BB concentrations >7 μg/L, the sensitivity decreases. For example, at 11 or 12 μg/L, specificity is 100%, but sensitivity falls to 48/50 or 96%. Although it is apparent from this overlap that the test is not perfectly accurate in discriminating between subjects with AMI and those without, it is difficult to characterize the degree of inaccuracy by using only this plot.

Figure 3 shows another commonly used representation, a frequency histogram instead of a dot diagram. This Figure is taken from a study of the ability of four assays to discriminate between subjects with and without acute pancreatitis (*17*). The two tests shown exhibit overlap in

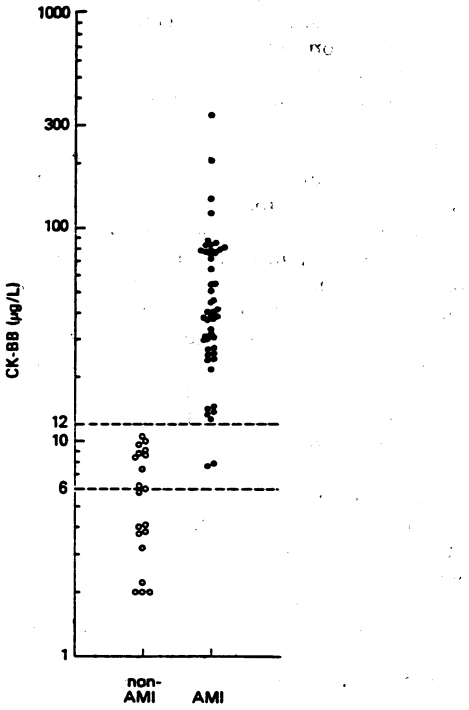

Fig. 2. Dot diagram of serum CK-BB concentrations 16 h after onset of symptoms in 70 subjects presenting to an emergency room with typical chest pain

Fifty were eventually considered to have had myocardial infarction; 20 were not

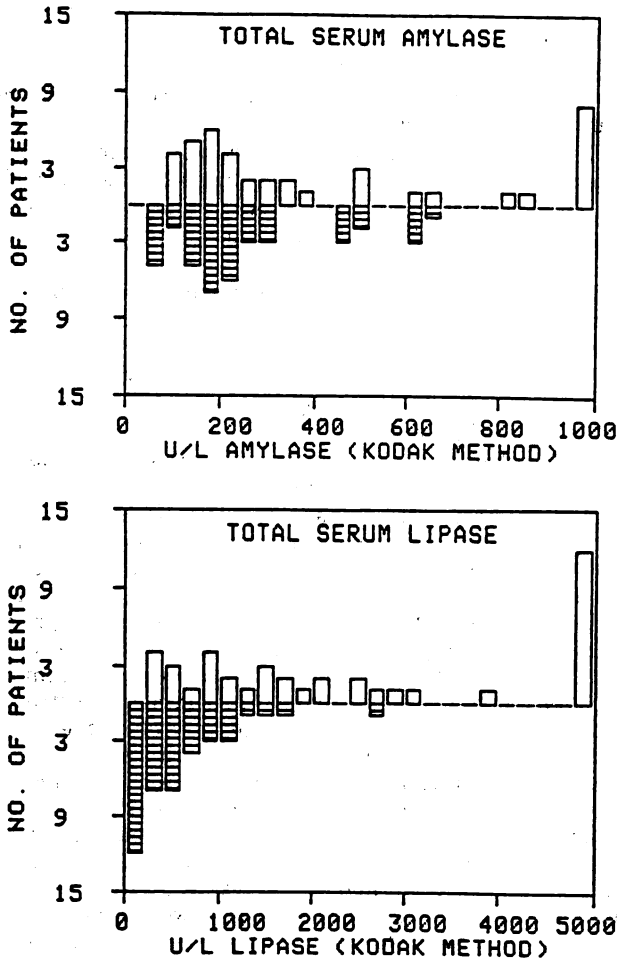

Fig. 3. Frequency histograms of serum enzyme concentrations in 41 patients with pancreatitis (*open bars*, above the line) and 40 without (19 gastrointestinal disease, 21 miscellaneous disorders; *striped bars*, below the line)

Reprinted with permission from Lott and Lu (*17*)

results, but it is difficult to describe or characterize the extent of overlap (accuracy) or to compare the accuracies of the two tests. Figure 4 shows two ROC plots corresponding to the two frequency-distribution histograms in Figure 3. The frequency histograms for each test are reduced to one ROC plot on a common scale.

The ROC Plot

The ROC graph is a plot of all of the sensitivity/specificity pairs resulting from continuously varying the decision threshold over the entire range of results observed. In each case, the ROC plot depicts the overlap between the two distributions by plotting the sensitivity vs 1 − specificity for the complete range of decision thresholds. On the *y*-axis is sensitivity, or the true-positive fraction [defined as (number of true-positive test results)/(number of true-positive + number of false-negative test results)]. This has also been referred to as positivity in the presence of a disease or condition. It is calculated solely from the affected subgroup. On the *x*-axis is the false-positive fraction, or 1 − specificity [defined as (number of false-positive results)/(number of true-negative + number of false-positive results)]. It is an index of specificity and is calculated entirely from the unaffected subgroup. (Note that some authors plot specificity, rather than 1 − specificity, on the *x*-axis.) Because the true- and false-positive fractions are calculated entirely

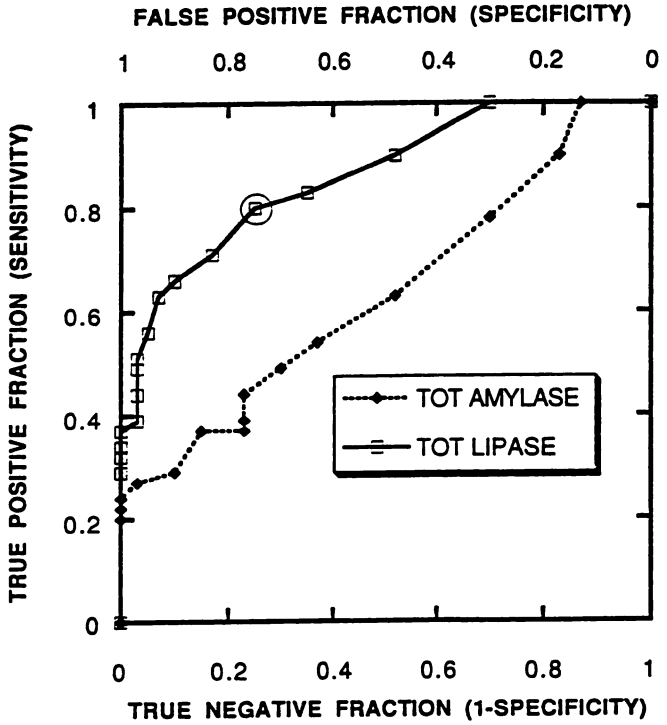

Fig. 4. Nonparametric ROC plots of two serum enzyme assays, based on the binned data as shown in Fig. 3

See text (*ROC analysis*) for discussion of circled point on ROC curve for lipase

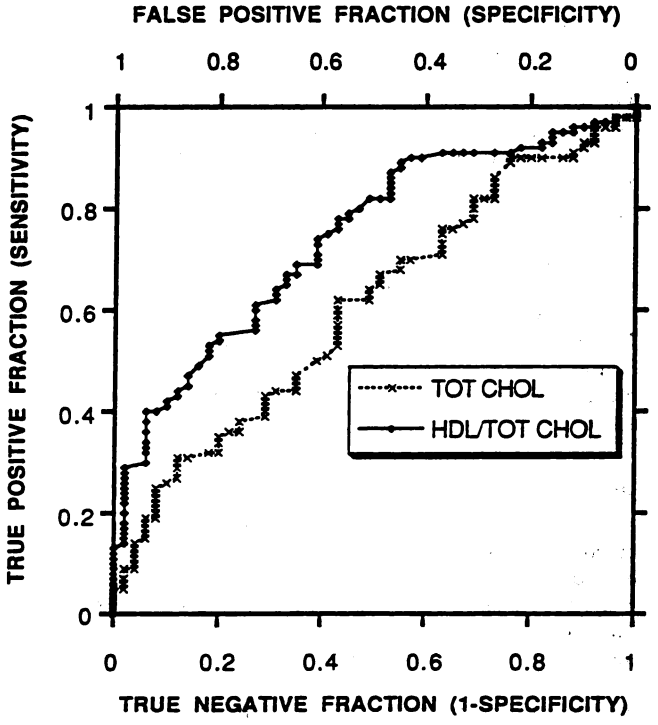

Fig. 5. Nonparametric ROC plots of the ratio of high-density lipoprotein (HDL) to total cholesterol and of total cholesterol concentration in 304 consecutive male patients who underwent coronary angiography for evaluation of suspected coronary artery (CAD) disease (255 had clinically significant CAD; 49 did not)

All were classified as having clinically significant CAD or not, based on angiographic findings. Data from Kottke et al. (*18*)

separately, by using the test results from two different subgroups, the ROC plot is independent of the prevalence of disease in the sample. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions of results) has an ROC plot that passes through the upper left corner, where the true-positive fraction is 1.0, or 100% (perfect sensitivity), and the false-positive fraction is 0 (perfect specificity). The theoretical plot for a test with *no* discrimination (identical distributions of results for the two groups) is a 45 ° diagonal line from the lower left corner to the upper right corner. Most plots fall in between these two extremes. (If the ROC plot falls completely below the 45 ° diagonal, this is easily remedied by reversing the criterion for "positivity" from "greater than" to "less than" or vice versa.) Qualitatively, the closer the plot is to the upper left corner, the higher the overall accuracy of the test.

Comparing Tests Visually with ROC Plots

When results from multiple tests have been obtained, the ROC plots can be graphed together, as in Figure 4. The relative positions of the plots indicate the relative accuracies of the tests. A plot lying above and to the left of another plot indicates greater observed accuracy. In Figure 4, the lipase assay exhibits greater observed accuracy than does the amylase assay. In Figure 5, the ratio of high-density lipoprotein (HDL) to total cholesterol is apparently more accurate than total cholesterol in identifying coronary artery disease (CAD) in a group of men. The ratio has a lower false-positive fraction at any given true-positive fraction; likewise, it has a higher true-positive fraction at any particular false-positive fraction. Figure 6 illustrates the ROC plots for two tests with greater accuracy than those in Figures 4 and 5. The plots pass closer to the upper left corner. The ability of dexamethasone suppression of urinary free cortisol excretion and of urinary 17-hydroxy-

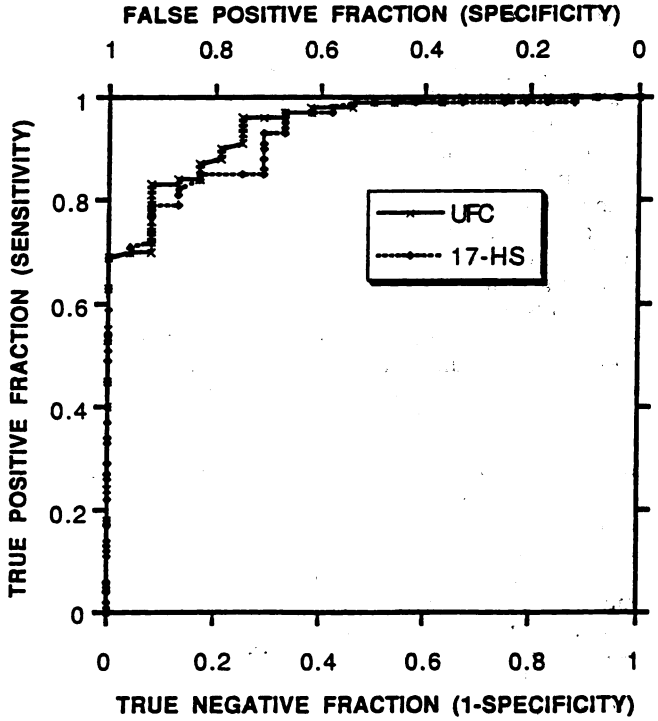

Fig. 6. Nonparametric ROC plots for dexamethasone suppression of the urinary excretion of free cortisol (UFC) and 17-hydroxysteroids (17HS) in distinguishing between pituitary (94) and nonpituitary (24) etiologies in 118 patients with Cushing syndrome

steroid excretion to discriminate those subjects with Cushing syndrome having a pituitary etiology from those hav-

ing a nonpituitary etiology is assessed here. (Diagnostic accuracy was evaluated by using surgical or histopathological diagnosis for the definitive classification of subjects.) The plots are virtually identical. It is easy to appreciate visually from Figure 6 that these two tests have essentially the same accuracy in making the intended discrimination. This assessment suggests that urinary free cortisol, a more convenient assay, could replace the older, more traditional 17-hydroxysteroid determination for addressing this clinical issue. Thus, the position of the plot (i.e., closeness to the upper left corner vs closeness to the 45 ° diagonal) provides qualitative information about the accuracy of a given test. The relative positions of two or more plots (e.g., Figures 4, 5, and 6) provide a qualitative comparison of accuracies of multiple tests. (The quantitative comparison with use of statistics will be discussed later.)

Generating the ROC Plot; Ties

Clinical data usually occur in one of two forms: discrete or continuous. Most clinical laboratory data are continuous, being generated from a measuring device with sufficient resolution to provide observations on a continuum. Measurements of electrolytes, therapeutic drugs, hormones, enzymes, and tumor-marker concentrations, etc., are essentially continuous. Urinalysis dipstick results are discrete data, as are rapid pregnancy-testing devices, which give positive/negative results. Scales in diagnostic imaging generally provide (discrete) ratings data with categories such as definitely abnormal, probably abnormal, equivocal, probably normal, and definitely normal.

A tie in laboratory data is of interest when a member of the diseased group has the same result as does a member of the nondiseased group. Such ties are more likely to occur when there are few data categories (i.e., different results), such as with coarse discrete data, rather than when the number of different results is large, as with continuous data. In radiology, where it may be convenient to categorize radiographs on a five-point rating scale, ties may be common; such discrete ordinal data are called ratings data. Both diseased and nondiseased individuals may have results in the "equivocal" category, for example. This tie comes from grouping or "binning" the data into ordered categories. In clinical laboratories, when observations are made on a continuous scale, ties are much less likely unless grouping into "bins" has occurred. Theoretically, if measurements are exact enough, no two individuals would have the same result on a continuous scale. However, the resolution of results in the clinical laboratory is often not so fine as to prevent this, and thus some ties will occur even with continuous data. Binning continuous data increases the chance for ties. Ties can be caused, then, either by the intentional binning of data or by the degree of analytical resolution of the method of observation. Figure 2 is an example of continuous data displayed in a dot diagram with no ties between AMI and non-AMI subjects. The interval or category size was small because concentrations were estimated to 0.1 μg/L and all results were considered individually. In Figure 3, however, where continuous data have been binned into a few intervals, the frequency histogram has introduced many ties. If individual results are used, fewer ties are likely.

For both tied and untied data, one merely plots the calculated (1 − specificity, sensitivity) points at all the possible decision thresholds (observed values) of the test. It is the graph of these points that is the ROC plot. For data with no ties, adjacent points can be connected with horizontal and vertical lines in a unique manner to give a staircase figure (Figure 7). As the threshold changes, inclusion of a true-positive result in the decision rule produces a vertical line; inclusion of a false-positive result produces a horizontal line. As the numbers of individuals in the two groups increase, the steps in the staircase become smaller and the plot usually appears less jagged. Because this ROC plot uses all the information in the data directly through the ranks of the test results in the combined sample, it can also be called the nonparametric ROC plot. The term nonparametric here refers to the lack of parameters needed to model the behavior of the plot, in contrast to parametric approaches, discussed later, which rely on models with parameters to be estimated.

When there are ties in continuous data, both the true-positive and false-positive fractions change simultaneously, resulting in a point displaced both horizontally and vertically from the last point. Connecting such adjacent points produces diagonal (nonhorizontal and nonvertical) lines. For tied data, the correct path (if it exists) between the two adjacent points is unknown. It could be the minimal path (horizontal first, then vertical) or the maximal path (vice versa). The straight diagonal line segment is the average of the two most extreme paths and tends to underestimate the plot for a diagnostically accurate test.

Often, ties are intentionally introduced in the display of the test results (Figure 3). A common approach often adopted in the clinical literature is to plot the ROC at only a few points, by using only a few decision thresholds, connecting adjacent points with straight line segments. This may be convenient, but when the data are collapsed into discrete categories with the number observed in each one being listed, the original measurement scale and the values of the individual test results are no longer used except through these counts. The data are reduced to a 2 × k table of counts, where k is the number of intervals. The originally continuous amylase data in Figure 3 has been collapsed into 25 intervals or categories of ~40 U/L each, not all of which contain data. The figure is essentially a 2 ×

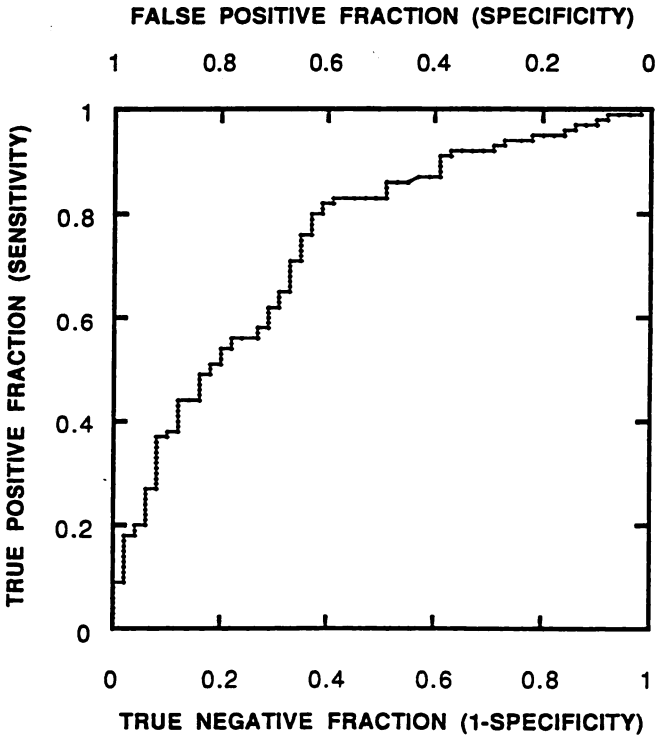

Fig. 7. Nonparametric ROC plot of serum apolipoprotein A-I to B ratios from same subjects as in Fig. 5

All results considered individually, with no binning. Data from Kottke et al. (*18*)

16 table of counts or observations, because 9 of the intervals contain no data. Figure 8 shows ROC plots for the ratio apolipoprotein A-I/B. If, for ease of plotting, the continuous data are grouped into intervals of 0–0.50, 0.51–1.00 g/L, etc., the data are reduced to a table of counts of observations in these bins, and the ROC plot reflects only these counts. Although this bin approach has the advantage of plotting ease, it discards much of the data and introduces many ties in the data (all data in the same bin are treated as tied). If the points are few and far between, this approximation can be poor and misrepresent the actual plot. Note in Figure 8 that the eight-bin ROC plot connected with diagonal lines is generally dominated from above by the more exact and accurate unbinned staircase ROC plot. Using all the data is more cumbersome; perhaps it is the lack of readily available software for using the original continuous data that encourages the display of such simplified ROC plots.

A more analytical method, usable only with tied data from the $2 \times k$ table of counts, employs some parametric model for fitting a curve. There are several good sources of information for parametric ROC curve construction and analyses based on such discrete data, including books by Green and Swets (*4*) and Swets and Pickett (*8*). One assumes that these counts are modeled by some parametric family of distributions and then estimates the parameters of the two distributions as well as the cutoffs that define the intervals. A popular model is the so-called binormal model, which assumes that the distributions are the normal (gaussian) parametric family, with usually different means and possibly different variances. (Note that this does not imply that the distributions of the original test results are gaussian.) Such assumptions cannot be completely verified because the imagined distributions are usually not observed directly but are reflected only through the counts. Authors have disagreed concerning the adequacy of these assumptions (*19–24*). Other theories, involving different distributions such as the logistic or negative exponential (*25*) to model the counts, suffer similar drawbacks. Many software packages facilitate the parametric approaches (see Table 2, later). Although these parametric approaches may be appropriate for radiologists dealing with ratings (discrete) data, they are much less so for clinical laboratorians, who usually already have a continuous scale for the data. In contrast to ratings data, it makes much less sense to collapse continuous data and introduce (or increase) ties simply to use the parametric modeling theory of signal detection, which was evolved to produce smooth curves for ratings data. A different approach is to fit the plotted ROC points directly to some mathematical function on the scale of the ROC plot. Such an approach implicitly presumes a parametric model. If this fit of the function to the ROC plotted points is least squares, an additional flaw is that the fit is based only on vertical errors, but here the errors are in both the horizontal and vertical directions of the ROC plot.

Table 1 lists the advantages and disadvantages of the nonparametric as well as the parametric ROC plots. Approaches that introduce ties in essentially continuous data from the laboratory have serious disadvantages. For continuous data, the nonparametric ROC is preferred. It passes through all the observed points. This is attractive because, for each observed threshold, the best (unbiased) guesses of sensitivity, specificity, and area are the nonparametric ones. (This issue of unbiasedness will be discussed later.) No data are discarded, unlike with the bin approach. There is no need to impose a model, either directly by picking a parametric family or indirectly by fitting a function to the points on the ROC plot. Although one would

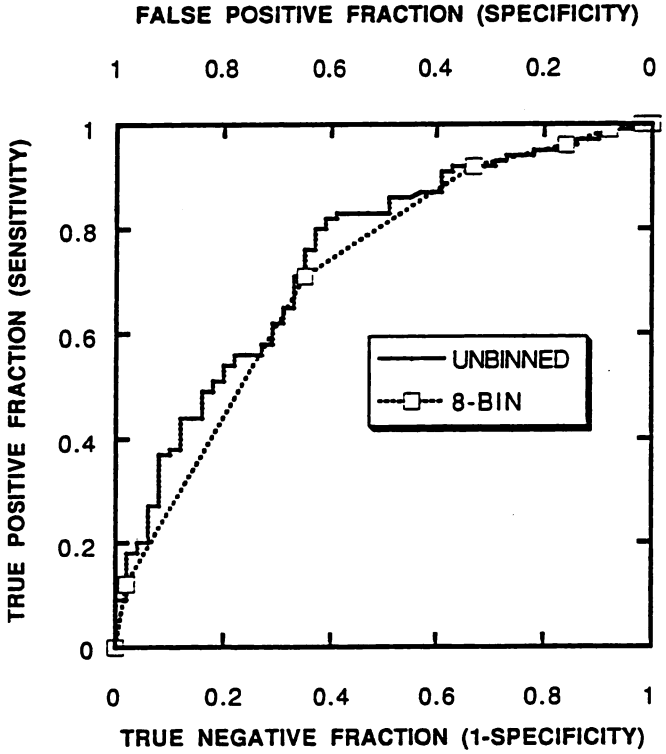

Fig. 8. Nonparametric ROC plots of ratios of serum apolipoprotein A-I to B, unbinned ("staircase") and grouped into eight bins (diagonals)

Same subjects as in Fig. 5. Data from Kottke et al. (*18*)

Table 1. Advantages and Disadvantages of Two ROC Approaches for Laboratory Data

	Nonparametric	Parametric
Advantages	Uses (ranks of) all data	Has smooth curve
	Makes no parametric assumptions	Compares curve at any sensitivity or specificity
	Plot goes through all the points	
	Provides unbiased estimates of sensitivity, specificity, ROC area	
	Computation is simple	
Disadvantages	Has staircase appearance	Discards data by binning
	Large data sets are plot-intensive	Assumes models
	Ties may be a problem	Curve does not necessarily go through actual points
	Compares plots only at observed sensitivity or specificity	Computation is complex
		ROC points and area are possibly biased
		Convergence problems for degenerate ROCs
		ROC slopes can be nonmonotonic

expect fewer ties with the nonparametric than with the discretized parametric approach, a substantial number of ties of test values between the diseased and the nondiseased groups may nonetheless occur. (For a few ties, the nonparametric plot with diagonal line segments is still very informative, especially if the diagonal line segments are short, e.g., as in Figures 5–7.)

Advantages of ROC Plots

The ROC plot has the following advantages: It is simple, graphical, and easily appreciated visually. It is a comprehensive representation of pure accuracy, i.e., discriminating ability, over the entire range of the test. It does not require selection of a particular decision threshold because the whole spectrum of possible decision thresholds is included. It is independent of prevalence: No care need be taken to obtain samples with representative prevalence; in fact, it is usually preferable to have equal numbers of subjects with both conditions. It provides a direct visual comparison between tests on a common scale, whereas both dot diagrams and frequency histograms require different plots if the scales differ. It requires no grouping or binning of data, as do frequency histograms. Its specificity and sensitivity are readily accessible, in contrast to dot diagrams and frequency histograms.

Disadvantages of ROC Plots

Of the 18 papers mentioned earlier, only 5 included ROC plots. Others had some data on sensitivity, specificity, efficiency, and/or predictive value, but without ROC plotting. Why such an elegant but simple tool has been underutilized by laboratorians is a puzzle. It is widely recognized in medicine as a powerful way to represent the accuracy of a signal detection system. Clinical laboratorians have written about it and utilized it for years. Nevertheless, although they embrace the concept, the laboratory community has been slow to use this tool. Even when authors include ROC plots in their publications, they frequently underutilize them or even present them without further comment, basing conclusions on other data.

There are apparent disadvantages of the ROC plot. Unlike dot diagrams and frequency histograms, actual decision thresholds are usually not displayed in the plot, though they are known and used to generate the graph. They are hidden from easy view. The number of subjects is also not shown on the display (although it is in the dot diagram), and as the sample sizes decrease, the ROC plots tend to become increasingly jagged and bumpy. However, even with large numbers of subjects, the plots may be bumpy. The generation of plots and calculation of parameters is cumbersome without computer software. With appropriate software, ROC plotting is quite readily done, but friendly, flexible software is not widely available.

ROC Analysis

Confidence Intervals for Sensitivity and Specificity

Because different groups of patients selected at random from a population can yield different ROC plots, such sampling variability for a single ROC plot is often indicated by reporting the variance or constructing a confidence interval about a point or points on the ROC plot. Moreover, because ROC plots can be treated either nonparametrically or parametrically, the statistical estimation (including confidence intervals) for points on the ROC plot is treated likewise.

For a chosen threshold, a point on the nonparametric ROC has the advantage of being an unbiased estimate of sensitivity and of 1 − specificity for that decision threshold. This means that, on average, the point neither over- nor underestimates the true (but unknown) values of sensitivity or specificity for that threshold. (This may not be true in the parametric approaches.) One could report a confidence interval about the sensitivity or a confidence interval about the specificity. Calculations of these nonparametric confidence intervals for the sensitivity and specificity have been described elsewhere (*14, 26*). For example, in the total lipase plot in Figure 4, at the circled ROC point with observed 1 − specificity = 0.250 and sensitivity = 0.805 (which corresponds to the threshold >800 U/L), a ~95% confidence interval for sensitivity is (0.684, 0.926). Of course, the variances and confidence intervals for the sensitivity and specificity shrink as the group sizes increase. A different but also correct approach fixes not the decision threshold but the true (theoretical) specificity at, say, 80% and then constructs a 90% confidence interval for the sensitivity (or vice versa) (*27*).

The parametric approach to confidence interval estimation relies on the initial estimation of the parameters. Under the assumption that the underlying distributions that give rise to the counts are normal (gaussian), a computer-intensive approach based on maximum likelihood estimation yields estimates of the parameters as well as their variances (*28, 29*). Computer programs are available to perform this complicated estimation. The output can be applied to a theory for doing inference (confidence interval and hypothesis testing) for specificity and for sensitivity, even for unobserved values (*30*).

Area under a Single ROC Plot

One convenient global way to quantify the diagnostic accuracy of a laboratory test is to express its performance by a single number. The most common global measure is the area under the ROC plot. By convention, this area is always ≥0.5 (if it is not, one can reverse the decision rule to make it so). Values range between 1.0 (perfect separation of the test values of the two groups) and 0.5 (no apparent distributional difference between the two groups of test values). The area does not depend only on a particular portion of the plot such as the point closest to the diagonal or the sensitivity at 90% specificity, but on the entire plot. This is a quantitative, descriptive expression of how close the ROC plot is to the perfect one (area = 1.0). The statistician readily recognizes the ROC area as the Mann–Whitney version of the nonparametric two-sample statistic (*31, 32*), introduced by the chemist Frank Wilcoxon. An area of 0.8, for example, means that a randomly selected individual from the diseased group has a laboratory test value larger than that for a randomly chosen individual from the nondiseased group 80% of the time. It does *not* mean that a positive result occurs with probability 0.80 nor that a positive result is associated with disease 80% of the time.

When there are no ties between the diseased and nondiseased groups, this area is easily computed from the plot as the sum of the rectangles under this graph. Analytical formulas to calculate the area are in reports by Bamber (*31*) and Hanley and McNeil (*32*). Alternatively, the area can be obtained indirectly from the Wilcoxon rank-sum statistic (*33*). Also, the area is related to the overlap index of Hartz (*34*) by the formula area = 1 − (overlap index/2). The area is an unbiased estimate of the true area under the theoretical ROC curve.

It is possible to test whether the diagnostic test is at all effective in distinguishing the two populations as well as to estimate the ROC area by a confidence interval. In partic-

ular, the rejection of the hypothesis that the theoretical area is 0.5 provides evidence that the laboratory test does have the ability to distinguish between the two groups. For example, for the data in Figure 7, the area is 0.747, with a standard deviation of 0.038. Here, the statistical test confirms that this area is significantly different from 0.5 ($P < 0.001$). It is also possible to form nonparametric confidence intervals about the area (*14*); the 95% confidence interval for the area in Figure 7 is (0.673, 0.821).

With clinical data, one must often contend with the possibility of ties, even though if the data were truly continuous, ties would not occur. If there are only a few tied values, it is reasonable to connect the adjacent points on the ROC plot (producing some diagonals) and to calculate the area in one of several ways: merely add up the areas under the trapezoids that comprise the entire area or use the Mann–Whitney version of the Wilcoxon statistic with average ranks. However, if the number of ties is considerable, this trapezoidal area tends to be a biased underestimate of the true area. This can be illustrated for continuous data that have been discretized. For example, Figure 8 presents the ROC plot based on the discretization of the data of Figure 7 into eight bins. Note that the eight-bin ROC plot generally lies below the unbinned ROC plot. Whereas the unbiased area estimate is 0.747 for the unbinned plot, the area under the eight-bin discretization is smaller (0.718). This is because the actual ROC plot tends to be convex and therefore to lie above, rather than below, the diagonals (*30, 32, 35–37*). If the number of ties is large, it is advisable to report not only the trapezoidal area but the maximal area where all diagonals are replaced by maximal vertical, then horizontal possible paths. For example, in Figure 8, whereas the unbinned maximal area = 0.747 (only one small diagonal), the maximal area for the eight-bin plot is much larger (0.854). Introduction of ties by binning the data biases the trapezoidal area estimate and increases the difference between that estimate and the maximal-area estimate. The standard deviation of the trapezoidal area estimate is always increased with discretization; in this case, it is 0.036 for the unbinned area and 0.041 for the eight-bin area.

For the parametric approach, there is a graphical method that estimates the parameters of the binormal model to obtain an area estimate (*2, 4*). A more exact approach involving the above maximum likelihood estimation provides not only an area estimate but also its standard error (*8, 28, 29*). The latter permits hypothesis testing and confidence interval estimation of the area. However, if a parametric model such as the binormal has been applied, the area obtained by estimation of the parameters can also be biased, unless the parametric assumptions on the counts are well satisfied. A comparison of nonparametric and binormal parametric areas has been done by Centor and Schwartz (*38*).

Area is in some sense an imperfect measure of the performance of the diagnostic test. For one thing, it is a single global measure—there is necessarily a loss of information in reducing the overall performance of the diagnostic test to a single number. A less global alternative is to restrict the area to a relevant portion, e.g., the area under the curve for observed specificity >0.6 or for sensitivity ≥0.7. This restriction analysis has been accomplished nonparametrically (*39*) as well as under the binormal model (*40*). Because the area under the ROC plot condenses the information of the graph to only a single number, it is usually undesirable to consider area without examining the plot itself. As Figure 9 illustrates, two ROC plots can be quite different in shape and yet have similar areas. At

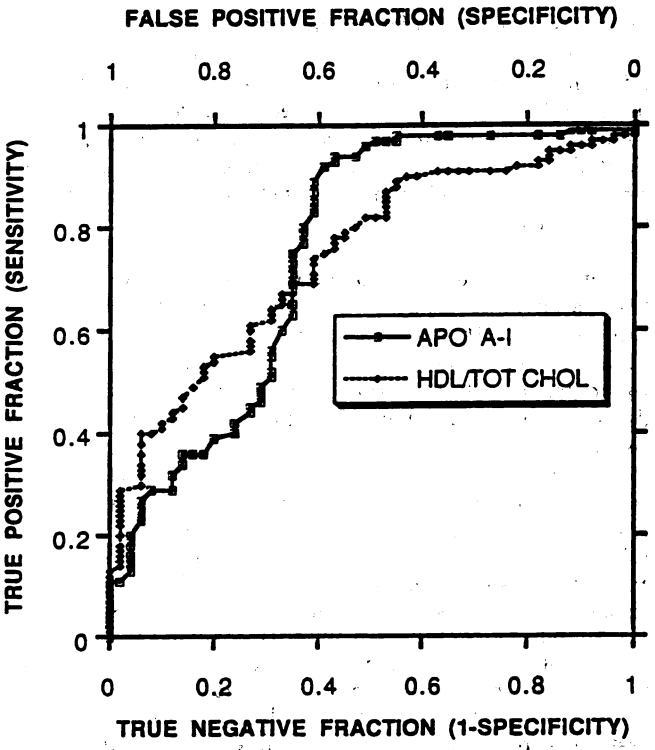

Fig. 9. Nonparametric ROC plots of serum apolipoprotein A-I concentrations and the ratio of HDL to total cholesterol

Same subjects as in Fig. 5 (data from Kottke et al., *18*). Areas under the ROC plots are 0.753 and 0.743, respectively

sensitivities below ~0.65–0.70, the HDL/total cholesterol ratio has better specificity (lower false-positive fraction), whereas at higher sensitivities apolipoprotein A-I has better specificity.

Statistical Comparison of Multiple Tests by Use of ROC Plots

Direct statistical comparison of multiple diagnostic tests is frequent in clinical laboratories. Two (or more) tests are usually performed on the same subjects, as in a split-sample comparison. In such cases, the results from two tests are usually correlated or associated. It is also possible, but less common, to have different individuals for the two tests, in which case the test results are independent (and hence uncorrelated). Of these two designs, the paired (split) design using the same individuals for the two tests is more efficient and also controls the patient-to-patient variation. For example, it may be possible to detect a real difference by examining 50 individuals with test A and another 50 with test B, whereas only 40 patients given *both* tests might have been sufficient. When comparison of test performance is accomplished statistically by using ROC plots or curves, it is referred to as ROC analysis. (See below for other ways to compare diagnostic tests.) ROC graphs for the two diagnostic tests, either nonparametric or parametric, can differ in shape but still may agree at a single point or have the same areas. Therefore, it is always advisable to visualize the entire performances through the ROC graphs.

Note that with only a single (specificity, sensitivity) pair for each test (i.e., only one point on the ROC plot for each test), a comparison of the performance of the two tests is usually impossible. Only if the two points (one for each test) are on the same vertical or same horizontal line on the nonparametric ROC plot can they be compared (i.e., at a common sensitivity or specificity). In such cases, McNe-

mar's test can be used for the paired data and a chi-squared test or Fisher's exact test can be used on the 2 × 2 tables if different subjects are used for the two tests (*14*). However, because this applies to only one decision level, this comparison is not global and the conclusions are consequently limited. Unlike the nonparametric approach, the parametric approach is able to compare two ROC curve sensitivities at a common (usually unobserved) specificity (or vice versa) (*30*).

A global approach is to compare entire ROC plots by using an overall measure such as area under the plot; this can be done either nonparametrically or parametrically. This can be especially attractive to laboratorians because the comparison does not rely on the selection of a particular decision threshold (which should consider prevalence and cost trade-off information).

For comparison of ROC plots by nonparametric area, the independent case (of different individuals for the two tests) is straightforward (*41*). The more difficult case occurs when all test results come from the same patients (paired data). Here the visual impression of the ROC plots may be misleading, because the determination depends on how correlated the two tests are; unfortunately, this correlation is invisible on the graph of the two ROC plots. Correlation is important because the more associated the two laboratory tests are, the sharper the ability of the hypothesis testing to pick out small area differences as statistically significant. (There may be statistically significant differences that are not clinically significant.) The computations are much more involved because one must estimate the correlation or covariance between either the two tests or the two areas (*39, 42, 43*). A computer program is essential for this sort of analysis. There is also an approximate procedure from Hanley and McNeil (*41*) for correlated test data in which Pearson (not Kendall) correlations of the test results are used to estimate the correlation of the two areas. For example, in Figure 5, the correlated analysis of the difference in the areas (0.743 − 0.606, two-sided *P*-value <0.0015) provides evidence that total cholesterol and the ratio of HDL to total cholesterol differ in the ability to detect CAD, based on the approximate procedure. Here the correlations of total cholesterol with the ratio of HDL to total cholesterol are −0.532 for the CAD patients and −0.367 for the non-CAD patients. The ROC areas are positively correlated because large total cholesterol values and small ratios both lead to the diagnosis of CAD. If the correlation of the two variables were zero, as would happen if different subjects were used for the two ROC plots, the *P*-value would be larger (0.014). For the same area difference, more subjects would have been required for the uncorrelated case to achieve the statistical significance that was obtained in the correlated case. In Figure 6, there is no indication, based on the area difference 0.9439 − 0.9302, that measurement of urinary free cortisol and of 17-hydroxysteroids differ, even though the correlations of these two measurements are large (0.777 and 0.720), because the areas are so close.

For using the parametric ROC approach to discrete data, whether the data have been intentionally discretized or are inherently ratings (discrete) data, areas can be compared in the independent case and for paired data. The comparison of multiple tests is important for discrete data, especially if one wishes to compare two diagnostic tests that are each on a different (ratings) scale or to compare two raters that are apparently using the same discrete scale but may have totally different definitions of what the categories represent. For independent tests, this comparison based on areas is straightforward (*30*). For paired tests, an extension of the maximum likelihood estimation for the parameters of two bivariate normal distributions is computer intensive and relies heavily on computer software (*44*). A simpler approach uses an approximate parametric procedure based on correlations (*41*).

Other Analyses: Relation to ROC Plots

Likelihood Ratios and the ROC Plot

Use of the likelihood ratio has been discussed in several publications (*5, 12, 45–48*). Of interest here is the relationship between likelihood ratios and the ROC plot. Like ROC plots, likelihood ratios do not depend on prevalence or on the ratio of the costs of false-positive and false-negative results.

Likelihood ratio can be defined as the ratio between the probability of a defined test result given the *presence* of a disease and the probability of a defined test result given the *absence* of a disease. Here defined result can mean a single result or a group of results. This is because the likelihood ratio can be calculated for a particular single test value, for results in a defined interval, or for results on one side of a particular threshold. Each of these represents results in some interval. If the interval is very narrow, e.g., >99 to <101 units, then the likelihood ratio is for the single result, 100, included in that interval. At the other extreme, the interval can be very large, e.g., >100, which includes all results >100 units. The likelihood ratios correspond to slopes on the ROC plot. For results on the diseased side of a particular threshold, this slope is simply the ratio of the true-positive fraction to the false-positive fraction, i.e., sensitivity/1 − specificity. More generally, the slope reflects the change in sensitivity divided by the change in specificity over the defined interval of test results.

In the discussion of CK by Radack et al. (*47*), likelihood ratios were calculated for the specific intervals or "slices" 1–120, 121–240, 241–360, 361–480, and >480 U/L. They also calculated the likelihood ratio for >120 U/L, because 120 was considered the upper limit of normal for their hospital. (Likewise, with sufficient data, a likelihood ratio could be calculated for a very small interval such as 239–241.) The likelihood ratio for the interval 241–360 U/L was 4.13 and represented the slope of the ROC plot between the two points on the plot corresponding to the decision thresholds 241 and 360 (Figure 10, line segment "a"). The likelihood ratio for the interval >120 was 1.57 and represented the slope of ROC plot between the origin and the point corresponding to the threshold of 121; in this case, it is the ratio of the true-positive fraction to the false-positive fraction (segment "b," Figure 10). All likelihood ratios are slopes that can be calculated from the ROC plot. In fact, the nonparametric ROC curve is a concise graphical way of presenting information that is ordinarily presented in the tables of likelihood ratios.

We have defined likelihood ratio, but what does it mean? The conceptual meaning of likelihood ratio is tricky and can be confusing. Like sensitivity and specificity (and also, then, the ROC plot), it is an expression of probability of test results, given the presence (and absence) of disease. In the case above, a CK concentration of 241–360 U/L (LR = 4.13) was about four times as likely to occur in a patient with AMI as in a patient without AMI. This *does not* necessarily mean that, given a result in that interval, the result is four times as likely to be from a patient with AMI as it is to be from a patient without an AMI. If the likelihood ratio is 4, then the fraction of diseased subjects having a test result of 241–360 U/L is four times as high as the fraction of nondiseased subjects having such a result. For example, of 100 subjects with acute appendicitis, 80 (0.8) might have a

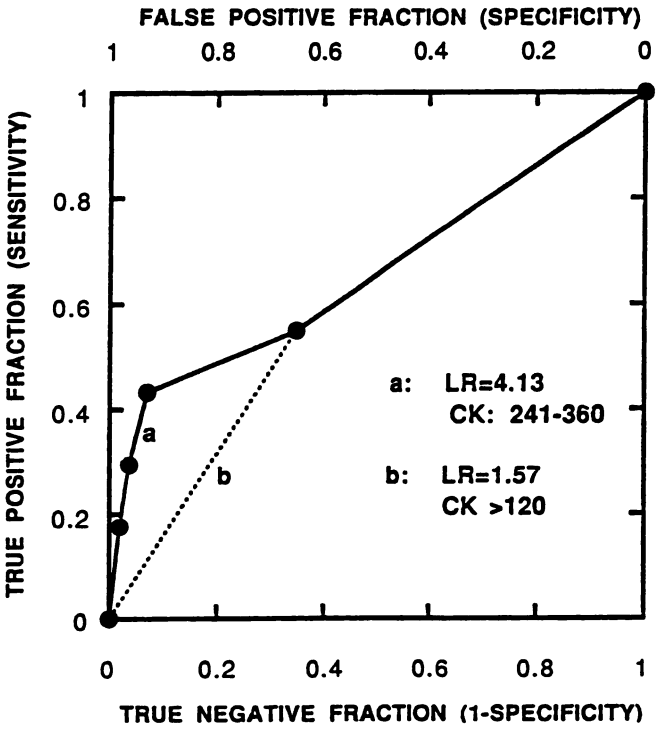

Fig. 10. Nonparametric ROC plot of serum creatine kinase activity in identifying acute myocardial infarction in patients presenting with chest pain

Fifty-one patients had infarct; 722 did not. *a* and *b* show differences in likelihood ratios for two different intervals for CK activity

positive result (falling in some defined interval) for test X. Of 100 subjects with abdominal pain not due to acute appendicitis, 20 (0.2) might have a positive test X. The likelihood ratio, true-positive fraction/false-positive fraction, is 0.8/0.2, or 4. However, this *does not* necessarily mean that a patient with abdominal pain and having a positive test result is four times as likely to have appendicitis as not. In other words, a likelihood ratio of 4 *does not mean* that the probability of appendicitis, given a positive result, is 0.8. To determine this probability (i.e., posttest probability, the probability of the presence of disease given a specific test result), the prior probability of disease (prevalence) must be included in the calculation. Such a calculation would be an expression of probability of disease given a positive test result, rather than the probability of a positive test result given the presence (or absence) of disease. *The likelihood ratio is the latter, not the former.* If at a particular hospital the prevalence of acute appendicitis among patients presenting with acute abdominal pain is 20%, then by Bayes' theorem, given a patient with a positive test result, the posttest probability of appendicitis is 50%. On the other hand, if the prevalence of appendicitis were actually 50% as in the example above with 200 subjects, then, given a positive result, the posttest probability of appendicitis is indeed 80%. The likelihood ratio itself is independent of prevalence. It is only when the likelihood ratio is used to calculate posttest probability from pretest probability (prevalence) that prevalence plays a role.

What is the usefulness, then, of the likelihood ratio? The likelihood ratio for a particular result or for an interval in which the result falls enables the revision of the pretest probability of disease. It is primarily a *tool* for calculating posttest probability of disease from pretest probability of disease, by using Bayes' theorem. In fact, the likelihood ratio is the minimal amount of information that is needed to revise prior disease probability by using Bayes' theorem. This is similar to but provides less information (true-positive, false-positive pairs) than the ROC plot, because the actual sensitivities and specificities are not evident in the likelihood ratio. The lower-most segment (CK >480 U/L) in the left-hand corner has a likelihood ratio of 9.26. Although this high ratio of the true-positive fraction (sensitivity) to the false-positive fraction (1 − specificity) may seem to indicate that the test is "effective" in this region, the ROC plot shows that the sensitivity reaches only ~0.18. Furthermore, of the 773 patients studied, only 23 (3%) had CK concentrations >480 U/L. Thus, although the probability of a person having an AMI will be increased considerably by a result in this interval (>480 U/L), only a small fraction of affected subjects will have such results (and some unaffected ones will, too). Because it does not locate the operating point on the curve, a likelihood ratio without an ROC plot may be misleading, no matter how high the ratio is. For example, Figure 11 contains the ROC plots of two hypothetical tests that have identical interval likelihood ratios (slopes) for the four segments shown. However, these two tests have very different diagnostic accuracy, as is evident from the ROC plot. The likelihood ratio is not a particularly good tool for assessing test performance or for comparing test performance.

It is important to note that the likelihood ratios referred to above are *estimated* nonparametric likelihood ratios because the true (theoretical) ratios are not known. (One can calculate parametric likelihood ratios by using an assumed binormal model.) Variability is associated with these estimated likelihood ratios; for a discussion of confidence intervals, see Centor (*49*).

Choosing Decision Thresholds: Trading off Sensitivity, Specificity, Prevalence, and Costs

The ROC plot may be used to observe the effect of different prevalences and different costs, and eventually to

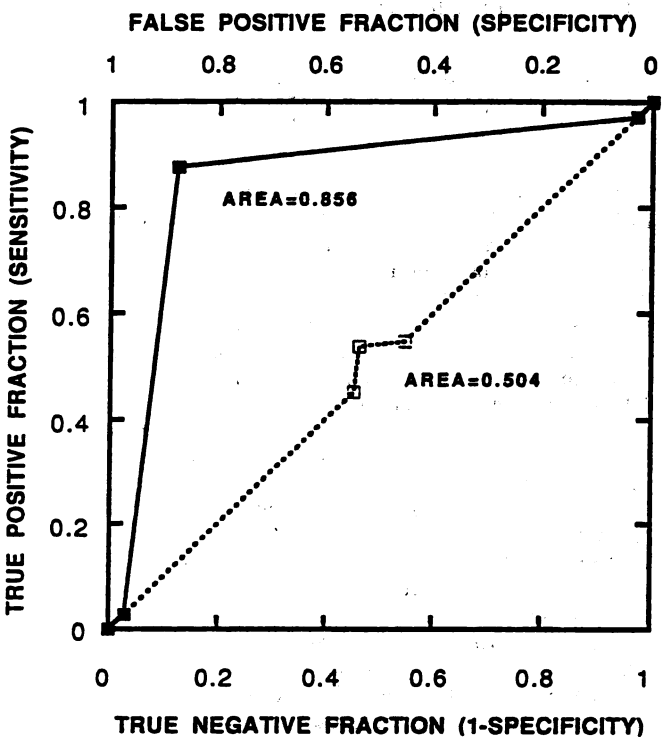

Fig. 11. ROC plots of two hypothetical tests that have identical likelihood ratios (four line segments with identical slopes), but different diagnostic accuracies

select a decision threshold (*5, 7*). As discussed above, each possible decision threshold for a test corresponds to a particular sensitivity/specificity pair. All of these pairs are graphed in the ROC plot. However, to use the test for patient management, a decision threshold must be selected. Two major elements determine which of the possible sensitivity/specificity combinations (and the corresponding decision threshold) is most appropriate for a particular application of the test (*7*): (*a*) The relative cost or undesirability of errors, i.e., false-positive and false-negative classifications; the value or benefits of correct classifications may also be considered (*6*). (*b*) The relative proportions of the two states of health that the test is intended to discriminate between. This is related to prevalence, or prior probability of disease, as shown below.

Assessing or assigning cost to false-positive or false-negative classifications is complex. This can be expressed in terms of financial costs or health costs and can be viewed from the perspective of the patient, the care providers, the insurers, dependents, society, etc. Nonetheless, some judgment about the relative costs of false results should be made when selecting rationally an operating decision threshold.

We use the simplified approach incorporating the ratio of the costs of false results (termed minimizing expected costs) to illustrate the relationship between ROC plots and selecting decision thresholds.

If, for example, the relative cost of a false-positive result is very much greater than the cost of a false-negative result, the appropriate sensitivity/specificity pair would favor specificity rather than sensitivity. However, to pick a sensitivity/specificity pair that yields false-negative and false-positive results in the optimal proportion, one must also incorporate the second factor, prevalence, because it interacts with the sensitivity and specificity, determining the actual probabilities of false-positive and false-negative results occurring in the population of interest.

The two elements *a* and *b* above are combined to calculate a slope (m) as follows (*7, 50*):

$$m = \left(\frac{\text{false-positive cost}}{\text{false-negative cost}}\right) \times \left(\frac{1 - P}{P}\right)$$

where P = prevalence or prior probability of disease. The point on the ROC plot where a line with this slope touches the curve is the best operating point, given the prevalence and the false-positive/false-negative cost ratio. If the ROC plot is smooth, as in the parametric model, the operating point is where the line is tangent to the curve. On the other hand, the nonparametric ROC plot for continuous data with no ties is a "staircase" of line segments having alternating slopes of zero and infinity. The operating point can be determined by the point where a line (with the above calculated slope), moving down from above and to the left, intersects the ROC plot. In both cases, this operating point corresponds to the decision threshold that will yield the optimal mix of false-positive and false-negative results, given P and the relative weights assigned to false results in the cost ratio. In Figure 12 the line with a slope m = 0.75 corresponds to a false-positive/false-negative cost ratio of 1/4 and a prevalence of 0.20. The line touches the plot at the point with a sensitivity of 0.890 and 1 − specificity of 0.551, corresponding to the following decision rule: a result is considered to indicate the presence of CAD if it is ≤0.192. Changing either P or the cost ratio changes the slope. If the prevalence were to be 0.10 instead of 0.20, *or* if the cost ratio were to be 3/4 instead of 1/4, then m = 2.25. Now this

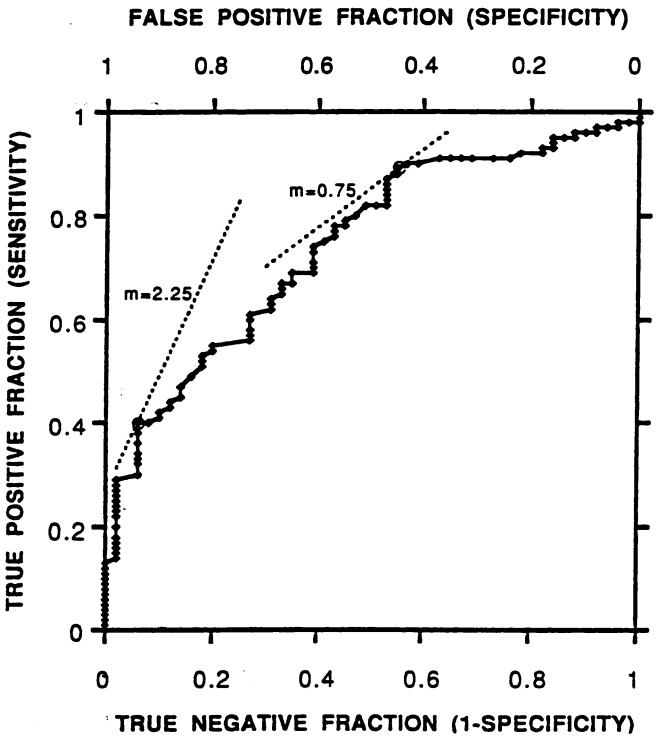

Fig. 12. Nonparametric ROC plot of the ratio of HDL to total cholesterol, showing optimal operating points for two sets of prevalence and costs (see text)

Subjects same as in Fig. 5 (data from Kottke et al., *18*)

line touches the ROC curve at a sensitivity of 0.404 and false-positive fraction of 0.061, corresponding to a shift of the threshold from 0.192 to 0.117. Thus, the ROC plot shows the effect of changing the false-positive/false-negative cost ratio or the prevalence and indicates the actual optimal operating point corresponding to a given set of circumstances (specified prevalence and cost ratio).

The operating point on the ROC plot as described above is the specificity, sensitivity pair that maximizes the function [sensitivity − m(1 − specificity)], where m is the above slope. If m = 1, the special case of the differential positive rate (PDR) as cited in a report by Pellar et al. (*51*) occurs. Grouping results into a small number of bins may yield a jagged plot that may, in turn, cause errors in locating the point where the line touches the curve, an issue noted by Pellar et al. (*51*). This is another reason to use the complete nonparametric ROC plot without binning data.

Efficiency and Predictive Value

If ROC analysis is the preferred approach to examine basic test performance, then how do concepts such as efficiency (percentage of correct results) and predictive value fit in? Whereas sensitivity and specificity describe the ability of the test to correctly distinguish between affected and unaffected subjects, respectively, predictive value and efficiency combine sensitivity, specificity, and prevalence to address the meaning of the results at one particular decision threshold. Sensitivity and specificity are properties inherent to the test; predictive value and efficiency are properties of the application once the context (decision threshold and prevalence) is established. Efficiency describes, for a given decision threshold and a particular prevalence, what fraction of all results are correct (true-positive and true-negative results). Predictive value of a positive result, PV(+), describes what fraction of all positive results are correct. Similarly, the predictive

value of a negative result, PV(−), describes what fraction of all negative results are correct results. Because prevalence is incorporated, efficiency and predictive value are not properties of the test alone, but the results of applying the test in one particular way. Predictive value, especially, is more an aid in interpreting a given test result than a measure of performance.

Efficiency has three important limitations:

- It is usually calculated at one decision threshold, although there are actually many efficiencies, one for each possible decision threshold (holding prevalence constant). It may be misleading to consider only one (or a few) because other, perhaps better, operating points may be overlooked.
- It is also highly dependent on disease prevalence in the study sample, and may appear high even when there is poor diagnostic sensitivity or specificity. The prevalence in the study sample may not be representative of the prevalence in the target population. Therefore, the prevalence used (or assumed) to calculate efficiency must be carefully chosen.
- It is defined as the percentage of true (or correct) results. Inherent in this definition is the concept that all true results are equally valuable and all false results (false-positive and false-negative results) equally costly or undesirable. This is often not true, however, and the greater the difference in the cost of a false-positive vs a false-negative result, the more the calculation of efficiency distorts the apparent clinical usefulness of the test.

Therefore, when efficiency is used or interpreted we must be mindful of the effect of prevalence and the possibility of misleading impressions if an inappropriate prevalence is being used. However, even when the correct prevalence is used, efficiency may present a distorted picture of test performance. Of the 18 papers in *Clinical Chemistry* mentioned earlier, one examining the use of fructosamine assays to distinguish between diabetic and nondiabetic workers involved a prevalence of diabetes of 2.3% in the study group (*52*). Merely by setting the decision threshold so high as to classify all subjects as negative, efficiency at that operating point would have been 97.7% despite a sensitivity of 0%! This result is due simply to the low prevalence. Whenever prevalence is high or low, it is possible to have a high percentage of correct results merely as a result of classifying one of the groups correctly. The farther prevalence is from 50%, the more this plays a role in observed efficiency.

Furthermore, when two tests are being compared, as in the above paper on fructosamine assays, and the prevalence is very low, both tests will have the potential for these very high efficiencies resulting from the effect of prevalence. This may mask real differences in accuracy between tests. Fortunately, in that paper ROC plots were generated, which provided a comprehensive picture of accuracy independent of prevalence for both assays.

Predictive value is essentially a calculation of the percentage of correct negative *or* of correct positive results. It shares, with efficiency, the first two of the above limitations. Whereas efficiency considers all results together, predictive value looks at only one class of results, either positive or negative, at a time. Therefore, when prevalence is low and the chosen decision threshold is high enough to classify most or all of the unaffected subjects correctly as true negatives, there will be so few false negatives (because there are so few affected subjects) that the PV(−) will be high even if every affected subject is incorrectly classified as negative (zero sensitivity)! Likewise, when prevalence is very high and the decision level is set low enough to classify all or most affected subjects correctly as true positives and all or most unaffected subjects incorrectly as false positives (poor specificity), the PV(+) will be very high. In this case, there are so few false-positive results, because the prevalence of unaffected individuals is so low, that the vast majority of positive results are correct! Thus, high efficiencies or predictive values may distort one's impression of the test's performance and may obscure a poor sensitivity or specificity.

In Figure 12, we examined two separate points on the ROC plot because when we changed either prevalence (P) or the false-positive/false-negative cost ratio, the slope (m) changed. Let's now examine the effect of this change in m on efficiency and predictive value. In the first case, where the line with m = 0.75 and prevalence = 0.20 touches the plot at a threshold of 0.192, the efficiency is 54% and PV(+) is 29%. However, when the prevalence changes to 0.10 from 0.20, it affects the calculation of m in the equation to optimize the decision threshold. The slope m changes from 0.75 to 2.25, and the line with that new slope touches the plot at a new point (and therefore a different specificity/sensitivity pair). Here the efficiency is 89% and PV(+) is 43%. As noted above, we can obtain the second slope (m = 2.25) by changing either prevalence to 0.10 *or* the false-positive/false-negative cost ratio from 1/4 to 3/4. When the second approach is used to change m to 2.25, the false-positive/false-negative cost ratio changes to 3/4 whereas prevalence remains at 0.20; now efficiency is 83% and PV(+) is 62%. Both efficiency and predictive value are derived from the interaction of the operating point on the ROC plot (sensitivity/specificity) with prevalence.

In summary, predictive value and efficiency provide limited information about the interaction of a particular point on the ROC plot with prevalence; in contrast, the ROC plot provides a more global comprehensive view of the test, independent of prevalence. Predictive value is more useful for interpreting a given result than for describing test performance.

Other Statistical Approaches

Other approaches besides predictive value and efficiency include discriminant analysis and logistic regression. The relationship of these to ROC plots is briefly discussed.

In logistic regression, the diagnostic test (and any covariates) is used in a prediction model that uses the logistic distribution to optimize the probability of a patient belonging to the diseased group. ROC plots are sometimes discussed in the context of logistic regression (*53, 54*). At the least, nonparametric ROC analysis can be used, in the absence of parametric assumptions, to evaluate the performance of logistic models. For a single test with no covariates, the ROC plot based on the logistic regression is identical to that for the original diagnostic test. The addition of other variables (covariate information) would create an improved ROC plot of the logistic model; these various ROCs could be examined by nonparametric ROC analysis to determine whether overall these other variables improve the predictive performance of the model. Here the logistic distribution is used to evaluate the predictive performance of the diagnostic test and other variables; an ROC plot of the logistic model could be used visually to describe its behavior as well as to aid in the identification of a decision threshold at which to operate.

Discriminant analysis is a statistical technique designed to arrive at a decision rule and is also related to the ROC plot. In fact, discriminant analysis for a single test variable can be thought of as a statistical device used, after making the visual ROC plot, to decide where on the ROC plot to select a decision rule. Discriminant analysis in its most

familiar form relies on the assumption that the (multiple) diagnostic test values are (multivariately) normally distributed for the two groups of patients. As such, it imposes a parametric model and then attempts to find the linear combination of the test values that best separates the two groups. Discriminant analysis is usually implemented so that one can control the prevalence and the trade-off of the errors in the decision rule in this model, as is also done in the ROC approach. In the case of a single variable (one diagnostic test), classical discriminant analysis is equivalent to assuming that both group distributions are normal (gaussian)—or in logistic-discriminant analysis, assuming that the ratio of densities is logistic—estimating the parametric ROC curve for this model, and then selecting a single optimal ROC point (decision threshold) based on the trade-off of error cost and prevalence. In contrast, the nonparametric ROC plot assumes no such parametric model and yet provides information on the entire performance of the diagnostic test. In the multivariate case (more than a single test on each subject), whereas the ROC plot can represent only the performance of one particular function (linear or quadratic) of the variables, classical or logistic-discriminant analysis is advantageous because it provides the linear- or quadratic-discriminant function that defines the optimal decision rule.

Importance of Study Design

One benefit of this concept of diagnostic accuracy, and of ROC plotting and analysis in particular, is that it calls attention to the issue of study design. Whether a test is being assessed alone or being compared with other tests, several aspects of study design have important influence on the outcome (*13, 55, 56*). The clinical question being addressed should be clearly defined and stated to help prevent misleading conclusions about what discrimination the test is capable of achieving and to aid in the proper selection of subjects to study. If the test is to be used to identify subjects with AMI among subjects presenting to an emergency room with typical chest pain and other symptoms suggestive of AMI, then the study group should be composed of a sample of just such subjects. Healthy laboratory workers or blood donors would not be appropriate because the clinical issue was not identification of AMI in asymptomatic volunteers. Similarly, if a test is being evaluated for identification of prostatic cancer in men older than 50 years or symptomatic men, then the subjects should not include younger men in the former case or asymptomatic men in the latter case. Once the proper group is selected, each member should be classified (e.g., AMI or no AMI; presence or absence of prostatic cancer) definitively and independently of the test. The true-positive fractions are then calculated from the group having AMIs, and the false-positive fractions are calculated from the subjects who did not have an AMI. Improper selection of subjects and inaccurate diagnostic classification of these subjects may distort the apparent accuracy of the tests and result in erroneous conclusions. A major source of bias is the so-called verification bias, in which attention is restricted only to confirmed (or verified) cases and controls (*57–60*). If only the easily diagnosed cases are classified and the others are discarded, then specificity and sensitivity are generally overestimated and hence the ROC plot is biased. There has been considerable effort in the literature to compensate for verification bias; some success has been obtained by using a logistic-regression model for the verified data on the covariates to predict the verification probability in the unconfirmed subjects (*61*). Generally, generating ROC plots or performing sophisticated statistical ROC analysis will not compensate for design flaws. The validity or usefulness of the ROC approach depends, ultimately, on the soundness of the study design.

Earlier, we noted that sensitivity and specificity (and thus the ROC plot) were properties of the test itself and independent of factors, particularly prevalence (prior probability of disease), that are properties of the circumstances. However, sensitivity and specificity do depend on the nature of the subjects selected for evaluating the test. For instance, in the above discussion referring to prostatic cancer, prostate-specific antigen may exhibit different sensitivities and specificities in symptomatic elderly men than in younger, asymptomatic men. In these two groups, the type and extent of symptoms will be different, and the prevalence and severity of both malignant and nonmalignant disease will differ. Thus, it may appear that sensitivity and specificity depend on prevalence; in fact, however, they depend on the spectrum of disease in the subjects studied. This effect of composition of the study sample will not introduce bias if the subjects recruited for the study are representative of the subjects relevant to the question being posed. Lachs et al. (*62*) clearly note this in their study of the performance of the leukocyte esterase and bacterial nitrite dipstick tests for detecting urinary tract infection. Sensitivities and specificities were higher in a group of patients with numerous and typical signs and symptoms (and a higher prevalence of infection) than in a group with fewer signs and symptoms (lower prevalence of infection). This was due to the spectrum of disease in each group, rather than to prevalence (*13, 55, 63*). Information such as age or number of symptoms can also be used to model specificity and sensitivity and hence the ROC plot based on the logistic regression model (*53*).

Some researchers have considered what to do if there is no gold standard; i.e., no errorless identification of disease and nondisease. One strategy is to define the diagnostic problem in terms of measurable clinical outcomes (*64*). A second approach is to use some sort of consensus, majority rule, or expert review to arrive at a less error-prone identification process (*65*). A third solution that is more applicable for parametric models is to assume for the comparison of several accurate tests that the subject population consists of some unknown mixture of diseased and control subjects and then to estimate this mixture parameter as well as the other parameters (*66*). A fourth approach, rather than definitively assigning each such patient to one of the groups, say, diseased or nondiseased (reference), is to assign to each a value between 0 and 1 that corresponds to the (subjective) assessment of how likely it is that this patient belongs to the diseased group (this could be accomplished by logistic regression). Then there is no need to discard the data from these gray, fuzzy cases where group assignment is not unequivocal. An outgrowth of the fuzzy approach is that one no longer needs to treat as the same two test results of, e.g., 16 and 80, for a test for which the decision threshold is 15. A fuzzy or probabilistic analysis for ROC plots has been undertaken (*67–69*).

Computer Software for ROC Plotting and Analysis

There are several commercial and public domain software products for ROC analysis. Table 2 lists most of these and highlights their capabilities. Contacts for these products are as follows:

CLINROC. Henry T. Sugiura and George A. Hermann, R. Phillip Custer Laboratories, Presbyterian University of Pennsylvania Medical Center, 39th & Market St., Philadelphia, PA 19104. CLINROC does not produce its para-

Table 2. Characteristics of ROC Software

	CLINROC	Metz	ROC ANALYZER	ROCLAB	RULEMAKER	SIGNAL	TEP-UH
Continuous or Bin data[a]	C(14)	B(11), C(20)	B(17)	C	C	B(21)	C
CI for sensitivity	—	P	—	NP	—	—	NP
ROC plot	NP	—	NP, P	NP	NP	NP, P	NP
ROC area estimate	NP	NP, P	NP, P	NP	NP	NP, P	NP
SD of area	NP	NP, P	NP, P	—	NP	—	NP
CI for area	—	—	—	—	NP	—	NP
Test if area = 0.5	—	—	—	—	NP	—	NP
MLE estimation	—	G	G	—	G	G, O	—
Choice of decision threshold	P	—	—	NP	—	—	NP[c]
Likelihood ratio (LR)	P	—	NP	—	NP	P	NP
CI for LR	—	—	NP	—	—	—	NP
Compare two ROC areas	—	P	NP, P	—	NP	—	—
Output file for ROC graph	—	Y	Y	Y	Y	Y	Y
Commercial, public domain, or shareware[b]	PD	PD	S	PD	C	C	S
Macintosh, PC, or mainframe (MF)	PC, MF	M, PC, MF	PC	PC	M	PC	PC, MF

CI, confidence interval; NP, nonparametric; P, parametric; G, gaussian (normal) distribution; O, other; Y, yes.

[a] C(n) denotes continuous input, binned into a maximum of n categories for the analysis; B(n) denotes binned (ordered categorical) input, with up to n bins for parametric analysis.

[b] Shareware implies a small fee but the enterprise is not commercial.

[c] m = 1.

metric analysis of likelihood ratios through maximum likelihood methods but rather based on the assumption of normality in the original or log-transformed scale.

Metz programs: LABROC1, CLINROC, ROCFIT, CORROC. Charles E. Metz, Department of Radiology, MC2026, The University of Chicago Medical Center, 5841 S. Maryland Ave., Chicago, IL 60637-1470 [FAX (312)702-6779, Internet address: c-metz[@]uchicago.edu]. The Metz programs are, for a single diagnostic test, LABROC1 for continuous data and ROCFIT for discrete data, and, for two correlated tests, CLABROC and CORROC, respectively. There are slight differences in Table 2 entries for versions on the different computer platforms. Program requesters are asked to specify platform and to include for microcomputer requests two appropriate floppy disks.

ROC ANALYZER. Robert M. Centor, 10806 Stoneycreek Drive, Richmond, VA 23233 [Bitnet address: Centor[@]VCUVAX on BITNET]. This program is described by Centor and Keightley (*70*).

ROCLAB. James M. DeLeo, Bldg. 12A, Room 2013, Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20892 [Bitnet address: deleo.nihdcrt[@]cu.nih.gov]. ROCLAB provides maximal as well as trapezoidal areas for ties. It has the ability to do ROC plots for fuzzy data as well.

RULEMAKER. Digital Medicine, Inc., Hanover, NH 03755 [FAX (603) 643-3686]. RULEMAKER has an interactive capability that enables one to point to a location on the ROC plot and obtain the exact sensitivity and specificity and the decision level to which it corresponds. New decision rules can be formed by combining results from two or more diagnostic tests and the corresponding ROC plot can be displayed. A release version is anticipated in the second half of 1993.

SIGNAL. SYSTAT, Inc., 1800 Sherman Ave., Evanston, IL 60201. SIGNAL is a module of a much larger commercial package SYSTAT.

TEP-UH (Test Evaluation Program-University Hospital). Thomas G. Pellar, Department of Clinical Biochemistry, University Hospital, P.O. Box 5339, 339 Windemere Road, London, Ontario, Canada N6A 5A5. Running TEP-UH requires the parent program MUMPS (Micronetics Design Corp., Rockville, MD).

Note that only three of the programs are designed to treat the continuous data directly, without binning (forcing into discrete intervals) the data.

Examples from the Literature

Van Steirteghem et al. (*16*) compared the accuracies of myoglobin, CK-BB, CK-MB, and total CK in discriminating among persons presenting to an emergency room with typical chest pain, with and without AMI. ROC plots could be constructed for any sampling time by using measurements on multiple closely sequential serum samples timed from the onset of pain. The plots showed clearly the superior accuracy of myoglobin at early times such as 5 or 8 h, as well as the impressive accuracy of CK and its isoenzymes at 18 h after the onset of pain. More recently, Leung et al. (*71*) performed a similarly detailed evaluation of total CK and CK-2 in 310 patients admitted to a cardiac care unit with chest pain. These authors also used ROC plots to describe the changing accuracy at various time intervals after the onset of pain.

Carson et al. (*72*) investigated the abilities of four assays of prostatic acid phosphatase to discriminate between those subjects with prostatic cancer and those subjects with either some other urologic abnormality or no known urologic abnormality. They concluded from comparisons of ROC plots and areas under the plots that there is little difference in diagnostic accuracy among the four assays. Because ROC was used, the conclusions were not influenced by choice of "upper limit of normal." They cited previous studies that had claimed superior performance of

certain assays of prostatic acid phosphatase, noting that these reports had generally not considered the influence of upper limits and may have overstated the differences between assays.

Hermann et al. (*73*) compared the diagnostic accuracies of two versions of a commercial assay for thyrotropin to test a claim that the newer one was superior for discriminating between euthyroidism and hyperthyroidism. On the basis of ROC plots and areas under the plots, the authors concluded that the newer version exhibited a small but significant superiority in diagnostic accuracy.

Kazmierczak et al. (*74*) used ROC plots in a study of the accuracies of lipase, amylase, and phospholipase A in discriminating acute pancreatitis from other diseases in 191 consecutive patients seen with abdominal pain. ROC plots clearly demonstrated that, in this group of subjects, amylase and lipase performed similarly and both were superior to phospholipase A.

Flack et al. (*75*) used ROC plots and areas to compare the abilities of urinary free cortisol and 17-hydroxysteroid suppression tests to discriminate between Cushing disease and other causes of Cushing syndrome. Evaluating sensitivity and specificity by conventional decision criteria suggested differences, whereas ROC plots showed clearly that the diagnostic accuracies of the two tests were essentially equivalent.

Guyatt et al. (*76*) studied the ability of seven tests including ferritin, transferrin, saturation, mean cell volume, and erythrocyte protoporphyrin to discriminate between iron-deficiency anemia and other causes of anemia in subjects older than 65 admitted to the hospital with anemia. ROC plots showed that serum ferritin "performed far better than any of the other tests." In calculating areas, a factor was used to correct for correlation because all plots were generated from the same cohort of subjects. Beck (*77*), studying iron-deficiency anemia, also used ROC plots to compare several tests for "predicting the presence or absence of bone marrow iron stores."

In summary, the ROC plot, representing the fundamental ability of a test to discriminate between two states of health, is an index of pure accuracy. A nonparametric ROC plot is an unbiased view of a test's performance (accuracy) in a defined clinical setting. The ROC plot itself and ROC analysis provide information useful to the clinical laboratorian in making practical decisions about laboratory operation. Furthermore, the ROC plot is a springboard to several pathways (Figure 1) to further exploring test performance and to clinical application.

We thank Edward Shultz and Digital Medicine, Inc., for the use of RULEMAKER™ software and Robert Centor for ROC ANALYZER, which we used for our ROC analyses. We thank James M. DeLeo for providing ROCLAB, which was used to generate the ROC plots.

References

1. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–93.
2. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
3. Lusted LB. ROC recollected [Editorial]. Med Decis Making 1984;4:131–5.
4. Green DM, Swets JA. Signal detection theory and psychophysics. New York: John Wiley & Sons, Inc., 1966.
5. Lusted LB. Decision making studies in patient management. N Engl J Med 1971;284:416–24.
6. Lusted LB. Signal detectability and medical decision-making. Science 1971;171:1217–9.
7. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975;293:211–5.
8. Swets JA, Pickett RM. Evaluation of diagnostic systems. New York: Academic Press, 1982.
9. Galen RS, Gambino SR. Beyond normality: the predictive value and efficiency of medical diagnoses. New York: John Wiley & Sons, Inc., 1975:9–11.
10. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
11. Turner DA. An intuitive approach to receiver operating characteristic curve analysis. J Nucl Med 1978;19:213–20.
12. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: WB Saunders, 1980:114–30.
13. Robertson EA, Zweig MH, Van Steirteghem AC. Evaluating the clinical efficacy of laboratory tests. Am J Clin Pathol 1983;79: 78–86.
14. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13–20.
15. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989;29:307–35.
16. Van Steirteghem AC, Zweig MH, Robertson EA, Bernard RM, Putzeys GA, Bieva CJ. Comparison of the effectiveness of four clinical chemical assays in classifying patients with chest pain. Clin Chem 1982;28:1319–24.
17. Lott JA, Lu CJ. Lipase isoforms and amylase isoenzymes: assays and application in the diagnosis of acute pancreatitis. Clin Chem 1991;37:361–8.
18. Kottke BA, Zinsmeister AR, Holmes DR Jr, Kneller RW, Hallaway BJ, Mao SJT. Apolipoproteins and coronary artery disease. Mayo Clinic Proc 1986;61:313–20.
19. Guignard PA, Salehi N. Validity of the gaussian assumption in the analysis of ROC data obtained from scintigraphic images. Phys Med Biol 1983;28:1409–17.
20. Swets JA. Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psych Bull 1986;99:181–98.
21. Rockette HE, Obuchowski NA, Gur D. Nonparametric estimation of degenerate ROC data sets used for comparison of imaging systems. Invest Radiol 1990;25:835–7.
22. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. Stat Med 1990;9;325–37.
23. Hanley JA. The robustness of the "binormal" assumption used in fitting ROC curves. Med Decis Making 1988;8:197–203.
24. Centor RM. Signal detectability: the use of ROC analysis. Med Decis Making 1991;11:102–6.
25. Egan JP. Signal detection theory and ROC analysis. New York: Academic Press, 1975.
26. Hilgers RA. Distribution-free confidence bounds for ROC curves. Methods Inf Med 1991;30:96–101.
27. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. Biometrics 1950;6:399–412.
28. Grey DR, Morgan BJT. Some aspects of ROC curve-fitting: normal and logistic models. J Math Psychol 1972;9:128–39.
29. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. J Math Psychol 1969;6:487–96.
30. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 1984;2:137–50.
31. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic curve. J Math Psychol 1975;12:387–415.
32. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
33. Hollander M, Wolfe DA. Nonparametric statistical methods. New York: John Wiley and Sons, 1973:67–78.
34. Hartz AJ. Overlap index: an alternative to sensitivity and specificity in comparing the utility of a laboratory test. Arch Pathol Lab Med 1984;108:65–7.
35. Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11:95–101.
36. Swets JA. ROC curve analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109–21.
37. Metz CE. Some practical issues of experimental design and

data analysis in radiological ROC studies. Invest Radiol 1989;24: 234–45.
38. Centor RM, Schwartz JS. An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. Med Decis Making 1985;5:149–56.
39. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76:585–92.
40. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–5.
41. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
42. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
43. Campbell G, Douglas MA, Bailey JJ. Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve. In: Ripley KL, Murray A, eds. Computers in cardiology. Washington, DC: IEEE Computer Society, 1989:267–70.
44. Metz E, Wang P-L, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, ed. Information processing in medical imaging: proceedings of the eighth conference. The Hague: Martinus Nyhoff, 1984:432–45.
45. Lusted LB. Introduction to medical decision making. Springfield, IL: Charles C Thomas, 1968.
46. Albert A. On the use and computation of likelihood ratios in clinical chemistry. Clin Chem 1982;28:1113–9.
47. Radack RL, Rouan G, Hedges J. The likelihood ratio: an improved measure for reporting and evaluating diagnostic test results. Arch Pathol Lab Med 1986;110:689–93.
48. Suchman AL, Dolan JG. Odds and likelihood ratios. In: Griner PF, Panzer RJ, Greenland P, eds. Clinical diagnosis and the laboratory. Logical strategies for common medical problems. Chicago: Year Book Medical Publishers, 1986:36–43.
49. Centor RM. Estimating confidence intervals of likelihood ratios. Med Decis Making 1992;12:229–33.
50. Linnet K. A review on the methodology for assessing diagnostic tests. Clin Chem 1988;34:1379–86.
51. Pellar TG, Leung FY, Henderson AR. A computer program for rapid generation of receiver operating characteristic curves and likelihood ratios in the evaluation of diagnostic test. Ann Clin Biochem 1988;25:411–6.
52. Baker J, Metcalf P, Scragg R, Johnson R. Fructosamine Test-Plus, a modified fructosamine assay evaluated. Clin Chem 1991;37:552–6.
53. Albert A, Harris EK. Multivariate interpretation of clinical laboratory data. New York: Marcel Dekker, Inc., 1987.
54. SAS Institute Inc. SUGI supplemental library user's guide, version 5 ed. Cary, NC: SAS Institute, 1986:273.
55. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978;299: 926–30.
56. Zweig MH, Robertson EA. Why we need better test evaluations [Opinion]. Clin Chem 1982;28:1272–6.
57. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.
58. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39: 207–15.
59. Begg CB, McNeil BJ. Assessment of radiological tests: control of bias and other design considerations. Radiology 1988;167:565–9.
60. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 1984;4:151–64.
61. Hunink MG, Begg CB. Diamond's correction method—a real gem or just cubic zirconium. Med Decis Making 1991;11:201–3.
62. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992;117:135–40.
63. Gur D, King JL, Rockette HE, Britton CA, Thaete EL, Hoy RJ. Practical issues of experimental ROC analysis. Selection of controls. Invest Radiol 1990;25:583–6.
64. Valenstein PN. Evaluating diagnostic tests with imperfect standards. Am J Clin Pathol 1990;93:252–8.
65. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983;18: 194–8.
66. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. Med Decis Making 1990;10:24–9.
67. Campbell G, DeLeo JM. Fundamentals of fuzzy receiver operating characteristic (ROC) functions. In: Malone L, Berk K, eds. Computing science and statistics: proceedings of the twenty-first symposium on the interface. Alexandria, VA: American Statistical Assoc., 1989:543–8.
68. DeLeo JM, Campbell G. The fuzzy receiver operating characteristic function and medical decisions with uncertainty. Proc. First Int Symp on Uncertainty Modeling and Analysis. IEEE Computer Society Press, 1990:694–9.
69. Campbell G, Levy D, Bailey JJ. Bootstrap comparison of fuzzy R.O.C. curves for ECG-LVH algorithms using data from the Framingham heart study. J Electrocardiol 1990;23(Suppl):132–7.
70. Centor RM, Keightley GE. Receiver operating characteristic (ROC) curve area analysis using the ROC ANALYZER. SCAMC Proc 1989:222–6.
71. Leung FY, Galbraith LV, Jablonsky G, Henderson AR. Reevaluation of the diagnostic utility of serum total creatine kinase and creatine kinase-2 in myocardial infarction. Clin Chem 1989; 35:1435–40.
72. Carson JL, Eisenberg JM, Shaw LM, Kundel HL, Soper KA. Diagnostic accuracy of four assays of prostatic acid phosphatase. Comparison using receiver operating characteristic curve analysis. J Am Med Assoc 1985;253:665–9.
73. Hermann GA, Sugiura HT, Krumm RP. Comparison of thyrotropin assays by relative operating characteristics analysis. Arch Pathol Lab Med 1986;110:21–5.
74. Kazmierczak SC, Van Leute F, Hodges ED. Diagnostic and prognostic utility of phospholipase A activity in patients with acute pancreatitis: comparison with amylase and lipase. Clin Chem 1991;37:356–60.
75. Flack MR, Oldfield EH, Cutler GB, et al. Urine free cortisol in the high-dose dexamethasone suppression test for the differential diagnosis of the Cushing syndrome. Ann Intern Med 1992;116: 211–7.
76. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med 1992;7:145–53.
77. Beck JR. The role of new laboratory tests in clinical-decision making. Clin Lab Med 1982;2:751–77.