RECEIVER OPERATING CHARACTERISTIC (ROC) METHODOLOGY: THE STATE OF THE ART*

Author: **James A. Hanley**
Department of Epidemiology and Biostatistics; and
Department of Medicine
McGill University
Montreal, Quebec, Canada

Referee: Charles E. Metz
Department of Radiology
University of Chicago
Chicago, Illinois

I. INTRODUCTION

A. What Is ROC Analysis?

Receiver operating characteristic (ROC) analysis is a procedure, derived from statistical decision theory, that was developed in the context of electronic signal detection.[1] It is used in medical applications to evaluate the accuracy of diagnostic and prognostic technologies. It is particularly relevant for conducting observer performance tests, with real or simulated clinical "cases", to evaluate diagnostic systems that require (subjective) operator interpretation; it can also be used to evaluate the discriminatory performance of the more straightforward quantitative diagnostic tests and prognostic indices.

B. Biomedical Applications: The Past 10 Years

The use of ROC analysis in the medical literature has grown considerably in the past 10 years. A search of the Medline data base for the most recent years, using the words "ROC" and "observer performance", yielded more than 30 articles per year containing such analyses; judging from the numbers of articles found through other sources that were not identified by the Medline search, the real number of articles using ROC is now probably at least 50 per year. In this 10-year time span, the methodology for the design and statistical analysis of ROC studies has itself undergone substantial development, although, as will be discussed, there are still some gaps. One can appreciate the methodologic progress by comparing the two papers by Metz,[2,3] published in 1978 and 1986, respectively.

Until 1982, ROC methodology was treated mainly in the psychophysical literature, although a small number of expository and applied articles had appeared in *Science* and in specialty journals in radiology and nuclear medicine. The 1982 text by Swets and Pickett[4] brought together the theory and the practice of ROC analysis for biomedical applications; it provided a comprehensive guide to study design, data analysis, and interpretation, and illustrated the principles and the practical considerations using case studies from the authors' own experience in the assessment of medical diagnostic technology. The two journals *Investigative Radiology* and *Medical Decision Making* (1981—) provide a continuing forum for tutorials and new methodologic developments as they apply to biomedical contexts; unfortunately, methods which are developed and reported in the psychology literature are often slower to reach biomedical investigators. To date, there has been only isolated methodological interest from biostatisticians, although some papers dealing with statistical analysis of ROC data have just recently begun to appear in professional statistical journals.

* Supported by an operating grant from the Natural Sciences and Engineering Council of Canada. Colin Begg, Carl Brewer, Barbara McNeil, and John Swets provided helpful comments.

Reprinted from the Critical Reviews in Diagnostic Imaging, Vol. 29, Issue 3, pages 307-335.

C. Adapting ROC Methodology to Biomedical Contexts

ROC methodology dates back to the early 1950s, when it was developed to describe and summarize the data from signal detection experiments in radar and other psychophysical research. It is closely linked to statistical decision theory and to the practice of quality control, where the term "operating characteristic" was used to describe the performance of an inspection rule which tried to distinguish bad batches from good on the basis of samples. The best textbook source for the basic fundamental theory of signal detection is still the book by Green and Swets.

While the basic concepts underlying ROC analysis do not change when it is applied to medical diagnosis, the experimental challenges and obstacles become substantially larger. The fact that the "signals" emanate from humans rather than from inanimate sources introduces many methodologic issues that are not encountered in detection tasks involving simple signals. In such psychophysical experiments, stimuli cost very little to generate; each subject can evaluate several hundred stimuli per session, and sessions can be repeated using different decision criteria. Contrast this with medical diagnostic tasks, where signals are highly variable, where appropriate biological "stimuli" are scarce and are complex enough that it may take several minutes to interpret each one, and where few suitable subjects are available at any one research location. Clearly, the original ROC methodology cannot just be ad*o*pted; it needs to be ad*a*pted to fit the increased complexity of medical diagnosis.

D. Scope, Intent, and Organization of This Review

It was difficult to decide toward whom, and at what technical level, this review should be aimed. One possible target is those investigators who need to be aware of ROC analysis but still are not; judging from the biomedical literature and from recent reviews of research proposals to compare diagnostic modalities, this is a large group. A second possible target are those who are already familiar with, or have used, ROC methodology. To many of those in this (much smaller) group, the statistical techniques required appear to be more complex than those in other biomedical applications; in particular, the parametric curve fitting and subsequent statistical inferences may seem especially mysterious and model intensive in relation to the small numbers of empirical data points. A third possible target is the (still smaller) group of individuals who are developing new analytic techniques: it is difficult to resist using some of this review as a forum for methodologic debate about their recent work.

Rather than address itself to just one of these groups, the review tries to include items of interest to all three. Many of the concepts will have been expressed before, but they bear repeating or being stated another way. Fortunately, in the last 3 years, a number of major articles and reviews have covered various aspects of ROC methodology.[3,5-7] This allows this reviewer to go lightly over some of these aspects and to concentrate a little more on others that are of personal methodological interest.

The review is organized along three main lines: the need for, and rationale behind, ROC studies; the design of ROC studies; and the analysis of ROC data. The treatment of data analysis may seem unduly long in relation to that of design; this is, in part, because it contains some items, such as consideration of sample sizes, that more properly belong under study design, but that are best appreciated when discussed in the context of the methods of analysis.

II. THE NEED FOR, AND BASIS OF, ROC ANALYSIS

The need for, and the basic principles of, ROC analysis in biomedical contexts have been described by several authors, notably Lusted,[8] McNeil et al.,[9] Metz,[3] Swets et al.,[10] Swets,[11] Turner,[12] and Weinstein and Fineberg.[13] The fundamental theory, based on the signal detection model, is given in the textbook by Green and Swets[1] and is best illustrated in the applications described in the applied text by Swets and Pickett.[4] However, in the spirit of accommodating "entry level" readers, a short introduction is given here.

A. Purpose of ROC

Swets[14] describes ROC analysis as "a method of quantifying how accurately experimental subjects, professional diagnosticians and prognosticators (and their various tools) perform when they are required to make a series of fine discriminations or to say which of two conditions or states of nature, confusable at the moment of decision, exists or will exist." By allowing the correlation between data-based diagnoses (prognoses) and actual states of present or future health of patients to be quantified, ROC analysis can be used to evaluate the accuracy of a medical diagnostic (or prognostic) system. As of now, ROC analysis is limited to discrimination between two states (outcomes); this review will refer to the two types of patients as diseased and nondiseased, or D+ and D–. Likewise, the term "diagnostic system" will be used generically. In some applications, it will refer to a combination of elements, such as a data-collection and data-display device, and a human interpreter or "reader", or "diagnostician", or "observer"; in others, it may simply refer to a test that yields a numerical result. Very often, the interest is in the comparative performance of two "systems", e.g., we may wish to compare the performances of two competing biochemical tests, a biochemical and an imaging test, or the performance of observers in two or more diagnostic modalities, which may differ in the way they collect, process, or display data, or in what additional information the observers have access to. ROC techniques can also be used to compare two *specific* observers using the same technology (as, for example, in deciding which one to use for a mammography or cytology study) or to test the signal detection performance of trainees. However, the more usual application is in studies where one wishes to generalize from a sample of observers to observers in general.

B. Measuring Diagnostic Accuracy: The Need for ROC

The considerable variation in the use of the term "diagnostic accuracy" illustrates the need for a sophisticated approach such as that provided by ROC. The simplest index of accuracy is the proportion of patients for which the diagnosis is correct; it is also the most naive, in that it is heavily influenced by the actual proportions (p) and (1 – p) of the two states D+ and D–. Gilbert pointed this out 100 years ago,[15] when he explained the inordinately high "accuracy" a fellow meteorologist claimed in predicting tornadoes.[16] Using a degree of politeness seldom seen in modern-day letters to the editor, he pointed out that because the actual frequency (p) of tornadoes was so low, this high accuracy could be achieved by simply "calling" for "no tornado" each day.

One possibility is to separate the decisions or predictions into two groups according to the two states of nature, i.e., to report separate "proportions correct" for the D+ and D– groups. These two quantities are known by various names; in the medical diagnostic literature, they are referred to as sensitivity and specificity or as the true-positive (TP) and true-negative (TN) fractions. Their complements, the false negative (FN) fraction = 1 – TP and false-positive (FP) fraction = 1 – TN, are also used. (The latter are called type II and type I errors by statisticians.) Incidentally, a recent article has pointed out that statistical tests are themselves forms of diagnostic tests and that the concept of statistical power can be viewed as the "statistical sensitivity" of an investigation.[17] The various terms are best defined using the kind of 2×2 table displayed in Table 1. The naive "overall proportion correct" is a weighted sum of the two state-specific accuracies with weights p and (1 – p).[3,4]

Unfortunately, using two separate performance measures still does not allow us to compare one diagnostic test with another, or even with the same test carried out in another setting or by another observer. Some of the reasons are illustrated in a study by Harris.[18] He noted that sensitivity and specificity values are often regarded as reliable mathematical "constants" in calculations such as those employed in decision analysis. He examined the literature on seven diagnostic tests and found that for five tests there was significant variation among the reported sensitivities and specificities. He warned the "significant variability among the reported results" raised difficult questions about most of the tests and about the decision-analytic approach in general. He acknowledged several reasons for this unacceptable variability, including (1)

Table 1
ILLUSTRATIVE DATA TABLE AND MEASURES OF PERFORMANCE DERIVED FROM A BINARY DIAGNOSTIC TEST

		Test result		
		–	+	Total
Data table	D–	360	40	400
	D+	20	80	100
Performance				
In D– group				
False-positive fraction			FP = 0.10 (40/400)	
Specificity, i.e., true-negative fraction		TN (= 1 – FP): 0.90		
In D+ group				
Sensitivity, i.e., true-positive fraction			TP = 0.80 (80/100)	
False-negative fraction		FN (= 1 – FP): 0.20		

Note: D+ denotes diseased; D– denotes nondiseased.

technical limitations that may prevent the test from being completely reproducible, (2) interobserver variability in test interpretation, (3) biased populations or biased interpretations, (4) comparison with a criterion or end point that also has inherent variability, and (5) uninterpretable test results. His message prompted a reply from Swets,[19] who felt that Harris was being unduly harsh, that such pessimism "misrepresents the inherent variability in data on diagnostic tests that can be collected in well-designed studies" and that it "exaggerates the effective variability of data on tests now in the literature." Swets expanded on reason (2) given by Harris by explaining that the sensitivity and specificity values associated with a diagnostic test are subject to variation on two independent dimensions: (1) the test's capacity to discriminate a given disease from nondisease, and (2) the decision criterion (confidence level or cutting score) that is adopted for declaring a test result to be positive. "Thus a test with constant discrimination capability can have measured sensitivity values that vary from 0.0 to 1.0 and specificity values that vary from 1.0 to 0.0, as the decision criterion varies from conservative to liberal."

This plot of the TP fraction as a function of the FP fraction is called a ROC curve. In his analysis, Swets took the pair of TP and FP values from each study and plotted them as two-dimensional data points on FP:TP axes; he was able to show that, after allowance for the sample sizes and other factors in the different studies, the data points from the different studies of any particular test fell on a single ROC curve, illustrating that the test's capacity to discriminate a given disease from nondisease is relatively constant over studies, and that the apparent variability may reflect different choices of the decision criterion. Using the ROC device of obtaining TPs for a range of FPs, accuracy studies can be designed expressly to control the decision criterion and remove the effect of its variation when evaluating a test's capacity to discriminate disease from nondisease.

C. Essence of ROC

ROC principles are based on the notion of a "decision variable".[1,3] This concept is needed because very few clinical diagnostic tests produce results which fall into two obviously defined categories with unequivocal boundaries. With some tests, an "explicit" variable exists; for example, a biochemical test may yield a numerical test result. In such situations, one can choose

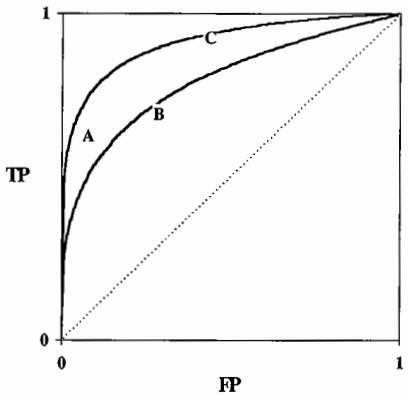

FIGURE 1. Three diagnostic tests on two ROC curves. Diagnostic tests with (TP, FP) values lying on the upper ROC curve are more discriminating than those lying on the lower one. Although test C appears to have higher sensitivity (TP) than test A, it has the same inherent discriminating potential; the difference between A and C is entirely explained by the fact that at the particular operating point, test C has poorer specificity (higher FP) than test A. In contrast, although test B also appears to have higher sensitivity than test A, it actually has poorer discriminatory ability than both A and C when specificity is held constant.

among an infinity of decision thresholds or cut points along the continuum of this variable to serve as the boundary above which one would declare the test "positive"; each different choice will yield a different sensitivity and a different specificity, with a decrease in one being accompanied by an increase in the other. When a diagnostic test can only be interpreted subjectively, the scale and the possible boundaries are "implicit" or "latent", i.e., they exist only in the mind of the observer. However, the concept of a ROC curve remains the same: it is the curve showing the trade-off between TP successes and FP errors (or equivalently, between sensitivity and specificity) as one employs different decision boundaries.

By isolating the perturbations due to the use of different decision criteria, and representing them as different points on the same curve, ROC analysis allows one to directly compare the inherent diagnostic capabilities of different diagnostic tests. As is shown in Figure 1, a more accurate test will be located on a curve closer to the top left corner than a less accurate one. A noninformative test will have a ROC curve that lies along the diagonal.

As was stated earlier, ROC analysis is best suited to evaluating observer performance in distinguishing between *two* states; it has not yet been adapted to deal simultaneously with *three or more* states. Although not reviewed further here, ROC methods can, however, be adapted to evaluate observer performance in detecting and localizing abnormalities — tasks which are frequently faced in imaging modalities.[3,4] Recently, statistical procedures have been suggested for the "free response operating characteristic" (FROC) often used to test detection of multiple tumor nodules on chest films.[20] The approach uses an older technique,[21] which was developed to deal with multiple signals embedded in a uniform background. It is not yet clear if the transfer of the methods to clinical studies (with such nonuniform anatomical backgrounds) can be successful.

III. DESIGN OF ROC STUDIES

The main issues to be faced in the design of a ROC study are deciding the exact nature of the question to be answered, choosing the appropriate type and amount of case material, deciding

the types and numbers of observers to employ, and establishing the protocol for the conduct of the reading tests. The text by Swets and Pickett covers many of these in detail.

A. Study Objectives

One must carefully specify what the diagnostic system consists of, the clinical setting, and what type of diagnostic task it is being used to perform. Rozanski et al.[22] remind us that the intent of a test, i.e., what types of patient it is trying to sort from one another, is very important. They studied the performance of three ejection fraction variables relative to three intentional goals: diagnosis of coronary artery disease in patients without previous myocardial infarction, prediction of multivessel disease in the same patients, and prediction of multivessel disease in patients with previous myocardial infarction. They found that neither the information content nor the accuracy of radionuclide angiocardiography remained fixed when the test was employed for different, although related, purposes. Fletcher[23] found similar differences in the performance of the CEA test in screening for, and monitoring the progression of, cancer. Thus, diagnostic performance may differ depending on types of patients and institutions in which the test is used, and who the operators are.

It is important to be clear whether one wishes to compare directly two diagnostic modalities that are competitive, or to evaluate the incremental accuracy of tests that are complementary and might be used serially.

Although one would like to have a study evaluate the actual performance of diagnostic tests in the clinic, most observer studies are conducted after the fact on a retrospectively assembled set of cases; the reasons have mainly to do with issues of efficiency and standardization. The amount of reality to be built into "off-line" studies will depend on whether one is interested in measuring the pure signal detection performance of a technology, or its performance when used in conjunction with the other clinical information that is usually available in practice.

B. Study Observers

Although we often tend to ingore it, differences in accuracy can be greater among, or even within, observers than they are (on average) between modalities.[24] Since observers are both the main source of variation in performance and at the same time the main unit of analysis, the choice of observers needs to be carefully made, particularly in comparative studies. Swets and Pickett devote Chapter 10 to this topic.

Incidentally, the constant references in this review to "readers/observers" and "images" may give the impression that the only application of ROC analysis is to radiology and pathology, or to tests that require some operator interpretation. This is not so. The techniques are generic and apply equally, and more readily, to quantitative tests; the only difference is that some of the sources of human variation do not exist and will be replaced by other ones (such as within and between instruments).

C. Study Cases

1. Gold Standard for "Truth"

The true state of each case (i.e., D+ or D–) must be known in order to score the correctness of each diagnosis. Obtaining such "truth data" for ROC studies of observer performance in dealing with real cases is often difficult and investigators sometimes resort to using surrogate truth data. Revesz et al.[25] investigated various methods of approximating the truth on the conclusions of a study that compared three radiographic techniques. They found that any of the three techniques could be shown to be more accurate than the others, depending on which method was used to define the truth. A similar pitfall is to compare two tests using cases who are positive or suspicious on one of the tests being studied.

Henkelman et al.[26] have proposed a method of ROC analysis that does not require truth data, for use when several very accurate tests are being compared, but their suggestion has not been

followed up. The main drawback seems to be that by essentially relying on the other tests to "define" the truth, it is difficult for the new modality to appear better, even if it allows one to detect disease that the other modalities do not.

2. Case Difficulty

Kundel[27,28] has emphasized that for the cases used to compare two modalities, the decision on each case should be of borderline difficulty, since "using an obvious case will not show up any differences in the two modalities, nor will a completely obscure one." Thus, one should aim for a ROC curve that is somewhere near halfway between the diagonal and the upper left corner. In simpler detection tasks, the case difficulty can be manipulated by altering the amount of time allowed for viewing each case.[29]

3. Source of Cases: Verification Bias

It is incorrect to think of a ROC curve as a "constant" that is somehow characteristic of the diagnostic test per se. Several authors have pointed out that nature of the cases used to estimate the ROC curve can have a large influence on its location. To use Diamond's term, the ROC curve is far from "steady". The variation among the four reports of the performance of the stress electrocardiogram (ECG) test examined by Harris[18] illustrates this quite dramatically. Harris agreed that in tests such as ultrasound, lymphangiography, and mammography, there is wide latitude for observer interpretation and, thus, the possibility of wide variability in sensitivities and specificities as a function of differing observer decision criteria. However, he argued that in the stress ECG test, there is virtually no room for observer variability, at least in the studies he had presented, since "it was rigidly defined as 1 mm of flat or downsloping ST depression below the baseline, which lasted at least 0.08 s." Thus, "if one is ever to expect 'quite reliable data', the stress ECG, using a well-defined endpoint, the Bruce protocol, and a standard definition should provide such data." However, to his dismay, the data points from the four studies did not fall at a single data point in "ROC space", or even on a single unifying ROC curve, that one would expect from the standardization. Instead, Harris noted that the plotted points seemed to generate a "ROC ellipse".

Much of the variation between published ROC curves can be traced to "*biased populations*" (to quote one of Harris' own suspicions) and "problems of spectrum and *bias in the selection of test cases*" (an explanation offered in Swets' letter[19]). Ransohoff and Feinstein[30] describe how "workup bias" occurs when the result of the test being studied affects the subsequent workup needed to establish a definitive diagnosis. A positive test result may make the clinician look intensely for a disease that would otherwise go undetected, while a negative result can cause the diagnosis to be missed (Diamond[31] refers to these latter patients as "Reverend Bayes' Silent Majority"; based on the survey by Greenes and Begg,[32] one in every four articles on diagnostic accuracy that they surveyed had this bias). The biased sample produces a high index of sensitivity and an erroneously high accuracy for negative prediction. Begg and Greenes,[33] who refer to this as verification bias, use the model that if the selection for verification is predictable from the test result and other clinical indicants (so that the statistical association between selection and disease status is merely a reflection of the information in the test and other clinical indicants), then it is possible to quantitatively remove the selection bias in the indices of diagnostic performance. This can be done if information is available on the distribution of test results and clinical indicants in the source population. Diamond el al.[34] provide data on the use of exercise radionuclide ventriculography for detecting coronary disease to illustrate a model for assessing the sensitivity and specificity of tests subject to selection bias. Gray et al.[36] extended the methods of Begg and Greenes to the case of test results interpreted on a rating scale, and showed how one can derive a ROC curve which is corrected for verification bias and which should, therefore, more accurately reflect the performance of the diagnostic modality in the source population.

Diamond[36] has suggested that one does not need data on the distribution of test results in the source population in order to estimate a ROC curve that is free of verification bias. Instead, he argued that the "de-biasing" can be achieved simply by the choice of a particular form of the ROC, namely, that based on the "equal variance" bilogistic model (see below). Unfortunately, as was pointed out by Hanley and Begg,[37] this claim is not valid: the algebra used to claim the invariance of the 1-parameter logistic model to selection bias relied on a formula from Begg and Greenes, which is appropriate for individual 2×2 tables (i.e., for binary test results), but which becomes inappropriate when applied to the successive (and, thus, dependent) 2×2 tables generated when constructing ROC points from rating data.

4. Other Biases

Other biases in the assessment of diagnostic tests, such as test interpretation biases, uninterpretable test results, absence of a definitive reference test, and the effect of case mix, are described in several texts and articles.[5,38,39] Guidelines for circumventing the problems in prospective studies, and areas requiring further research, are described by Begg and McNeil.[40]

Fortunately, the considerable cost and task of setting up a retrospective observer study forces one to confront, and minimize, many of the biases. Also, by and large, biases are less of an issue in *comparative* studies, where the concern is with relative rather than absolute levels of diagnostic performance. The same issue is regularly faced in assessing studies of therapeutic performance, where there can be large variations in the performance of one treatment from center to center and subgroup to subgroup, but, presumably, the comparative performance of two treatments is more constant across these settings.

D. How Many Cases? How Many Readers?

1. General Principles

The third and fourth chapters of Swets and Pickett give a very valuable generic overview of the statistical analysis of any index derived from ROC curves. The authors emphasize that statistical precision depends not only on the number of cases (n) used in the study, but also on the size of the sample of observers (l) and the number of times (m) each case is interpreted in each modality. Including multiple observers is part of the general scientific requirement of "replicability" that enhances the believability and acceptability of findings, in the same way that the general tendency seen over many individual clinical trials or centers is more convincing than the results of a large, single-center trial. By studying and describing the range of performance across observers and sessions, one can avoid atypical results and produce a better estimate of the performance of the "average" or "typical" observer on an "average" session. Sample sizes are discussed further below.

2. Comparison of Two Modalities

When designing a study to compare the performance of two modalities, one tries as much as possible to keep all other factors constant in an effort to "compare like with like". Usually this means comparing the two modalities on the same set of patients, observed, when possible, by the same observers or, if necessary, by observers with the same degree of experience. As will be seen below, the analysis of the resulting data should take full advantage of this matching; this is done by using standard errors that reflect the fact that the difference in performance between two modalities should not be as variable in the same cases as it would be in different cases, or with matched observers as it would be with unmatched observers. The Swets and Pickett text has a detailed analysis of the increased power from using case and reader matching across the modalities being compared, as well as the gain from replication (reading each set of images in each modality more than once).

3. More Readings or More Readers?

The authors show that if the total number of readings is fixed, it is always more statistically

Table 2
RATING DATA FROM ONE OBSERVER IN ONE MODALITY[54]

	Test result (rating category)					
	−−	−	+/−	+	++	Total
D−	33	6	6	11	2	58
D+	3	2	2	11	33	51

efficient to add another reader than to have a reader read the cases a second time in the same modality. However, they emphasize that this logic is somewhat shortsighted, since it depends on the availability of an estimate of the variability when the same reader reads the same cases in the same modality more than once. If one does not have an estimate of this "within-reader" variability, one is penalized in the calculated standard error for not being able to estimate how much of the observed difference between two modalities could be due the fact that, even when the same cases are read twice by the same reader in the same modality, the reader does not achieve exactly the same accuracy.

E. Data Acquisition

The most economical and widely used method of collecting an observer's impression of each case is through the use of a rating scale,[4] i.e., using graded levels of confidence that the case is D+. Since it is not required that the different rating categories correspond to stated numerical intervals on the probability scale, or even that all readers use the ratings with same meanings,[41] labels such as "very likely", "probably", ...are sufficient. If a reader does not use all of the rating categories or uses some very sparingly, the resulting ROC curve will be based on fewer points, or points that are not well spread out across the axes. Swets and Pickett recommend a 2:1 ratio of D+ to D− to "coax readers away from a conservative criterion" and to encourage them to spread out the ROC points as much as possible over the rating scale. Judging from the poor spread of points on some of the ROC curves displayed in the literature, investigators have not been fully successful. A recent paper[42] reported that in each reading session, a random number of films were viewed more than once in order to change the prevalence of positive cases from one session to the next. Incidentally, in that study each session was devoted to a different modality; one would prefer to try to counterbalance the modalities over sessions in order to minimize any trends.

IV. DATA ANALYSIS

Following Swets and Pickett, the general procedure for comparative studies is to (1) form two-way frequency tables (such as the 2×5 table in Table 2), with one data table for each of the m readings of the entire set of cases by each of the l readers in each of the (say) two modalities, (2) fit a smooth ROC curve and derive associated accuracy indices for each data table to yield a multiway layout of $l \times m \times 2$ indices (indices are discussed in detail later), (3) average each reader's index over the m rereadings in each modality (if $m > 1$), (4) average the index over readers within each modality, and (5) compare the difference d in this average index between the two modalities to the standard error (SE) of d. Step (2) will be illustrated for one data table.

A. Empirical ROC Points

As is seen in Table 3, the coordinates for each empirical ROC point are the TP and FP fractions obtained from the 2×2 table formed by each different reexpression of Table 2. The two columns are formed by regarding all columns to the right of a designated cut point in Table 2 as a

Table 3
CONSTRUCTION OF EMPIRICAL ROC DATA POINTS (USING DATA FROM TABLE 2)

Rating data

	Test result (rating category)					
	−−	−	+/−	+	++	Total
D−	33	6	6	11	2	58
D+	3	2	2	11	33	51

ROC data points

	Test "result"		FP	TP
	"− − or − or +/− or +"	"++"		
#1				
D−	56	2	2/58	
D+	18	33		33/51
	"− − or − or +/−"	"+ or ++"		
#2				
D−	45	13	13/58	
D+	7	44		44/51
	"− − or − or"	"+/− or + or ++"		
#3				
D−	39	19	19/58	
D+	5	46		46/51
	"− −"	"− or +/− or + or ++"		
#4				
D−	33	19	25/58	
D+	3	48		48/51

"positive" test result and those to the left as a "negative" one. As one changes the cut point from the very strict one (where only the rightmost ++ category would be regarded as positive) to successively laxer ones, the sensitivity (TP fraction) increases, but obviously at the expense of the FP fraction, which also rises (and so specificity falls). If there are five rating categories, one can "cut" the axis in four ways, yielding four separate (although obviously not independent) two-dimensional data points.

Incidentally, many investigators treat each empirical sensitivity as a sample proportion and use the binomial distribution to accompany it by a vertical confidence interval (CI) around the estimated TP. There are two problems with this approach: first, confidence intervals for different sensitivities are not independent of each other; in fact, the true sensitivities must follow a monotonic sequence as the FP rate is increased, i.e., as the criterion for test positivity is relaxed. Second, this uncertainty calculation fails to take into account the sampling variability of the estimated FP point.[43] Some investigators attempt to incorporate this two-dimensional uncertainty by producing a second confidence interval, this time horizontal, for the FP point, so that the (TP, FP) point is enclosed in what appears to be a rectangular confidence region. This does not mean that the true TP for a given (theoretical) FP must fall within the rectangle or oval formed

by the "cross hairs" on the point; as was pointed out by Greenhouse et al.,[44] the SE used to form a CI for TP at a given FP is of the form

$$SE(TP_{FP}) = [TP(1 - TP)/n_+ + s^2\, FP(1 - FP)/n_-]^{1/2}$$

where s is the slope of the ROC at TP_{FP}. One will recognize this as a sum of the separate components: the first is the usual binomial variance for TP, based on n_+ D+ cases; the second, reflecting the horizontal uncertainty, is the binomial variance for FP, based on n_- D– cases, but weighted by the steepness of the ROC slope.

Even if there are many rating categories, or even if the data scale is continuous, the finite number of cases n will still cause the empirical ROC curve to have a jagged appearance. Thus, it makes sense to fit a smoothed curve; this curve estimates the true curve which would be produced if test data could be obtained on a continuous scale on an entire population of cases, rather than on just a sample.

B. Curve Fitting

Some have argued that the formal fitting of ROC curves, other than a simple eye-fit, is unnecessarily sophisticated to describe what is, after all, only a 2 × 5 data table. If the sole purpose is to give a very rough, almost qualitative statement and picture of whether the ROC points are close to the upper left corner, close to the diagonal, or halfway in-between, one would tend to agree. However, if a more quantitative description is required, as in a formal comparison of two modalities, some form of objective curve fitting and inferential procedure is necessary.

One author[77] has fit ROC curves using explicit equations, e.g., a mixture of power-law curves $TP = pFP^{1/a} + (1 - p)[1 - (1 - FP)^a]$. However, unless one uses jackknifing of cases (see below) it is not possible to describe the SEs of the resulting indices. Thus, the first, and still the only fully satisfactory, approach to fitting a smooth ROC curve to rating data is to posit the existence of two overlapping distributions (as we instinctively do in the case of numerical test results), one for the universe of D– cases and one (arbitrarily located more to the right) for D+ cases. Since we cannot visualize these scales, we cannot speak of the "true" shapes of the distributions. Any scale and any distributional form that produces a sensible ROC curve are suitable provided they reproduce the observed data table.

By positing the scale and the forms of these two overlapping distributions, one can convert the observed TP and FP proportions to points along this scale, and vice versa. (The choice of scale is not much more arbitrary than the choice one has to make, between, say linear and log, before plotting histograms for the concentrations of a biochemical substance.) For example, consider a scale where the two distributions are normal (Gaussian), with the nondiseased population having a mean of 0 and a standard deviation of 1, and the diseased population having a mean μ and a standard deviation σ. If the proportion of a normal distribution with mean 0 and standard deviation 1 that lies to the left of x is given by the function $\Phi(x)$, then a cutoff of x = 1.645 will yield a FP fraction of $1 - \Phi(1.645) = 0.05$ and a TP fraction of $1 - \Phi[(1.645 - \mu)/\sigma]$. The expected distribution of rating data generated by such a model is illustrated in Table 4; it is assumed for the purposes of illustration that the parameter values are $\mu = 2$, $\sigma = 1.5$, and that the four cutoffs, numbered from right to left (as in Table 3), are $x_1 = 1.645$, $x_2 = 1.28$, $x_3 = 0.84$, and $x_4 = 0.0$ (these four cutoff values were chosen so that they would be familiar to those who use cutoffs or critical values for statistical tests). These six parameter values generate ten entries (two sets of "five-nomial" frequencies) in the data table.

Table 5 shows the reverse process, i.e., how one fits the parameters by searching for the values that produce expected frequencies that "best" fit the observed data. The steps in the process are the same as those of Table 4, except that the order is reversed. The ten table entries, which, because each row total is fixed, constitute only 8 degrees of freedom, are used to estimate six parameters a, b, x_1, x_2, x_3 and x_4.

Table 4
EXAMPLE OF "EXPECTED" ROC DATA TABLE GENERATED BY SIX PARAMETERS OF A BINORMAL MODEL [n(D–) = 50, n(D+) = 60]

Parameters

μ(= ab)	σ(= 1/b)	x_4	x_3	x_2	x_1	
2	1.5	0.00	0.84	1.28	1.65	∞

Expected values

Performance

FP = 1 – Φ (x)	0.50	0.20	0.10	0.05	0.00
TP = 1 – Φ ([x – μ]/σ)	0.91	0.78	0.68	0.59	0.00

		Rating category					
		– –	–	+/–	+	++	**Total**
Proportions							
	D–	0.50	0.30	0.10	0.05	0.05	1.00
	D+	0.09	0.13	0.10	0.09	0.59	1.00
Data table							
	D–	25.0	15.0	5.0	2.5	2.5	50
	D+	5.4	7.8	6.0	5.4	35.4	60

Note: Φ (x) is the proportion of the normal distribution with mean 0 and standard deviation 1 falling below (to the left of) x. Thus, Φ ([x – μ]/σ) is the proportion of the normal distribution with mean μ and standard deviation σ falling below (to the left of) x. The expression [x – μ]/σ is written as (1/σ) x – μ/σ = bx – a in Dorfman and Alf.

At least two criteria for best-fit could be used, that of minimum chi-square[45] and that of maximum likelihood.[46] Both methods estimate all six parameters simultaneously and require an iterative search. It is important to realize that both techniques fit the proportions or expected counts to the observed 2 × 5 data table, rather than to the points per se, and both recognize the multinomial structure of the data.

While the performance data appear to have the traditional linear form "y = bx – a" when plotted on normal-deviate axes, the x's are *parameters* to be estimated rather than known constants. Also, for rating data, the data points were not independent, since the cumulation over successive rating categories induces a correlation. The nuisance parameters x_1 to x_4 are not needed to draw the smooth or fitted ROC curve; it is simply the locus of points $1 - \Phi[(x - \mu)/\sigma]$, i.e., $1 - \Phi(bx - a)$, vs. $1 - \Phi(x)$ as x moves from ∞ to -∞, and so it only needs the values of a and b.

As will be discussed below, one could just as profitably have chosen logistic rather than normal distributions. Although their mathematical form is less familiar, the two models are very close in shape, and, indeed, for the former it is mathematically easier to transfer between the FP and TP fractions on the (0,1) scale and the modeled decision (x) scale, since the logistic function corresponding to Φ(x) is simply $\Phi(x) = 1/[1 + \exp(-x)]$. ROC curves fit from a logistic model will be linear on logistic axes.

C. ROC Axes: Linear or Transformed?

Some authors prefer to plot the empirical and the formally fitted ROCs on these transformed

Table 5
SIX BINORMAL PARAMETERS ESTIMATED FROM OBSERVED ROC DATA TABLE (DATA AS IN TABLE 2)

Observed

		Rating category					
		−−	−	+/−	+	++	**Total**
Data table							
	D−	33	6	6	11	2	58
	D+	3	2	2	11	33	51
Proportions							
	D−	0.57	0.10	0.11	0.19	0.03	1.00
	D+	0.06	0.04	0.04	0.21	0.65	1.00
Performance							
(Linear axes)							
	FP	0.43	0.33	0.22	0.03	0.00	
	TP	0.94	0.90	0.86	0.65	0.00	
(Binormal axes)							
$x = \Phi^{-1}(1 - FP)$		0.20	0.47	0.77	1.88	∞	
$[x - \mu]/\sigma = \Phi^{-1}(1 - TP)$		−1.65	−1.34	−1.13	−0.38	∞	

Parameter estimates

	$\mu(= \mathbf{ab})$	$\sigma(= \mathbf{1/b})$	x_4	x_3	x_2	x_1	
Initial[a]	1.16	1.42	0.17	0.45	0.76	1.82	∞
Final[b]	1.18	1.40	0.17	0.46	0.77	1.80	∞

Fitted values

$FP = 1 - \Phi(x)$	0.43	0.32	0.22	0.04	0.00	
$TP = 1 - \Phi([x - \mu]/\sigma)$	0.94	0.91	0.87	0.65	0.00	

		Rating category					
		−−	−	+/−	+	++	**Total**
Proportions							
	D−	0.57	0.11	0.10	0.18	0.04	1.00
	D+	0.06	0.03	0.04	0.22	0.65	1.00
Data table							
	D−	33.1	6.4	5.8	10.4	2.3	58
	D+	3.1	1.5	2.0	11.2	33.2	51

[a] Estimated by least squares method from data on binormal axes.
[b] Estimated by method of maximum likelihood.

(normal, logistic, ...) axes rather than on the conventional linear (0,1) ones. The main reasons are that it is easier for the human eye to compare straight lines than curves, and that the lack of fit of the fitted model is more readily apparent on the transformed axes. Also, these axes make it easier to plot several curves, since the scales expand as one approaches the extremes, i.e., the crowding at the upper left corner of the unit square is avoided. Of course, both scales can be shown on the same graph.

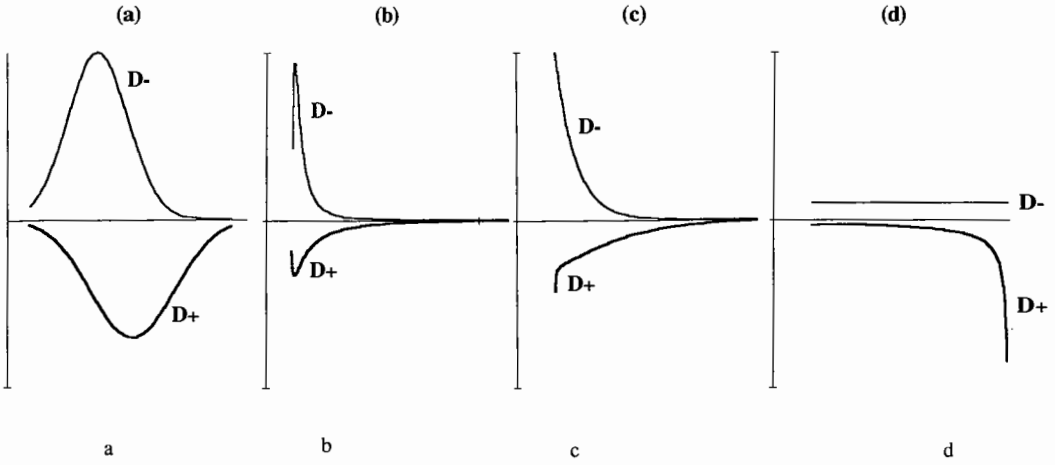

FIGURE 2. Four representations of the same overlapping pair of distributions. The "D–" frequency distributions are shown in lighter lines above the x-axis; the "D+" frequency distributions are shown in heavier lines below the x-axis. (a) Binormal distributions using the parameter values $\mu_{D+} - \mu_{D-} = 1.18$, and $\sigma_{D+}/\sigma_{D-} = 1.18$ (the same as those derived from Table 5); (b to d) the same distributions, but with the x-axis rescaled in three different ways. They show that the "binormal" assumption is not a strong one for rating data.

D. Choice of Models for Fitting ROC Curves

One might ask whether the (somewhat arbitrary) choice of the binormal model is justified, especially when the data scale posited to underlie the observed 2×5 table is unobservable. A review of the history of some of the available models and how these models are really used may help.

Swets and, especially, Metz have repeatedly emphasized that whether one fits a binormal model (or a bilogistic, bichi-squared, or biexponential one), one does not assume that if one could observe the latent distributions, they would have this exact form. The ratings are ordinal rankings, and the spacings between them do not imply any particular numerical scale. Investigators sometimes try to attach a probabilistic value to the verbal label associated with each rating category, but even if these were taken literally (as in weather forecasting), the distributions can be made to look very different by simply changing from a probability to other scales, such as the odds, log odds, probit, or logit.

This is illustrated in Figure 2, where by simple changes in the decision variable scale, the familiar pair of fitted normal distributions can be deformed into very different shapes, one like a pair of exponential distributions, one more like a pair of chi-square distributions, and one pair involving a rectangular distribution. Conversely, had one begun by fitting a pair of negative exponential distributions, one could easily, by the same changes of scale, produce a pair of distributions very close to the binormal ones. No matter which scale, however, the ten fitted data frequencies would be the same and the ROC curves would be identical.

This latitude in the depiction of the same distributions means that a ROC curve does not uniquely determine the underlying decision-variable distributions. Any monotonic transformation of the decision variable generally yields different underlying distributions with different shapes, but does not change the functional form of the ROC curve. Thus, the "binormal" assumption concerns only the functional form of the ROC curve (which can always be examined empirically) and not the form of the underlying distributions themselves (which cannot be determined in many applications of ROC analysis).[3] It only specifies the underlying distributions up to a monotonic transformation of the decision axis.[47,48] As one can see from the figure, the coarse nature of the rating data leaves a lot of freedom to fit distributions to tables. This freedom is not a function of sample sizes (numbers of cases), but of the number of rating categories.[49]

Judging from Figure 2, it is no surprise that the binormal model fits rating data so well. Swets[50] recently reviewed a sample of empirical ROC curves from both discrimination tasks in experimental psychology and from diagnostic tasks in several practical fields. He found that they all seem to be well fitted by a straight line, of varying slope, on a binormal graph. Hanley[49] recently compiled all of the arguments, both theoretical and pragmatic, given for the binormal model in the various signal detection texts. These ranged from theoretical considerations of probability laws and signal detection theory, to mathematical tractability and convenience, to empirical results showing that "it fits!". He supplemented them with a numerical study of how well a binormal ROC fitted datasets that had been generated by other models. He found that even if an alternative formulation based on another underlying form (e.g., power law) or model (e.g., binomial, Poisson, or gamma-type distributions) were, in fact, correct, the binormal fit differs so little from the true form as to be of no practical consequence. Moreover, the small lack of fit is unlikely to be demonstrated in practice: it is obscured by the much larger variation that can be attributed to sampling of cases. For these reasons, the choice of the normal or logistic model can be on pragmatic grounds. Ogilvie and Creelman chose to fit the logistic model;[51] Dorfman and Alf chose the normal.[46] Both used the method of maximum likelihood to fit them. The availability of computer software to implement the binormal model probably explains why it became the more popular of the two.

E. Fitting an ROC Curve to Data from Different Studies

Occasionally, one may wish to construct a ROC curve for a diagnostic modality using single (TP, FP) points from different reports in the literature. Swets did just this, albeit in an informal way, using a simple eye fit, with the data presented in Harris' compilation of (TP, FP) pairs from different studies of a single modality. One can produce a more formal fit using a fitting technique developed to estimate a ROC curve from a series of (independent) 2×2 tables.[52] The technique assumes that the different 2×2 tables in the series are generated from separate "sessions" using a *binary* response category; a different criterion level for positivity (signal presence) is used in each session. More often than not, since its use was mainly in psychophysics, the set of cases varied from session to session (unless they had been recorded). However, the nature of the stimuli and, thus, the level of detectability of the "signals" remained the same over sessions. Because it requires one session per data point, this technique is seldom used nowadays.

However, if one considers the 2×2 data table from one medical diagnostic study as analogous to one 2×2 data table from one session of the psychophysics study, and if the nature (degree of difficulty) of the cases was reasonably constant over the different reports, the algorithm can be applied to fit a curve to a series of data points from the literature. The procedure uses the same binormal model used to fit rating method data, and produces estimates of the same parameters a and b.

Obviously, there is the possibility that the points from different investigations do not really belong on the same ROC curve, just as there is often controversy over whether discrepant results of several studies of the performance of a therapeutic maneuver should not be merged into a single overall summary. The output from the fitting procedure can help one to judge the contribution of each data point to the overall lack of fit statistic. However, just like in any pooling of therapeutic results from different studies, the decisions as to what results to pool involve more than just purely statistical criteria.

It must be emphasized, however, that it is very difficult to fairly compare the "meta-ROC" curve for one modality (using one set of reports) with that for another modality (estimated from other reports). Any attempt to do so is subject to the same kind of criticism that would be engendered if one compared the pooled reports on the performance of one therapeutic maneuver with the pooled results of (another series of) reports on the performance of another therapeutic maneuver.

F. Pooling Rating Data from Several Readers

When using a rating scale, readers are very likely to differ in their choices of decision criteria; for example, one observer's "+/–" may be another's "–". Thus, it is unwise to pool the data tables of individual observers, unless their decision criteria are very close. The single index calculated from the overall pooled raw data is usually attenuated compared to the average of the reader-specific indices, as is explained in Chapter 2 of Swets and Pickett. More recently, Macmillan and Kaplan[53] have analyzed ways to obtain composite ROCs from binary-response data. Their results have been taken as support for the pooling of rating data, although they caution against "averaging together data from subjects whose (accuracies) or response criteria are very disparate." As will be discussed later in this review, these cautions are not always heeded, even though ignoring them can produce unwelcomed effects. Raw data should only be pooled in situations where they are too sparse to be analyzed separately, and then only after careful inspection.

G. Indices Derived from ROC Curves

1. Reproducing the ROC Curve

The values of the binormal (or bilogistic) parameters a and b are sufficient to reproduce the smooth ROC curve. For example, to calculate the TP points corresponding to FP = 0.05 to 0.95, one first uses either the normal probability tables or the function $\Phi^{-1}(p)$ (which gives the x value below which a proportion p of the normal distribution lies) to calculate $x_{.05} = \Phi^{-1}[1.00 - (FP = 0.05)]$ to $x_{.95} = \Phi^{-1}[1.00 - (FP = 0.95)]$. One then uses the complementary normal probability tables or the complementary function $\Phi(z)$ (which gives the proportion of the normal distribution that lies below z), to obtain the TP fractions

$$TP_{(FP = 0.05)} = 1 - \Phi(bx_{05} - a) \text{ to } TP_{(FP = 0.95)} = 1 - \Phi(bx_{95} - a)$$

2. Indices to Summarize a ROC Curve

As an index of accuracy, one can use either the TP fraction corresponding to a single selected FP fraction, which one might refer to as TP_{FP}, or the area under the ROC curve, which is generally referred to as A.

The main advantage of the TP_{FP} index is that it is readily understood. However, it is unlikely that different authors will standardize their TPs to the same value of FP, and one cannot always gather from a published report whether a FP value was chosen in advance of a study or after inspection of the curve(s). Moreover, as will be discussed below, the statistical reliability may be lower than that of other univariate indices.

Thus, the area A (and, in particular, the area A_z) under the ROC curve plotted on ordinary axes is recommended as the index of choice.[4] The "z" in A_z is used to denote that it is calculated using the formula $A_z = \Phi[a/(1 + b^2)^{1/2}]$ from the parameters a and b of the ROC curve fitted using the binormal model; we would need a different suffix (say *l*) and a different formula if the parameters are fit using the *l*ogistic model. The suffices z and *l* serve to distinguish these model-based areas from the trapezoidal area under the "curve" that has been formed simply by joining the empirical ROC points. Although this latter area, called P(A) by Green and Swets, is nonparametric and much easier to calculate,[54] it is more likely to be affected by the location or spread (or small number) of the ROC points and yields a smaller area than one derived from a smooth curve.[4] Few would advocate using P(A) as a summary index for already collected ROC data. However, because it is nonparametric, and because it is related to the well-studied Wilcoxon statistic, its sampling variation can be used to project the variation to be expected in P(A), or in A_z, from one set of cases to another (assuming no within-reader or between-reader variation).[54]

Apart from its obvious graphical meaning, the A index does have an interpretation in signal detection theory as the proportion of correct choices in a two-alternative forced choice (2A-FC) experiment,[1] i.e., an experiment where in each trial the reader is presented with a pair of images,

one from a randomly chosen D+ and one from a randomly chosen D–, and is asked to decide which image derives from which. Bamber[55] and later Hanley and McNeil[54] exploited this three-way identity between A, the quantity Prob(Y > X) measured in the 2A-FC experiment, and the proportion estimated by the Wilcoxon statistic to characterize the variability of A due to case sampling.

Despite its interpretation as a probability of correct choice in the 2A-FC experiment, the area index does not appeal to all investigators. Some find it too "slick" and too "mathematical"; they argue that since a large part of the area comes from the rightmost part of the curve, it includes FPs that are of no clinical relevance, and so is likely to be insensitive when used to compare the performance of two modalities. They also worry that curve 1 may have higher TPs than curve 2 in the region of relevant FPs, but that the two curves could conceivably cross, and that this superiority in the left portion of the curve may be lost, or even reversed, when the curves are ranked on the basis of the overall curve.

3. Why A_z?

One set of negative reactions to A_z is worth recounting. One investigator[56] conceded that the (ROC) method to describe the diagnostic efficiency of a test used in the paper by Swets et al.[10] is an elegant way to graph the sensitivity and the specificity of a diagnostic test where observers are not or cannot be standardized to diagnostic criteria. However, he worried that, unfortunately, the "fundamental index", A_z, is less useful for comparing diagnostic methods than is visual inspection of the ROC curves. He further suggests that, even when the sensitivity-specificity lines, plotted on normal-deviate axes, do not cross, comparison of A_z index is meaningless unless there is a good likelihood that the sensitivity-specificity lines are actually parallel. He argues that "even when the use of the A_z index is statistically permissible, it is not the logical measure of comparison when choosing a method for its cost-benefit in screening, for its precision in estimating prevalence, for its sensitivity in monitoring change, or any other characteristic" and asks "is there then, any use for the A_z index in comparing medical diagnostic methods?"

Since this reviewer suspects that many investigators have this same reaction to the area index, but may be unwilling to voice it publicly, the authors' reply[57] is particularly instructive. Not only does it focus on which ROC index to use, but it also serves to put ROC methodology itself into broader perspective. Thus, it is reported and quoted at some length here.

While agreeing that in any one particular situation one should try to define the operating point on the curve that is optimal for that situation and use this point in further analyses, they stress

> We do not agree, however, that all important questions of accuracy pertain to a particular situation, nor that optima can usually be clearly established, and so we find useful an index (A_z) that reflects accuracy in general, through a range of possible operating points.
>
> A general index of accuracy is required when evaluating a diagnostic method from a general vantage point, say, a government health agency. The agency must recognize that the optimum operating point for a given disease will vary considerably from one diagnostic setting to another (for example, community hospital to teaching hospital) and, within one setting, from one patient population to another (for example, from a population being screened to a population at high risk). The agency no doubt will also recognize that optimum operating points for particular situations are frequently not precisely determined, and that the operating point used in a given type of diagnostic situation will vary across locations and perhaps across diagnosticians at one location. Moreover, translating an established optimum into consistent practice is not simple. Thus an index of accuracy that represents a range of operating points is desirable, if not necessary. A_z has convenient and well-studied statistical properties, which make it preferable to the several other general indices that have been considered.

Swets et al. do not dismiss Habicts's concern at the possibility that the slopes of ROC curves being compared may differ; they emphasize that a general index should not be used if the slopes are materially different. However, they point out that the difference in their particular study could easily have been sampling variation, and that in any case the curves did not cross anywhere near an operating point of possible interest. In their experience, ROC slopes are essentially similar in most comparisons.

4. A Compromise between TP_{FP} and A

It has been suggested that, rather than choosing either the index TP_{FP} (the TP for a selected FP point) or, at the other extreme, the index A (which can be thought of as TP averaged across all FPs), one should compute the average sensitivity (TP) over a range of relevant FPs. Part of the statistical theory for this type of index, which one could term $TP_{FPrange}$, has recently been developed.[58] The statistical procedures are established in the context of a class of indices and are illustrated using continuous rather than rating data; however, they can be adapted to rating method data. In its generic form, all of the indices are viewed as weighted integrals of the sensitivities (s) over the specificities (p), i.e.,

$$\text{Index} = \int S(p)dw(p)$$

where the integral runs from 0 to 1, and w(p) is any nondecreasing function satisfying w(0) = 0 and w(1) = 1. Thus, the sensitivity for a given specificity p_0, which the authors call $S(p_0)$ and which is called TP_{FP} in this review, is calculated by letting w(p) have unit point mass at the fixed value p_0. The area index A is obtained by letting dw(p) = dp for all p, while the average sensitivity for specificities between p_1 and p_2, which is called $TP_{FPrange}$ in this review, is given by letting w(p) = $dp/(p_2 - p_1)$ for p between p_1 and p_2, and 0 elsewhere. SEs are developed for the various forms of the index and for between-modality differences in the index.

As they stand, the methods are not entirely suited to the heavy grouping inherent in rating data. However, the simulations that were carried out on continuous data suggest that in terms of statistical power to distinguish between two diagnostic modalities, the A index had the highest power, the $TP_{FPrange}$ was intermediate, and the TP_{FP} index was the least sensitive. This ranking is to be expected, given the successively smaller amounts of data used in the three indices. It is also borne out in the examples supplied with the computer program ROCPWR[59] which calculates statistical power for a variety of accuracy indices derived from rating data.

H. Standard Errors Used for Inferential Procedures

1. Single Modality

The main purpose in reporting an index of performance of a diagnostic system with a sample of cases, a sample of observers, and a sample of readings is not to tell the journal readership how well it performed in this particular sample that *was* studied, but to provide an estimate of how it would perform "on the average" in those similar cases and observers and readings that were *not* studied. The uncertainty of this estimate of performance, usually quantified using a SE, should be an amalgam of all of the sources of its sampling variation. (Nonrandom perturbations [i.e., those of a systematic nature] are referred to as "biases"; these are often more difficult to quantify.)

In a study of n cases, *l* observers or "readers", and m readings/case/observer, the SE of the index (averaged over m and *l*) is given by Equation 2 of Swets and Pickett,

$$SE(\text{index}) = [V_c/n + V_{br}/l + V_{wr}/lm]^{1/2}$$

where V_c, V_{br}, and V_{wr} are the variance components introduced by sampling individual cases, observers, and reading occasions, respectively. This textbook should be consulted for a fuller explanation and illustration of these components. As a general rule, however, the index is more trustworthy, i.e., the standard error is smaller, if it is averaged over more cases (n), more observers (*l*), and more occasions (m). In the form of the equation given here, I have used V instead of S^2 for variance and changed the meaning of "V_c" so that it refers to the variance in the difficulty of *individual* cases, rather than to the "variation due to differences in *mean* difficulty from one case sample of *n* to another case sample of *n*" (Swets and Picketts' usage). Because the index is derived from a *group* of n cases, it contributes a term V_c/n to the SE equation. I have

presented "V_c" with this altered meaning in order to emphasize its similarity to the $(\sigma^2/n)^{1/2}$ used as the SE of a simple group mean. This multicomponent form of the SE is not peculiar to indices derived from ROC curves: if one estimated the average blood pressure of some population from a sample of n individuals, each measured *l*m times by *l* observers on m different instruments, the associated SE would have a somewhat similar form.

The quantity V_{br} must be estimated through a components of variance analysis, since the *observed* variation between readers includes both V_{br} and V_{wr}. Likewise, the quantity V_c must be estimated by subtracting V_{wr} from the variance V_{c+wr} that is calculated by the Dorfman and Alf procedure. In a study which does not include repeat readings, the two components cannot be separated and the estimated SE will be too large by some unknown amount.

2. Comparison of Two Modalities

Statistical tests and confidence intervals for the difference d in accuracy of two modalities is based on the SE of this difference (Swets and Pickett advocated the use of the z tables for such inferences, although, as will be discussed below, it will sometimes be more appropriate to use the *t* tables). If the two modalities are tested on two completely different sets of cases by different readers, the SE of the difference in the mean performance index of two modalities is simply the square root of the sum of the squares of the two separate SEs. If cases and/or readers are matched across the two modalities, the relevant components of the SE are reduced accordingly. This is done by using correlation terms (r's) which multiply the variance terms by factors of $(1 - r)$. Again, following Swets and Pickett, if the V's, n's, *l*'s, and m's are equal in the two modalities, the SE of the estimated difference d in the performance index between two modalities has the compact form

$$SE(d) = [2\{(1 - r_c)V_c/n + (1 - r_{br})V_{br}/l + V_{wr}/lm\}]^{1/2}$$

where r_c and r_{br} are correlations reflecting the degree to which a set of cases which are above (below) average difficulty in one modality will be likewise in the other modality, and the degree to which a set of observers who are above (below) average accuracy in one modality will be likewise in the other modality.

Methods for estimating the different variance and covariance components, and for obtaining conservative estimates for V_{wr} (if the study does not include repeat readings) are described in detail in Swets and Pickett. Their method of estimating V_c/n and r_c can be difficult when n is small, because it involves splitting the data set into subsets of cases, and applying the Dorfman and Alf estimation method to each subset; Hanley and McNeil[60] used a largely non-parametric approach to estimating these two quantities for the "area under the curve" index. They also illustrated the use of jackknifing of cases, coupled with the Dorfman and Alf procedure, to directly estimate the SE of any index difference measured on the same cases.[61] Metz et al.[62] provide a bivariate extension of the binormal model which allows one to estimate r_c for any accuracy index derived from the same set of cases.

I. Sample Size Considerations

If one wishes to have a power of $(1 - \beta)$ to detect a difference (Δ) of δ in the accuracy of two modalities, using a "critical ratio" test based on $z = d/SE(d)$ with a type I error of α, then the SE of the observed difference (d) must be small enough to satisfy the equation:

$$CR_\alpha\, SE(d \mid \Delta = 0) + CR_\beta\, SE(d \mid \Delta = \delta) \leq \delta$$

(The quantities CR_α and CR_β [with CR_β taken as a positive quantity] are the values of the critical ratio associated with probability levels of β and α; they should be based on the z or t distributions depending on how many degrees of freedom are available to calculate SE[d].)

Because SE(d) involves three sources of variation (Vs) and three sample sizes (n, *l*, and m), there are many possible sample size configurations. Estimates of the Vs can be obtained from previous studies; indeed, investigators are urged to include them in their reports. Some sense of the tradeoffs among n and *l* can be gained from Figure 24 of Swets and Pickett, which uses the Vs estimated from their own study to present the relative power of different designs using n's from 60 to 240 and *l*'s from 1 to 7. A computer program, ROCPWR, developed by Metz[59] can simulate the power of studies of various size n's (cases) to detect differences between modalities in single- and two-parameter indices of ROC curves. Likewise, the procedure of Hanley and McNeil can be used for the area index. However, the sample sizes determined by these two approaches may be unrealistically small in that the calculations only take account of the term $(1 - r_c)V_c/n$, i.e., they, in effect, assume that $(1 - r_{br})V_{br}$ and V_{wr} are zero. It would be helpful if more data were made available on the three components so that those planning studies can assess which of the three variance components contribute the most to the overall SE(d).

J. A Simpler Approach?

Recently, Dorfman and Berbaum described a method and supplied an accompanying computer program to compute jackknife estimates (and their SEs) of indices extracted from a single ROC curve fitted to rating-method data that have been pooled over a group of observers[63] (such pooling might be the last resort if there are not sufficient cases to fit separate ROC curves for each observer). The approach is more noteworthy for the implications it has for unpooled than for pooled data, and is considered at some length here.

Usually, jackknifing has been used to assess sampling variation due to cases.[60,64] However, Dorfman and Berbaum use it to assess the variation due to readers. In order to understand their jackknife approach, it is best if one first reviews the four steps one might take if there *were* enough data to estimate a separate curve for each of the l observers. These are

1. Calculate a separate value of the index, which we could denote Index_1 to Index_l, for each of the observers.
2. Calculate and report the average value, $\text{Index}_{\text{average}}$, of this index.
3. Assume that V_c/n and V_{wr}/lm are negligible and base the SE of the reported $\text{Index}_{\text{average}}$ on the number (l) of readers making up this average and the observed variation (S^2_{br}) in the index between readers, i.e.,

$$\text{SE}(\text{Index}_{\text{average}}) = [S^2_{br}/l]^{1/2}$$

where S^2_{br} is computed from the indices Index_1 to Index_l in the usual way as

$$S^2_{br} = [\Sigma(\text{Index}_i - \text{Index}_{\text{average}})^2]/(l - 1)$$

4. Base inferences on the *t* distribution.

In their jackknife method, one proceeds by these same four steps, except that, instead of Index_1 to Index_l, one uses "pseudovalues", called Index_{*1} to Index_{*l}, of them. These pseudovalues can be regarded as the contributions of the individual observers to the single ROC curve estimated from the pooled data. Thus, to obtain the pseudovalues, one must first calculate the index, which they denote $\text{Index}_{\text{all}}$, from the entire pooled rating data. To obtain the pseudovalue for observer i, one deletes the rating data produced by the ith observer from the overall pool of data, and uses the index, which they denote Index_{-i}, calculated from this reduced pool, to calculate

$$\text{Index}_{*i} = (l)\text{Index}_{\text{all}} - (l - 1)\text{Index}_{-i}$$

Table 6
COMPARISON OF RESULTS OBTAINED FROM INDIVIDUAL AND POOLED RATING DATA[a]

Observer	A_z	A_{z*}
1	0.809	0.779
2	0.867	0.837
3	0.840	0.829
4	0.852	0.726
1—4 pooled	0.804	
Average of 1—4	0.842	0.793 ← jackknife
SD of 1—4	0.024	0.051
SE (average)	0.012	0.025 ← SE (jackknife)

Note: A_{z*} denotes the pseudovalue obtained by jackknifing.

[a] Data source Dorfman and Berbaum.[63]

The analysis of these pseudovalues then proceeds in the same way as outlined in steps 1 to 4 above.

The dangers of unnecessary pooling: an aside — Permit a comment on pooling per se. As was mentioned earlier, this reviewer sees little need for the pooling of rating data from individuals or for formal inferences based on them, unless cases are very sparse. The response data used to test the implementation of the RSCORE-J program[63] are a good illustration of the distortion that can occur when one pools raw data. These are the the same data described in the original Dorfman and Alf paper[46] and again in Appendix D of Swets and Pickett: n = 1188 and $l = 4$. The values of the A_z index, calculated separately for each observer, are shown in the "A_z" column of Table 6. They range from 0.809 to 0.867 and average 0.842. However, the single A_z estimated from the $nl = 4752$ ratings is 0.804, which is 0.038 lower than even the lowest of the four individual A_z's. The reason for this large attenuation is that the four datasets that were pooled had very different decision criteria. This can be seen in the binormal plots (see Figure 1 of Dorfman and Alf, or the printouts shown on pages 216 to 219 of Swets and Pickett). Observer 4 used much more conservative decision criteria than the others, so much so that his rightmost criterion value is near the leftmost criterion value of observer 3. Under the jackknife approach, one would use as a summary index the equally attenuated value of $\text{Index}_{*\text{average}*}$ (0.793) rather than $\text{Index}_{\text{all}}$. One could dismiss this attenuation as inconsequential as long as one compares attenuated estimates of the index for one modality with attenuated indices of another. However, one has to be more worried about the standard deviation of 0.051 and the SE of 0.025 ($0.051/\sqrt{4}$) calculated from the 4 "between observer" pseudovalues. The standard deviation and associated SE obtained by the jackknife method are more than twice those obtained from the individual A_z's. The reason that the jackknife method overestimates the between-reader variation can be traced to the pooling of data with very different decision criteria. As a result, as one can see in the A_{z*} column of Table 6, the pseudovalue for observer 4 ranks far below that of the others. The result is that while observer 4 would rank as the most accurate in the individual A_z's, he appears least accurate when ranked using the pseudovalues. In this example, the distortion resulting from the pooling of data has falsely increased the estimate of between-reader variation by a factor of 2, and this, in turn, would lead to very conservative statistical tests. If observer 4 operated in the same way in a second modality, one might speculate that the distortion would be partially cancelled out. However, the distortion would have a very serious impact in an "unmatched readers" study.

1. Which SE? An Important Principle

It is instructive to look beyond any objections to the pooling of raw rating data when one has enough data that one doesn't need to pool, or any distortions that pooling can produce when used with jackknifing. Consider instead the *principle* that Dorfman and Berbaum use to calculate SEs; the principle is the same whether one uses the traditional Index_i's or the pseudovalue Index_{*i}'s. The important point to note is that, in contrast to Swets and Picketts' formula, the case variance V_c and the rereading variance V_{wr} do not appear in the form of SE implied by the Dorfman-Berbaum approach. By analogy with SE calculations for more conventional measurement data, one suspects that V_{wr} is already contained in the *observed* between reader variance S^2_{br}. Whether

V_c should be included is less of a worry if one is estimating the SE of the *difference* in an index between two modalities which use the same set of cases. This is especially so if the case-correlation r_c is high, since the contribution $(1 - r_c)V_c$ to the overall SE will be greatly reduced. Dorfman and Berbaum explicitly acknowledge the issue of reader matching by stating in their user's guide, "if two independent pools of observers are to be compared, jackknife estimates can be used to construct a *t* test for independent samples." This means that the difference d in the index between modalities 1 (evaluated by l_1 observers) and 2 (evaluated by l_2 unmatched observers) could be tested statistically by computing the usual critical ratio

$$CR = \frac{Index_{average}(1) - Index_{average}(2)}{\{SE^2[Index_{average}(1)] + SE^2[Index_{average}(2)]\}^{1/2}}$$

and comparing it to the Student *t* distribution with $l_1 + l_2$ degrees of freedom.

For the study with matched readers, they recommend: "If two pools of data are obtained from a single group of observers presented with the same stimuli under two experimental conditions, a *t* test for paired observations can be performed upon the paired pseudovalues of the two groups." This means that the average of the *l* paired differences $d_i = Index_i(1) - Index_i(2)$ could be tested statistically by a one-sample Student *t* (with $l - 1$ degrees of freedom) by computing the usual critical ratio

$$CR = \frac{\text{average d}}{\text{SE[average d]}}$$

where the SE is calculated from the individual d_i's in the usual way.

It is not clear to this reviewer whether the term "the same stimuli" refers to case matching (as in medical images) or to "unmatched stimuli, but of the same type" (i.e., as might be presented in some "live" psychophysics experiments). Further investigation and clarification are needed to be sure that case and rereading variances are, or should be, included in Dorfman and Berbaum's formulas for the SEs. If they need to be included, the principle of jackknifing can be extended to cover both cases and readers, as has been suggested by Hanley.[65]

The Dorfman-Berbaum approach does emphasize, in a statistical way, the important point that the strength of an observed difference between two modalities depends on the number of observers in whom the difference is observed and (maybe) less so on the number of cases used in the study. In other words, the unit of analysis is as much the observer, as it is the objects that are being observed. Indeed, this is implicit in the terminology used by Dorfman and Berbaum, who use the term "subject" to refer to an observer.

The simpler Dorfman and Berbaum approach is not to be taken as a license to ignore the larger SEs calculated by Swets and Pickett's formulation. The two approaches will agree if $(1 - r_c)V_c/n$ and V_{wr}/lm are negligible compared to the term $(1 - r_{br})V_{br}/l$. Some will recognize the square root of this dominant term as the denominator (in disguise) of a paired *t* test involving the $2l$ indices of *l* matched readers.

2. Readers as the Unit of Analysis?

If readers are the main contributors to the SE(d), and a *t* test or paired *t* test is, therefore, warranted, this means that the more complex ML estimation procedures required to generate n-based SEs are no longer as crucial. As was pointed out by Grey and Morgan,[45] simple iterative techniques, suitable for implementation on a spreadsheet, can be used to estimate the parameters a and b of the logistic ROC model using the method of minimum chi-square; or one could use nonlinear estimation routines from standard statistical packages to fit either the binormal or the

bilogistic ROC by maximum likelihood. The fact that such fitting technique may not yield a case-based SE does not matter if SE(d) is to be based only on the variation among, and the number of, observers.

3. Implications for Sample Sizes

It is unfortunate if the early preoccupation with SEs and with statistical inference, in general, beginning with Dorfman and Alf and continuing with Hanley and McNeil and with Metz, was focused on n rather than *l*. These "n-based" SEs do have a place, albeit in special situations. They are appropriately used, when $l = 1$, to statistically compare the performance of a *specific* (named) reader in one modality with the performance of the same (or another named) reader in a second (or perhaps in the same) modality. However, no matter how large the n, an *l* of 1 cannot be used to make inferences to a whole class of readers (unless, of course, there is absolutely no between-reader variation). "Operatorless" diagnostic systems that invariably give the same test results on a case are one situation where the inferences can be based solely on n-based statistics. Examples of such systems are automated computer procedures which use objective features of images to detect abnormalities; and clinical prediction techniques such as discriminant analysis, logistic regression, and other patient-sorting algorithms that use unequivocal clinical indicants to generate diagnoses or prognoses. In such situations, where V_{br} and V_{wr} are both zero, the statistical conclusions are determined solely by the case-variance V_c, the degree of case matching (r_c), and the size (n) of the case sample. The n-based SEs from the estimation procedures of Dorfman and Alf, Hanley and McNeil, and Metz and the formulas and programs for calculating sample size developed by Hanley and McNeil, and Metz are directly applicable in such situations.

Dorfman and Berbaum's approach also makes another important point, namely that when using an estimate of variance based on $l - 1$ degrees of freedom, the critical ratio (CR) which uses this variance estimate in its SE is more correctly referred to the Student *t* distribution than to the normal (z) tables. This recommendation will not be welcomed by investigators, since the CR required for statistical significance is much higher for *t* tests than *z* tests, if the number of degrees of freedom is small (contrast the CR of 1.96 required for a two-sided *p* value of 0.05 based on normal tables, with the CR of 4.3, 2.5, and 2.2 required for the same *p* value based on *t* distributions with 2, 5, and 10 degrees of freedom, respectively).

4. The Correct Unit of Analysis — An Example

Since many investigators in a variety of fields have had considerable difficulty choosing the correct unit of analysis, i.e., knowing which source of variation and which n to use in the numerator and denominator of the SE,[66] it is worth considering an illustration. Imagine that we wish to test whether a particular modality of academic training leads to better performance than another modality. Performance is to be assessed (estimated) on a sample of n examination questions. If we compare the results of the modalities using $l = 1$ trainee per modality, the only effect of increasing the number of questions (n) used in the examination is to make us more convinced as to which of *these two* trainee-modality combinations performs better, and by how much. Indeed, unless we were not sure that these two candidates would stand in the same relation to each other on another day, we might need to augment the number (m) of examinations or sessions to take account of "within candidate" variability. Either way, no matter how much we increase n and m, we still cannot infer how well trainees, *in general*, will perform in each of the two modalities. This can only be achieved by increasing *l*, the number of trainees assessed. Of course, the numbers of questions (n) should also be large enough that they avoid the situation where, somehow by chance, the limited number of questions used favored one modality over the other; one hopes that any observed difference between two readers (either in the same modality or in different ones) is not due to the questions selected.

Since cases and readers in an observer performance study are analogous to questions and

trainees in our example, the use of a sufficient number of the same (or matched) cases in both modalities should allow one to consider that the contribution of $(1 - r_c)V_c/n$ to the composite SE(d) is minimal compared to the that of $(1 - r_{br})V_{br}/l$. Then, in planning the statistical power of an observer performance study, one can be guided by the same calculations used for simple comparisons of means taken over observers: one can simply consult nomograms or tables for two sample *t* tests showing the number of subjects (observers) required to have a specified probability of detecting a difference of δ when the projected between reader standard deviation is σ.[67] For matched readers, one consults the table for the one-sample *t* test, where σ is the projected standard deviation of the pair differences. The tables are tabulated in terms of the "signal-to-noise" ratio δ/σ.

The bounded nature of indices (such as A_z and TP_{FP}) means that they are unlikely to have a normal distribution over readers, even if the number of cases used is large. The histogram of A_z values from 168 respondents to the 1983 Nuclear Medicine Imaging set TSA-A (liver) is a graphic illustration of this skewness.[68] Incidentally, it is amusing that the commentary on the histogram noted that it was "satisfyingly Gaussian (although there is no really good reason why it should be) with a mean of 0.88 and a standard deviation of 0.11)." One does not expect this range of variation in experimental studies. If necessary, however, one can transform the indices to a scale (such as the probit or logit) that will make the distribution more symmetric (and thus, make the *t* test more suitable, in spite of the small numbers of readers involved). The lack of normality of indices across readers is less an issue when analyzing a set of within-reader differences, since, at least under the null hypothesis, they should have a distribution which is symmetric and, thus, closer to Gaussian.

5. Being Convinced by Consistent Differences

For those who are uncomfortable with performing parametric tests on such few numbers, the nonparametric equivalents of the *t* tests are an attractive alternative. Indeed, they illustrate the minimum number of readers needed to reach a "significant" difference: if, in a study that uses three readers in one modality and three (unmatched) in the other, the indices for the three readers in one modality all rank higher than those of the three in the other modality, and if this was the hypothesized direction, such a pattern is associated with a *p* value (one sided) of 1/20 or 0.05 using the rank sum test. If a study using five matched readers produces five intermodality d's that are consistently in the hypothesized direction, this pattern is associated with a *p* value (one sided) of 1/32 or 0.03 using the sign test. Many investigators have reported such patterns without formal statistical tests, knowing instinctively that they must be "real". In fact, if they are achieved in spite of the statistical noise caused by low n's and m's, one could argue that they are all the more remarkable. Although they leave the choice of the number of readers up to an investigator's scientific judgement, Swets and Pickett suggest that, even apart from issues of power requirements, a reading test should, as a rule, have "at least several" readers; in another chapter, they "emphasize again that one should strive to work with a reasonably large sample of readers", since small samples can easily give rise to "sampling oddities".

6. More Readers, Fewer Cases?

Often, the number of cases a reader is expected to read limits the number of readers willing to participate. Fortunately, the increasing emphasis on, and increased understanding of, the value of a larger selection of readers and of the fact that the number of cases may be secondary will make it easier to reduce the case numbers and thereby lure more readers into participating. Metz[78] has mentioned the arguments of one colleague that a small, but carefully selected, case sample of as low as 40, but interpreted by a large number of observers, can yield a very respectable study. Berbaum and Dorfman recently conducted a study that was small enough (n = 43) that it allowed them to list the individual characteristics of the 25 D+ cases![69] However, even apart from the practical difficulty of fitting reader-specific ROC curves from such a small

n (which these two authors overcame by pooling the rating data from $l = 6$ observers and extracting six pseudovalues), the number of cases cannot be allowed to be so small that it is impossible to generalize from them. Again, this is a matter for the individual investigator to judge.

K. Toward a General Regression Strategy for ROC Analysis

Many investigators use modern regression methods (such as those developed in the early 1970s for binary response and survival data) to assess, or adjust for the effect of, covariates. They have come to expect that the same level of sophisticated software should be available to deal with unbalanced ROC data, comparison of modalities using small samples, and the effect of patient and characteristics (and possibly other test results) on the performance of a test. Unfortunately, regression methods for ordinal (rating) data were developed much later,[70] and they have not yet appeared in commercial statistical packages. Indeed, even when there are no covariates, rating data are complex enough that one needs a specialized fitting program to formally fit an ROC curve (although the idea of plotting a straight line on binormal axes would suggest a regression approach).

The work of Tosteson and Begg[71] is a promising step toward the flexibility of a regression approach to ROC data. They explain how the model of McCullagh can be used to estimate the basic parameters of a signal detection model, along with parameters which reflect the effect of relevant covariates on the location and shape of the ROC curve. Just like the generalized linear models embodied in the approach that underlies GLIM,[72] several different regression models (ROC curves) can be fit within the same estimation framework simply by changing the form of the "link" (e.g., probit, logit, log[–log]) between the weighted combination of regressors variables and the expected rating response. Thus, one can perform a large variety of regressions within the same computer program. Provided one uses a scale parameter to allow for asymmetric ROC curves, the probit link (the function Φ used above) leads to the familiar Dorfman and Alf model, while the logit link leads to the logistic ROC fitted by Ogilvie and Creelman.[51] The log(–log) link, without a scale parameter, leads to the power-law or biexponential ROC curve. (Incidentally, this last model is the only one that can be fitted directly in the GLIM regression package: as soon as one postulates a scale term, i.e., a model based on bz-a rather than just 1.z-a, the model is no longer linear [since both b and z are parameters].)

As of now, the fitting can be carried out using an interactive computer package called PLUM, which is written in FORTRAN, developed by McCullagh of the University of Chicago for the analysis of ordinal data.

Part of the appeal of their regression approach is being able to use indicator variables to estimate separate ROC curves for distinct groups of cases within the same fit. One can also adjust for relevant covariates, and thus assess the incremental value of a test from its apparent value (the latter may simply reflect the contribution of other relevant clinical information). It is not clear whether this approach can be used to directly compare the performance of two modalities, as is done by the program CORROC. Presumably, one would have to use the rating data from the first modality as a categorical covariate, and test if there was any further gains in accuracy, once this covariate had been accounted for. The authors also point out that by including both direct and interaction terms for variables representing different observers, one can assess whether different observers simply operate at different ROC points on the same curve or at the same points on different curves.

The comments made above concerning the "reader-based" standard errors used by Dorfman and Berbaum also apply here. If one is interested in how diagnostic performance is affected by the characteristics (such as age, sex, hospital) of a sample of n cases, then the n-based inferences of the ordinal regression model apply. In the terminology of analysis of variance, the different age and sex groups (and possibly even the designated readers in those hospitals) can be thought of as "fixed" effects. If, however, the interest is in how readers (in general) perform in one

modality over another, rather than in how one reader's performance is affected by the characteristics of the cases read, then the readers are, in statistical terminology, random effects. If a large source of variation in accuracy is the readers, then the SE in any comparison needs to incorporate it.

L. Using ROC Techniques to Evaluate Quantitative Tests

Although ROC techniques were developed for discrimination tasks that required subjective interpretation and that are, therefore, based on "latent" decision scales, they have become increasingly popular in the evaluation of the performance of tests or prognostic instruments that yield numerical results on "observed" scales.[73-76] If one wishes to take a distribution-free ("nonparametric") approach, the paper by Wieand[58] gives the most comprehensive guide to the statistical treatment of the various accuracy indices. Alternatively, one can take a parametric approach by fitting the two overlapping distributions and algebraicly calculate the required indices from the parameters of these fitted distributions. The drawback to this approach is that, particularly if one estimates a percentile-based quantity such as TP_{FP}, obtaining a good estimate may be quite dependent on correctly specifying the forms of the distributions. The reason is that whereas the A index (or any other "central" index) is quite robust to distributional forms, indices based on extremes are difficult to estimate precisely, even when the correct form of the distribution is known.

A third approach is to categorize the observed scale (i.e., to simulate rating data) and fit ROC curves and indices, using the binormal model for rating data. To some, this replacement of the observed scale of measurement by a latent scale is a step backward in that it ignores the full "grain" inherent in the continuous scale, and they might rightly ask if one would not be better served fitting distributional models directly to the observed data. However, Metz has suggested using the finest possible categorization compatible with the fitting algorithm, i.e., converting the raw data into $2 \times k$ data tables with the largest possible k. A program "LABROC4" which uses this approach is available from him. He surmises that since this preserves the ranking of the observed data, the process may induce a useful transformation to a more inherently Gaussian scale; such a scale might be elusive if one works only with parametric transformations of the observed interval scale. This hypothesis fits with the demonstrated flexibility of the binormal model.

V. CONCLUSION

The use of ROC techniques to evaluate the accuracy of medical diagnosis and prognosis has grown considerably in the past decade. Methods for statistical inference have been established for most situations. However, a number of issues remain to be clarified further. Foremost among these is the need to recognize and discuss the types of questions that ROC studies can and cannot answer. Can ROC techniques be adapted further to the conduct of multicenter imaging trials? Are there simpler substitutes? What should be the role of regression and adjustment techniques? What is the best way to evaluate quantitative tests, and to compare quantitative with qualitative tests? Can we more precisely delineate the incremental contributions of serial tests? How can subjective impressions and objective indicants be merged to provide better detectability? These larger questions and applications must be kept in mind as ROC-based techniques are further developed.

REFERENCES

1. **Green, D.M. and Swets, J.A.,** *Signal Detection Theory and Psychophysics,* John Wiley & Sons, New York, 1966.
2. **Metz, C.E.,** Basic principles of ROC analysis, *Semin. Nucl. Med.,* 8(4), 283, 1978.
3. **Metz, C.E.,** ROC methodology in radiological imaging, *Invest. Radiol.,* 21(9), 720, 1986.
4. **Swets, J.A. and Pickett, R.M.,** *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory,* Academic Press, New York, 1982.
5. **Begg, C.B.,** Statistical methods in medical diagnosis, *CRC Crit. Rev. Med. Inf.,* 1(1), 1, 1986.
6. **Swets, J.A.,** Indices of discrimination or diagnostic accuracy: their ROCs and implied models, *Psychol. Bull.,* 99(1), 100, 1986.
7. **Swets, J.A.,** Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psychol. Bull.,* 99(1), 181, 1986.
8. **Lusted, L.B.,** General problems in medical decision making with comments on ROC analysis, *Semin. Nucl. Med.,* 8(4), 299, 1978.
9. **McNeil, B.J., Keeler, E., and Adelstein, J.S.,** Primer on certain elements of medical decision making, *N. Engl. J. Med.,* 293, 211, 1975.
10. **Swets, J.A., Pickett, R.M., Whitehead, S.F., Getty, D.J., Schnur, J.A., Swets, J.B., and Freeman, B.A.,** Assessment of diagnostic technologies, *Science,* 205, 753, 1979.
11. **Swets, J.A.,** ROC analysis applied to the evaluation of diagnostic techniques, *Invest. Radiol.,* 14, 109, 1979.
12. **Turner, D.A.,** An intuitive approach to receiver operating characteristic curve analysis, *J. Nucl. Med.,* 19(2), 213, 1978.
13. **Weinstein, M.C. and Fineberg, H.V.,** *Clinical Decision Analysis,* W.B. Saunders, Philadelphia, 1980.
14. **Swets, J.A.,** Indices of discrimination or diagnostic accuracy: their ROCs and implied models, *Psychol. Bull.,* 99(1), 100, 1986.
15. **Gilbert, G.K.,** Finley's tornado predictions, *Am. Meteorol. J.,* 1, 167, 1885.
16. **Finley, J.P.,** Tornado predictions, *Am. Meteorol. J.,* 1, 5, 1884.
17. **Browne, W.S. and Newman, T.B.,** Are all significant p values created equal? The analogy between diagnostic tests and clinical research, *JAMA,* 257, 2459, 1987.
18. **Harris, J.M.,** The hazards of bedside Bayes, *JAMA,* 246, 2602, 1981.
19. **Swets, J.A.,** Sensitivities and specificities of diagnostic tests, *JAMA,* 248, 548, 1982.
20. **Mitchell, C.R. and Sorenson, J. A.,** Application of the Bunch transform in clinical FROC nodule detection studies, in Proc. Chest Imaging Conf.-87, University of Wisconsin, Madison, 1987.
21. **Bunch, P.C., Hamilton, J.F., Sanderson, G.K., and Simmons, A.H.,** A free response approach to the measurement and characterization of radiographic observer performance, *SPIE Vol. 27, Opt. Instrum. Med.,* 6, 124, 1977.
22. **Rozanski, A., Diamond, G.A., Forrester, J.S., Berman, D.S., Morris, D., Jones, R.H., Okada, R., Freeman, M., and Swan, H.J.C.,** Should the intent of testing influence its interpretation?, *J. Am. Coll. Cardiol.,* 7, 17, 1986.
23. **Fletcher, R.H.,** Carcinoembryonic antigen, *Ann. Intern. Med.,* 104, 66, 1986.
24. **Kundel, H.L.,** Evaluation of observer performance, in Proc. Chest Imaging Conf. '87, University of Wisconsin, Madison, 1987.
25. **Revesz, G., Kundel, H.L., and Bonitatibus, M.,** The effect of verification on the assessment of imaging techniques, *Invest. Radiol.,* 18, 194, 1983.
26. **Henkelman, R.M., Kay, I.B., and Bronskill, M.J.,** Receiver Operator Characteristic (ROC) Analysis without Truth, unpublished document, Department of Medical Biophysics, University of Toronto, Toronto, 1986.
27. **Kundel, H.L. and Revesz, G.,** The evaluation of radiographic techniques by observer tests: problems, pitfalls and procedures, *Invest. Radiol.,* 9, 166, 1974.
28. **Kundel, H.L.,** The evaluation of observer performance, Proc. Chest Imaging Conf. '87, Peppler, W. W. and Alter, A., Eds., Department of Medical Physics and Medical Physics Publishing, Madison, WI, 1988, 291.
29. **Oestmann, J.W., Greene, R., Kushner, D., Bourgouin, P.M., Linetsky, L., and Llewellyn, H.J.,** Viewing time and detectability of lung lesions, *Radiology,* in press.
30. **Ransohoff, D.F. and Feinstein, A.R.,** Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *N. Engl. J. Med.,* 299, 926, 1978.
31. **Diamond, G.A.,** Reverend Bayes' Silent Majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography, *Am. J. Cardiol.,* 57, 1175, 1986.
32. **Greenes, R.A. and Begg, C.B.,** Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selectively verified patients, *Invest. Radiol.,* 20, 751, 1985.
33. **Begg, C.B. and Greenes, R.A.,** Assessment of diagnostic tests when disease verification is subject to verification bias, *Biometrics,* 39, 207, 1983.

34. **Diamond, G.A., Rozanski, A., Forrester, J.S., Morris, D., Pollock, B.H., Staniloff, H.M., Berman, D.S., and Swan, H.J.C.,** A model for assessing the sensitivity and specificity of tests subject to selection bias, *J. Am. Coll. Cardiol.*, 7, 17, 1986.
35. **Gray, R., Begg, C. B., and Greenes, R. A.,** Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Med. Decision Making*, 4, 151, 1984.
36. **Diamond, G.A.,** ROC STEADY, a receiver operating characteristic curve which is invariant relative to selection bias, *Med. Decision Making*, 7, 238, 1987.
37. **Hanley, J.A. and Begg, C.B.,** Response to ROC STEADY, *Med. Decision Making*, 7(4), 244, 1987.
38. **Feinstein, A.R.,** *Clinical Epidemiology: The Architecture of Clinical Research*, W.B. Saunders, Philadelphia, 1985.
39. **Begg, C.B.,** Biases in the assessment of diagnostic tests, *Stat. Med.*, 6, 411, 1987.
40. **Begg, C.B. and McNeil, B.J.,** Prospective assessments of radiologic tests: control of bias and other design considerations, *Radiology*, submitted.
41. **Kong, A., Barnett, O., Mosteller, F., and Youtz, C.,** How medical professionals evaluate expressions of probability estimates, *N. Engl. J. Med.*, 315, 740, 1986.
42. **Hannequin, P., Liehn, J.C., Fortier, A., Elaerts, J., and Valeyre, J.,** Comparison of phase analysis with factor analysis in equilibrium gated radionuclide angiography, *Nucl. Med. Commun.*, 7, 857, 1986.
43. **Linnet, K.,** Comparison of quantitative diagnostic tests: type I error, power, and sample size, *Stat. Med.*, 6, 147, 1987.
44. **Greenhouse, S.W. and Mantel, N.,** The evaluation of diagnostic tests, *Biometrics*, 6, 399, 1950.
45. **Grey, D.R. and Morgan, B.J.T.,** Some aspects of ROC curve fitting: normal and logistic models, *J. Math. Psychol.*, 9, 128, 1972.
46. **Dorfman, D.D. and Alf, E.,** Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data, *J. Math. Psychol.*, 6, 487, 1969.
47. **Egan, J.P.,** *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
48. **Swets, J.A., Tanner, W.P., and Birdsall, T.G.,** Decision processes in perception, *Psychol. Rev.*, 68, 301, 1961.
49. **Hanley, J.A.,** The robustness of the binormal model used to fit ROC curves, *Med. Decision Making*, 8, 197, 1988.
50. **Swets, J.A.,** Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance, *Psychol. Bull.*, 99(1), 181, 1986.
51. **Ogilvie, J.C. and Creelman, C.D.,** Maximum likelihood estimation of ROC curve parameters, *J. Math. Psychol.*, 5, 377, 1968.
52. **Dorfman, D.D. and Alf, E.,** Maximum likelihood estimation of parameters of signal detection theory—a direct solution, *Psychometrika*, 33, 117, 1968.
53. **Macmillan, N.A. and Kaplan, H.L.,** Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates, *Psychol. Bull.*, 98, 185, 1985.
54. **Hanley, J.A. and McNeil, B.J.,** The meaning and use of the area under an ROC curve, *Radiology*, 143(1), 29, 1982.
55. **Bamber, D.,** The area above the ordinal dominance graph and the area below the receiver operation characteristic graph, *J. Math. Psychol.*, 12, 387, 1975.
56. **Habicht, J.-P.,** Assessing diagnostic technologies, *Science*, 207, 1414, 1980.
57. **Swets, J.A. and Pickett, R.M.,** Assessing diagnostic technologies: reply to Habicht, *Science*, 207, 1415, 1980.
58. **Wieand, S., Gail, M.H., James, K.L., and James, B.R.,** A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data, *J. Am. Stat. Assoc.*, submitted.
59. FORTRAN programs ROCFIT, CORROC, and ROCPWR, available from C. Metz, Department of Radiology, University of Chicago, Chicago.
60. **Hanley, J.A. and McNeil, B.J.,** A method of comparing the areas under receiver operating characteristic curves derived from the same set of cases, *Radiology*, 148(3), 839, 1983.
61. **McNeil, B.J. and Hanley, J.A.,** Statistical approaches to the analysis of receiver operating characteristic (ROC) curves, *Med. Decision Making*, 4(2), 137, 1984.
62. **Metz, C.E., Wang, P.-L., and Kronman, H.B.,** A new approach for testing the significance of differences between ROC curves from correlated data, in *Information Processing in Medical Imaging*, Deconink, F., Ed., Nijhoff, The Hague, 1984, 432.
63. **Dorfman, D.D. and Berbaum, K.S.,** RSCORE-J: pooled rating-method data: a computer program for analyzing pooled ROC curves, *Behav. Res. Methods Instrum. Comput.*, 18, 452, 1986.
64. **McClish, D.K.,** Comparing the areas under more than two independent curves, *Med. Decision Making*, 7, 149, 1987.
65. **Hanley, J.A.,** Observer performance statistics, Proc. Chest Imaging Conf. '87, Peppler, W. W. and Alter, A., Eds., Department of Medical Physics and Medical Physics Publishing, Madison, WI, 1988, 291.
66. **Whiting-O'Keefe, Q.E.,** Choosing the correct unit of analysis in medical care experiments, *Med. Care*, 22, 1101, 1984.

67. *Handbook of Tables for Probability and Statistics,* 2nd ed., CRC Press, Cleveland, 1968.
68. **Hermann, G.A.,** Old pathways and new directions in nuclear medicine imaging, *Pathologist,* 147, 1984.
69. **Berbaum, K.S., Franken, E.A., Dorfman, D.D., Ell, S.R., Lu, C.H., Smith, W., and Abu-Yousef, M.M.,** Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs, *Invest. Radiol.,* 21, 532, 1986.
70. **McCullagh, P.,** Regression models for ordinal data (with discussion), *J. R. Stat. Soc.,* 42, 109, 1980.
71. **Tosteson, A.N.A. and Begg, C.B.,** A general regression methodology for ROC curve estimation, *Med. Decision Making,* 1988.
72. **Baker, R. J. and Nelder, J.A.,** The GLIM System, Release 3, Generalised Interactive Modelling, manual, Numerical Algorithms Group, Oxford, 1978.
73. **Erdreich, L.S. and Lee, E.T.,** The use of relative operating characteristic analysis in epidemiology, *Am. J. Epidemiol.,* 114, 649, 1981.
74. **Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A.,** Regression modelling strategies for improved prognostic prediction, *Stat. Med.,* 3, 143, 1984.
75. **Slap, G.B., Connor, J.L., Wigton, R.S., and Schwartz, J.S.,** Validation of a model to identify young patients for lymph node biopsy, *JAMA,* 255, 2768, 1986.
76. **Westerfield, B.D., Pals, G., Lamers, C.B.H.W., Defize, J., Pronk, J.C., Frants, R.R., Ooms, E.C.M., Kreuning, J., Kostense, P.J., Eriksson, A.W., and Meuwissen, S.G.M.,** Clinical significance of pepsinogen A isozymogens, serum pepsinogen A and C levels, and serum gastrin levels, *Cancer,* 59, 952, 1987.
77. **England, W.,** personal communication.
77a. (Note added in press) **England, W. L.,** An exponential model used for optimal threshold selection on ROC curves, *Med. Decision Making,* 8, 120, 1988.
78. **Metz, C.E.,** personal communication.