Med Decis Making
Vol. 4, No. 2, 1984

Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves

Barbara J. McNeil, M.D., Ph.D.,
and James A. Hanley, Ph.D.

In this article we review published and some unpublished work in statistical analyses of ROC curves. We describe both single and joint indices and indicate the approaches that have been taken to consider between-reader variations and correlations, within-reader variations, and variations and correlations between cases.

We then discuss in detail a single index, the *TP* ratio at a fixed *FP* ratio (designated TP_{FP}), or the *FP* ratio at a fixed *TP* ratio (designated FP_{TP}). We show how to calculate confidence limits around *any* point on the curve; we further show, using the conventional Dorfman and Alf program and the jackknifing technique, how to calculate these confidence limits for multiple curves derived from the same sample of patients. (Med Decis Making 4:137–150, 1984)

Over the past 20 years investigators have proposed a number of indices to describe receiver operating characteristic (ROC) curves, and have developed statistical techniques to compare two or more curves [1–3]. In this article we will briefly review past work in this field and will indicate an approach to dealing with the problem of comparing differences between two or more ROC curves at a single operating point in either the true positive (*TP*) or false positive (*FP*) dimension. We shall first review commonly used indices and discuss their statistical evaluation for experiments with unpaired and with paired designs.

Indices to Describe ROC Curves

Area-Related Measurements. Much recent work has involved the

This work was supported in part by a grant from the Hartford Foundation and the National Cancer Institute (CA 28847).

From the Departments of Radiology and Preventive Medicine and Clinical Epidemiology, Harvard Medical School and the Brigham and Women's Hospital, and the Department of Epidemiology and Health, McGill University. Please address requests for reprints to Dr. McNeil, Department of Radiology, Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, USA.

area under the ROC curve. When discrete *rating* data are used (e.g., 5-point or 6-point rating scales for radiology imaging experiments), and when ROC curves are assumed to be based on two underlying Gaussian distributions, a maximum likelihood estimation program by Dorfman and Alf [4] can be used to fit data points to a smooth curve and to derive thereby (along with other indices) the area under this fitted curve and its associated standard error. This area is designated A_z and ranges in value from 0.0 to 1.0.

If the ROC curve is drawn by connecting the pairs of observed *TP* and *FP* ratios, and if the trapezoidal areas are summed, the resulting nonparametric area is designated $P(A)$. The common availability of the Dorfman and Alf program, underestimation of the area, and undue dependence on extreme points have decreased the use of the $P(A)$ index for rating data.

When *continuous* data are available (as from chemistry laboratory tests, white cell counts, individualized predictions from logistic regression or discriminant analysis) and ROC curves are created, no assumptions on underlying distributions need be made to obtain area measurements. Instead, Bamber's recognition of the equivalence between the area under the ROC curve and the Wilcoxon statistic W allows immediate and direct calculation of W and hence the area [5]. Hanley and McNeil's [6] derivation of a closed form approximate expression for the standard error associated with the Wilcoxon statistic can be used to approximate the standard error of the area.

Slope-Related and Intercept-Related Indices. When ROC curves are assumed to be based on underlying Gaussian distributions, the expected ROC points should follow a straight line when plotted on binormal coordinate paper [2]. This assumption has led to the development of a series of indices related to the slope and intercept of the straight line fit to the observed ROC points. In general, these are derived after observed data are fitted with the Dorfman and Alf program. The true slope of the line is designated b and its true intercept a (estimates of these are designated $\hat{a}$ and $\hat{b}$; however, for simplicity, throughout this manuscript the estimate sign will be omitted and all symbols a and b will designate estimates); a divided by b is called Δm, an index commonly used in radiology phantom studies. Conceptually, Δm is a form of a standardized difference between the means of two normal distributions (which may have different variances). Other derived indices are also available (see [2] for a complete summary). The Dorfman and Alf maximum likelihood program provides estimates for a and b, var(a), var(b), and covar(a, b). Either of the above pairs of indices (e.g., Δm and the slope or a and b), or any other pair derived from them, is sufficient to describe fully a binormal ROC curve.

Statistical Treatment: Area Index

Unpaired Data. When the area index is used and we have only one reader (or if we have more than one reader but negligible between-reader variations), statistical comparisons between two areas can be made by testing their difference using the formula

$$\text{critical ratio} = \frac{\text{Area}_1 - \text{Area}_2}{\text{SE}(\text{Area}_1 - \text{Area}_2)} \tag{1}$$

and comparing the critical ratio with the table of the normal distribution. The areas and associated standard errors can be obtained directly from the Dorfman and Alf program if data consist of ratings, or from the Wilcoxon statistic if data consist of measurements [6]. The Wilcoxon statistic can also be used to provide a closed form expression for the standard error and thereby to estimate sample size and power.

Paired Data. Statistical tests for paired comparisons are more difficult, whatever the index used. Swets and colleagues have provided a general expression to take into account three types of variances (and associated correlations) that may be present in a paired comparison: the variances and covariances induced by using the same cases; the variance induced by having a reader read the same set of cases more than once (within-reader consistency); and the variances and covariances induced by having multiple readers read the same set of cases (between-reader consistency) [1]. For the area index, Hanley and McNeil [7] suggested a more feasible method of calculating the correlation between areas induced by studying the same cases, a quantity that is otherwise difficult to assess.

Slope-Intercept Index

Comparing ROC curves simultaneously on both their slopes and intercepts is the most rigorous statistical approach in that identity between two curves can exist *only* if there is a complete coincidence of the curves. (This coincidence is not necessary, as indicated below, for area measurements to be equal.) For unpaired comparison of slope-intercept pairs, Metz has developed a test statistic that follows a chi-square distribution if the two sets of rating scale data (and hence their summary indices) derive from the same underlying ROC curve [8,9]. For paired data, Metz has extended this work by using a "two-dimensional" Dorfman and Alf approach to estimate two correlated slope-intercept pairs [10]; this is not yet generally available, however.

TP Point Index

All of the above indices and resulting statistical techniques assume that the investigator is interested in an overall one-dimensional index of performance for the entire ROC curve, or in a description of the *entire* curve. Such may not always be the case, however. In some cases, for example, two ROC curves might cross, and although the areas for the two imaging modalities may be almost the same, in the clinical range of interest one may be superior to the other. In addition, even if two curves do not cross, one could imagine that differences would exist at one point (the clinically relevant one, perhaps) on the curve but would not be detected in any global test. In either of these two situations raw rating scale data may not allow direct comparisons to be made, since it is unusual for indentical interpretive critera (i.e., identical *FP* ratios) to exist in different experiments. In particular, this means that the observed *TP* ratio on one curve at a particular criterion cannot be directly compared with that on another curve, because the associated *FP* ratios observed may be different.

Methods

In this section we elaborate an approach to comparing differences between two ROC curves at one point (either *TP* or *FP*). This is also part of a comprehensive computer package being prepared by Metz [10]. Our approach is based on fitted *TP* and *FP* ratios, obtained from fitted parameters from the Dorfman and Alf maximum likelihood estimation program for rating-based ROC curves [1,4]. The method yields confidence intervals around true positive ratios at a fixed false positive ratio (designated TP_{FP}) or around false positive ratios at a fixed true positive ratio (designated FP_{TP}). Illustrative examples are presented.

The maximum likelihood estimation program of Dorfman and Alf provides parameters that allow calculation of *TP* ratios at any *FP* ratio and thus provides the basis for comparing two ROC curves at either the same *TP* ratio or the same *FP* ratio. In brief, the relevant outputs for this purpose are: (1) *a*, the normal deviate value of the intercept of the ROC curve with the *y* axis; and (2) *b*, the slope of the ROC curve obtained from ROC curves plotted on normal deviate axes. The equation for this purpose in normal deviate space (Z) for the *TP* ratio is:

$$Z_{TP} = bZ_{FP} - a. \tag{2}$$

The quantity Z_{TP} on the Z scale can be converted to *TP* on the 0–100% scale by determining what percentage of the normal probability distribution lies above (i.e., to the right of) Z_{TP}.

The Dorfman and Alf output also provides the variance and covariance terms var(a), var(b), and covar(a, b). These can be used to calculate the sampling variance or uncertainty of Z_{TP} and thus of *TP* itself. This is done in two steps. First we calculate a confidence interval for Z_{TP} (in the Z scale), and second we transform the confidence interval back into the usual 0–100% scale. The relevant equation for calculating the variance around a Z_{TP} ratio is:

$$\mathrm{var}(bZ_{FP} - a) = Z_{FP}^2\mathrm{var}(b) + \mathrm{var}(a) - 2Z_{FP}\mathrm{covar}(a, b). \tag{3}$$

The confidence interval for Z_{TP} becomes $Z_{TP}\pm$ some multiple of the SE of Z_{TP}, i.e., ± some multiple of $\sqrt{\text{variance }(Z_{TP})}$. The "multiple" can be chosen from the tables of the normal distribution; since the parameter estimates a and b are maximum likelihood estimates, if the sample is large they should have Gaussian distributions, regardless of the underlying models. (Moreover, Metz has shown empirically [8] that a and b have distributions reasonably near normal even for n as low as 50.) Once one has obtained a confidence interval for Z_{TP}, it is a simple matter of translating the upper and lower limits for *TP*, using the normal probability tables. Generally, the confidence interval around the *TP* (and also *FP*) ratios will be asymmetric, especially as the values move away from 50 percent.

SIGNIFICANCE TESTS FOR COMPARING TWO *TP*'S AT A COMMON *FP*. Because there is a one-to-one relation between Z_{TP} and *TP* itself, two fitted *TP*'s (from two different experiments) can be statistically compared on either the Z_{TP} or the *TP* scale. Since the sampling distributions tend to be more symmetric in the (open-ended) Z scale, it is more appropriate to perform tests on this scale. To test whether there is a statistically significant difference between two fitted *TP*'s (e.g., TP_1 and TP_2, both at the same *FP* ratios), the following critical ratio is calculated

$$CR = \frac{Z_1 - Z_2}{\text{SE }(Z_1 - Z_2)} \tag{4}$$

and compared to the normal distribution. The denominator of the critical ratio will depend on whether the Z_1 and Z_2 are from two modalities evaluated on the same or a different sample of patients and on whether there are several readers or several rereadings within one reader. For the moment we ignore the latter two sources of variation in the denominator of equation (4).

In comparisons of *TP*'s involving separate (independent, unpaired) samples, the SE of the differences is simply $[\text{var}(Z_1)+\text{var}(Z_2)]^{1/2}$, where $\text{var}(Z_1)$ and $\text{var}(Z_2)$ are each calculated as in equation (3). For comparisons involving paired samples, Z_1 and Z_2 will tend to be correlated on repeated samples of patients. There are two ways to calculate this (reduced) variance.

The first method uses the variance and covariance terms given as output from Metz's "two-dimensional" Dorfman and Alf program [10]. For this purpose we write $Z_1 = b_1 Z_{FP} - a_1$ and $Z_2 = b_2 Z_{FP} - a_2$, so that

$$\begin{aligned} Z_1 - Z_2 &= (b_1 - b_2)Z_{FP} - (a_1 - a_2) \\ \text{var}(Z_1 - Z_2) &= Z_{FP}^2\,\text{var}(b_1 - b_2) + \text{var}(a_1 - a_2) \\ &\quad - 2Z_{FP}\,\text{covar}(b_1 - b_2,\ a_1 - a_2) \\ \text{SE }(Z_1 - Z_2) &= [\text{var}(Z_1 - Z_2)]^{1/2}. \end{aligned} \tag{5}$$

To obtain the above components we use the following identities:

$$\begin{aligned}\mathrm{var}(b_1-b_2) &= \mathrm{var}(b_1)+\mathrm{var}(b_2)-2\mathrm{covar}(b_1,\ b_2)\\ \mathrm{var}(a_1-a_2) &= \mathrm{var}(a_1)+\mathrm{var}(a_2)-2\mathrm{covar}(a_1,\ a_2)\\ \mathrm{covar}(b_1-b_2,\ a_1-a_2) &= \mathrm{covar}(b_1,\ a_1)-\mathrm{covar}(b_1,\ a_2)-\mathrm{covar}(b_2,\ a_1)\\ &\quad+\mathrm{covar}(b_2,\ a_2)\end{aligned}$$

The second method is useful if Metz's program is unavailable. We can approximate $\mathrm{var}(Z_1-Z_2)$ by the method of jackknifing (see Fleiss [11] for a general introduction to jackknifing, and Efron [12], equations 6.11 and 6.17, for the jackknife variance in two-sample problems). When the rating data come from n_N normals and n_A abnormals, the jackknife method consists of obtaining n_N+n_A different estimates of Z_1-Z_2 and using the quantity

$$\sum_{i=1}^{n_N+n_A} [(Z_1-Z_2)-(Z_1^i-Z_2^i)]^2 \qquad (6)$$

as the jackknife variance of (Z_1-Z_2). The quantity (Z_1-Z_2) is obtained from the entire data set. The ith jackknife estimate $(Z_1^i-Z_2^i)$ is obtained by fitting two separate ROC curves to the data set (of n_N+n_A-1 subjects) formed by deleting the paired ratings of the ith subject from the original data set. Although this may sound computer-intensive, the numbers of pairs of ROC curves to be fitted depend on the number of rating categories and not on the number of patients. For example, data on a five-point rating scale will involve at most 10 distinct values for each modality; the n_A+n_N quantities being summed in equation (6) will occur in multiples.

Extension of the Method to Multiple Readers. As we have done so far, we still base the analysis on equation (4), but now need to include between-reader (S_{br}) and within-reader (S_{wr}) variances. For this purpose we use equation (5) (Chapter 4) from Swets [1], as shown here:

$$\mathrm{SE}_{(\mathrm{diff})} = 2^{1/2}\left[S^2_{c+wr}(1-r_{c-wr})+\frac{S^2_{br+wr}}{\ell}(1-r_{br-wr})-S^2_{wr}\right]^{1/2}, \qquad (7)$$

r_{br-wr} = the observable correlation between the Z_{TP}'s obtained when a set of readers reads the same cases in the two settings

r_{c-wr} = the observable correlation between the Z_{TP}'s obtained when a single reader reads the same set of cases in two settings

$S^2_{c+wr}=S^2_c+S^2_{wr}$, the observable variance in Z_{TP} that would be found by having one reader read once each of a set of different case samples

$S^2_{b+wr}=S^2_{br}+S^2_{wr}$, the observable variance in Z_{TP} that would be found by having one case sample read once by each of a set of different readers

S^2_{wr} = the observable variance in Z_{TP} that would be found by having one reader read one case sample on two or more independent occasions

ℓ = the number of independent readers

The quantity $2S^2_{c+wr}(1-r_{c-wr})$ (taking the $2^{1/2}$ inside the square brackets) is the same quantity whose calculations we have described as $\text{var}(Z_1 - Z_2)$. The different notation has two explanations: (1) Swets and Pickett assume equal variances for each of the three components; and (2) Swets and Pickett talk explicitly in terms of the correlation r_{c-wr}, while our variance formula calculates it implicitly, using covariances of the component parameters a_1, a_2, b_1, and b_2. Also, Swets and Pickett work in the closed-ended *TP* scale, whereas we prefer the open-ended Z_{TP} scale. The reader is referred to [1] for calculation of the other components; in all cases estimates are made in the *Z* space rather than in the *TP–FP* space. (The reader should note that even though the examples in [1] are based on area measures, they apply equally well to other measures.)

Results

CALCULATION OF CONFIDENCE LIMITS FOR A SINGLE POINT ON AN ROC CURVE. Consider the sample ROC curve data in Table 1. The Dorfman and Alf maximum likelihood program [4] provides the following data:

$$\begin{aligned} a &= 1.657 & \text{var}(a) &= 0.0974 \\ b &= 0.713 & \text{var}(b) &= 0.0467 \\ & & \text{covar}(a, b) &= 0.0478 \end{aligned}$$

Equations (2) and (3) allow calculation of Z_{TP} and TP_{FP} ratios for any *FP* values; they are shown in Table 2 at *FP* values of 5%, 10%, and 20%. Consider one detailed calculation for illustrative purposes. At an *FP* ratio of 5%, Z_{FP} (the value above which lies 5% of the normal distribution) equals 1.645 and Equation (2) becomes

$$\begin{aligned} Z_{TP} &= 0.713(1.645) - 1.657 \\ &= -0.48. \end{aligned}$$

The value -0.48 becomes 68.44 percent by referring to the table of normal distribution, and represents the probability to the right of the value -0.48.

Table 1. ROC Curve Obtained on a Five-Point Rating Scale

			Rating*			
	1	**2**	**3**	**4**	**5**	**Total**
Nondiseased	2	11	6	6	33	58
Diseased	33	11	2	2	3	51

*A value of 1 corresponds to a rating of definitely abnormal and a value of 5 to definitely normal.

Table 2. Predicted True-Positive Values at Fixed False-Positive Values, Using Equation (2)

FP	5%	10%	20%
Z_{TP}	−0.48	−0.74	−1.06
TP	68.44	77.04	85.54
SE of Z_{TP}	0.26	0.23	0.35
95% CI on Z_{TP}	(−1.03, 0.07)	(−1.19, −0.28)	(−1.76, −0.36)
95% CI on *TP*	(84%, 49%)	(88%, 39%)	(96%, 36%)

Equation (3) gives confidence limits about this value as follows:

$$\begin{aligned}\mathrm{var}(Z_{TP}) &= (1.645)^2(0.0467)+0.0974-2(1.645)(0.0478)\\ &= 0.0665\\ \mathrm{SE}(Z_{TP}) &= 0.2579.\end{aligned}$$

The 95 percent confidence interval for Z_{TP} is thus $-0.48 \pm 1.96(0.2579)$, or $(-0.99$ to $0.03)$. Using the same method used to transform $Z = -0.48$ to a rounded *TP* value of 68 percent, we can transform the $Z = -0.99$ and 0.03 back to upper and lower *TP* limits of 84 percent and 49 percent.

Table 2 summarizes the standard errors obtained in this way for three points on the above ROC curve.

COMPARING TWO CURVES (UNPAIRED DATA) AT ONE TP_{FP}. Table 3 displays two sets of rating data for previously published results on gallium

Table 3. Ranking Data For an Unpaired Experiment [13]

			Rating* – BWH		
	1	**2**	**3**	**4**	**5**
Nondiseased	4	0	4	13	12
Diseased	19	1	1	6	5

			Rating* – JHH		
	1	**2**	**3**	**4**	**5**
Nondiseased	3	1	3	2	11
Diseased	18	1	3	6	12

BWH		**JHH**	
a	= 0.6665	a	= 0.7631
b	= 0.4316	b	= 0.6969
var(a)	= 0.07234	var(a)	= 0.1822
var(b)	= 0.03639	var(b)	= 0.2021
covar(a, b)	= 0.0163	covar(a, b)	= 0.1257

*1 = definitely abnormal; 5 = definitely normal

Table 4. **Rating Data on CT Scans from a Paired Experiment [14]**

		Nondiseased patients, $n=54$							**Diseased patients, $n=35$**					
		Read with History							**Read with History**					
		1	2	3	4	5	Total		1	2	3	4	5	Total
	5	–	1	–	–	–	1	5	–	–	–	2	24	26
Read	4	5	4	1	1	–	11	4	–	–	–	3	2	5
Without	3	3	1	1	–	–	5	3	–	–	2	–	–	2
History	2	1	3	–	1	–	5	2	–	–	–	–	–	–
	1	31	–	–	1	–	32	1	–	1	–	1	–	2
	Total	40	9	2	3	–	54		–	1	2	6	26	35

ROC curve parameters (1 = with history; 2 = without history)*

Maximum likelihood estimates		**Variances and covariances** $\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$
$\hat{a}_1 = 3.60$	$\hat{a}_1$	1.2288	0.6495	0.1712	0.0757
$\hat{b}_1 = 1.29$	$\hat{b}_1$		0.4043	0.0542	0.0378
$\hat{a}_2 = 1.80$	$\hat{a}_2$			0.1552	0.0681
$\hat{b}_2 = 0.59$	$\hat{b}_2$				0.0533

*Derived using the method from [10].

citrate imaging in the search for a focal source of sepsis [13]. The top set of data was obtained at the Brigham and Women's Hospital (BWH) using an Anger camera, and the bottom set at Johns Hopkins Hospital (JHH) using a rectilinear scanner. The output from the Dorfman and Alf program is shown in Table 3 and allows calculation, as shown above, of Z_{TP} (at FP values of 10%) for each of the two settings. They are as follows:

$$\begin{aligned}(\text{BWH})Z_{TP} &= 0.4316(1.28)-0.6665 = -0.11, \text{ or } 54\%.\\ (\text{JHH})Z_{TP} &= 0.6969(1.28)-0.7631 = +0.13, \text{ or } 45\%.\end{aligned}$$

Using the approach described above, 95 percent confidence limits for the BWH are 31.2 percent to 76.1 percent, and for the JHH data 15.6 percent to 77.3 percent.

COMPARING TWO CURVES (PAIRED DATA) AT ONE TP_{FP}. Table 4 displays two sets of rating data from a single individual reading computed tomograms of the head without (rows) and with (columns) clinical history [14]. The values of a_1 and b_1 from the Metz program were 3.60 and 1.29, respectively, for the data with history, and the values for a_2 and b_2 were 1.80 and 0.59, respectively, for the data without history. The covariances are shown in the bottom of Table 4. Thus, at $FP = 10\%$ ($Z_{FP} = 1.28$),

$$\begin{aligned}Z_1 &= 1.29(1.28)-3.60 = -1.95 \text{ (or } 97.44\%)\\ Z_2 &= 0.59(1.28)-1.80 = -1.05 \text{ (or } 85.31\%)\\ Z_1 - Z_2 &= (1.29-0.59)(1.28)-(3.60-1.8) = -0.90.\end{aligned}$$

The three subcomponents, calculated as for equation (5), yield values of 0.3820, 1.0416, and 0.5877, so that equation (5) yields

$$\begin{aligned}\mathrm{var}(Z_1 - Z_2) &= (1.28)^2 0.3820 + 1.0416 - 2(1.28)\ 0.5877\\ &= 0.1630\\ \mathrm{SE}(Z_1 - Z_2) &= \sqrt{0.1630} = 0.40.\end{aligned}$$

Thus the critical ratio is $-0.90/0.40$, or 2.25, indicating that at $FP = 10\%$ TP's obtained with history are statistically higher.

With the jackknife technique, the paired ratings of successive patients were eliminated, as described in the Appendix, to create $54 + 35 = 89$ different data sets, each with 88 patients. Using the steps in the Appendix, we obtained the jackknife estimate of $\mathrm{SE}(Z_1 - Z_2) = \sqrt{0.1966} = 0.44$, which is only 10 percent higher than the more parametric SE of 0.40 calculated by the "paired binormal" model of Metz.

Discussion

The work was motivated by two concerns with the use of area indices for comparing two ROC curves: (1) That two curves might cross and in such cases similar areas might result; and (2) that even in the absence of crossing curves similar areas might result when, in fact, statistical differences could exist in the region of clinical interest. These concerns pointed to the need to make comparisons at single points on either the *TP* or the *FP* axis. In the process of developing the analysis discussed here we realized that there was a need to provide a brief overview of commonly used indices for ROC analysis, whether paired or unpaired experimental designs were used.

The major point of our review and analysis is this: Once we assume binormal distributions of an ROC curve, all statistical properties are determined by the parameters a and b of the maximum likelihood fit to the data. This paper has emphasized the use of these parameters for calculating confidence limits around single *TP* or *FP* points *anywhere* along the ROC curve. Others, particularly Swets and Pickett [1], have discussed confidence limits explicitly in relationship to area measurements and *TP* points corresponding to observed *TP-FP* pairs.

To put this work in perspective it is worthwhile to summarize previous work in the general area of statistical analyses of ROC curves. The work falls along two lines: (1) the use of a single index (for example, the area or TP_{FP}) versus joint indices (for example, slope and intercept); and (2) consideration of between-reader variations and correlations, within-reader variations, and variations and correlations between cases.

Swets and Pickett [1] give formulas for single indices that cover all possible experimental designs. Hanley and McNeil [6] elaborated on statistical considerations relating to a specific single index, namely the area; because of the unique relationship of the area to the Wilcoxon statistic it is possible to calculate explicitly the associated standard error due to case sampling.

The second paper of Hanley and McNeil [7] also deals with areas, but for ROC curves derived from the same set of cases. It thus complements the work of and general formulas suggested by Swets and Pickett [1]. The present investigation deals with another single index, this time TP_{FP}, and gives explicit formulas for confidence intervals at any point on a single ROC curve. For comparing two TP_{FP}s derived from the same sets of patients, we were able to use the conventional Dorfman and Alf program coupled with the jackknifing technique. We could obviously extend this approach to multiple readers by using the general approach of Swets and Pickett.

Metz [8-10] has emphasized joint indices (slope and intercept), although of course any single index can be derived from them. His programs consider only variations due to case sampling.

One other point is worth emphasizing. All of the comparisons made in this paper have been done in the Z space rather than in the linear TP–FP space. The latter has the disadvantage of having asymmetric sampling distributions around 0 percent and around 100 percent, because of the closed nature of the scale. The Z scale, being more open-ended, seems more appropriate because the sample distributions are more likely to be symmetric throughout all TP and FP ratios of interest.

Appendix: Illustration of Jackknife Technique

To obtain $\mathrm{var}(Z_1 - Z_2)$, where Z_1 and Z_2 refer to the two Z_{TP}'s obtained from the same set of patients evaluated by two modalities.

(1) Data from Table 4 are used to fit separate ROC curves for each modality (ND = nondiseased; D = diseased). Note that the Z_1 and Z_2 values here are derived from two separate runs of the Dorfman and Alf program, whereas those shown earlier in connection with Table 4 were estimated jointly from a single run of the Metz program. Thus, they are slightly different.

	With history						Without history						
	Rating						Rating						
	1	2	3	4	5	Z_1	1	2	3	4	5	Z_2	$Z_1 - Z_2$
ND	40	9	2	3	–	−1.9144	32	5	5	11	1	−1.0285	−0.8859
D	–	1	2	6	26		2	–	2	5	26		

(2) The new data sets are obtained by leaving out successive patients.

Delete nondiseased patient who received ratings of 2 with history and 5 without history.

						Z_1^i						Z_2^i	$Z_1^i - Z_2^i$
ND	40	8	2	3	–	−1.9128	32	5	5	11	–	−1.1623	−0.7505
D	–	1	2	6	26		2	–	2	5	26		

Delete nondiseased patients with ratings of 1 with history and 4 without history; there are five such patients.

Table 5. Computation of var$(Z_1 - Z_2)$ by Jackknifing

Type of patient deleted		Number of such patients	Jackknife estimates			
Status	Ratings		Z_1^i	Z_2^i	$Z_1^i - Z_2^i$	$[(Z_1 - Z_2) - (Z_1^i - Z_2^i)]^2 \times$ Number of patients
ND	2,5	1	−1.9128	−1.1623	−0.7505	0.01833316
ND	1,4	5	−1.9033	−1.0385	−0.8648	0.00222605
ND	2,4	4	−1.9128	−1.0385	−0.8743	0.00053824
ND	3,4	1	−2.0019	−1.0385	−0.9634	0.00600625
ND	4,4	1	−2.0661	−1.0385	−1.0276	0.02007889
ND	1,3	3	−1.9033	−1.0278	−0.8745	0.00038988
ND	2,3	1	−1.9128	−1.0278	−0.8850	0.00000081
ND	3,3	1	−2.0019	−1.0278	−0.9741	0.00777924
ND	1,2	1	−1.9033	−1.0254	−0.8779	0.00006400
ND	2,2	3	−1.9128	−1.0254	−0.8874	0.00000675
ND	4,2	1	−2.0661	−1.0254	−1.0407	0.02396304
ND	1,1	31	−1.9033	−1.0225	−0.8808	0.00080631
ND	4,1	1	−2.0661	−1.0225	−1.0436	0.02486929
D	4,5	2	−1.9008	−1.0090	−0.8918	0.00006962
D	5,5	24	−1.9018	−1.0090	−0.8928	0.00114264
D	4,4	3	−1.9008	−1.0548	−0.8460	0.00477603
D	5,4	2	−1.9018	−1.0548	−0.8470	0.00302642
D	3,3	2	−1.9782	−1.1027	−0.8755	0.00021632
D	2,1	1	−2.2965	−1.1677	−1.1288	0.05900041
D	4,1	1	−1.9008	−1.1677	−0.7331	0.02334784
						Σ = 0.19664119

						Z_1						Z_2	$Z_1 - Z_2$
ND	39	9	2	3	–	−1.9033	32	5	5	10	1	−1.0385	−0.8648
D	–	1	2	6	26		2	–	2	5	26		

Delete nondiseased patients with ratings of 2 with history and 4 without history; there are four such patients. Notice that the four Z_1 values for these last four patients are each equal to the Z_1 value for the very first patient. Similarly, the Z_2 values for these four are equal to the Z_2 values for the previous five patients.

ND	40	8	2	3	–	−1.9128	32	5	5	10	1	−1.0385	−0.8743
D	–	1	2	6	26		2	–	2	5	26		
			⋮						⋮				

Delete diseased patients with ratings of 4 with history and 5 without history; there are two such patients.

ND	40	9	2	3	–	−1.9008	32	5	5	11	1	−1.0090	−0.8918
D	–	1	2	5	26		2	–	2	5	25		

Delete diseased patients with ratings of five with history and five without history; there are 24 such patients.

ND	40	9	2	3	–	−1.9018	32	5	5	11	1	−1.0090	−0.8928
D	–	1	2	6	25		2	–	2	5	25		
			⋮						⋮				

Delete diseased patients with ratings of 4 with history and 1 without history (last entry in Table 4).

ND	40	9	2	3	–	−1.9018	32	5	5	11	1	−1.1677	−0.7341
D	–	1	2	5	26		1	–	2	5	26		

Computation of var$(Z_1 - Z_2)$ is shown in Table 5.

Acknowledgment

The authors are indebted to Dr. Charles E. Metz for his helpful comments.

References

1. Swets JA, Pickett RM: Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. New York: Academic, 1982
2. Swets JA: ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 14:109-121, 1979
3. Metz CE: Basic principles of ROC analysis. Semin Nucl Med 8:283-298, 1978
4. Dorfman DD, Alf E Jr: Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. Rating method data. J Math Psychol 6:487-496, 1969

5. Bamber D: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 12:387–415, 1975
6. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36, 1982
7. Hanley JA, McNeil BJ: Method for comparing the area under two ROC curves derived from the same cases. Radiology 148:839–843, 1983
8. Metz CE, Kronman HB: Statistical significance tests for binormal ROC curves. J Math Psychol 22:218–243, 1980
9. Metz CE, Kronman HB: A test for the statistical significance of differences between ROC curves. Inserm 88:647–660, 1979
10. Metz CE, Wang P-L, and Kronman HB: A new approach for testing the significance of differences between ROC curves measured from correlated data. In, Deconick F, ed: Information Processing in Medical Imaging VIII. The Hague: Martinus Nijhof, 1984
11. Fleiss J, Davis M: Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. Am J Epidemiol 115:841–845, 1982
12. Efron B: Bootstrap methods. Another look at the jackknife. Ann Stat 7:1–26, 1979
13. McNeil BJ, Sanders R, Alderson PO, et al: A prospective study of computed tomography, ultrasound and gallium imaging in patients with fever. Radiology 139:647–653, 1981
14. McNeil BJ, Hanley JA, Funkenstein HH, Wallman J: The use of paired ROC curves in studying the impact of history on radiography interpretation. CT of the head as a case study. Radiology 149:75–77, 1983