STATISTICS IN MEDICINE, VOL. 13, 499–508 (1994)

GENERAL METHODOLOGY I

ADVANCES IN STATISTICAL METHODOLOGY FOR THE EVALUATION OF DIAGNOSTIC AND LABORATORY TESTS

GREGORY CAMPBELL

Biometry and Field Studies Branch, National Institute of Neurological Disorders and Stroke, Federal Building, Room 7A08, Bethesda, MD 20892, U.S.A.

SUMMARY

The ROC plot is a useful tool in the evaluation of the performance of medical tests for separating two populations. For a two-state decision rule based on such a test, the ROC plot is the graph of all observed (1-specificity, sensitivity) pairs. Each point on this empirical plot can be represented by a 2×2 contingency table. The non-parametric statistics of Mann–Whitney and Kolmogorov–Smirnov can be immediately identified on this plot. Local non-parametric confidence interval procedures related to the theoretical ROC curve are briefly reviewed. For continuous data, two new simultaneous confidence regions associated with the ROC curve are presented, one based on Kolmogorov–Smirnov confidence bands for distribution functions and the other based on bootstrapping.

Two different tests on the same patients can be compared on the ROC scale. For continuous data, one important problem concerns the comparison of two ROC plots (as would arise from two correlated diagnostic tests on each patient) using a sup norm (this metric can detect differences that the ROC area cannot). The distribution of a statistic based on this norm is studied, using the bootstrap. A biomedical example illustrates the methodologies.

1. INTRODUCTION

There are many biomedical situations in which it is important to evaluate a diagnostic or laboratory test using data that are ordinal or continuously distributed. For two well-defined groups, diseased patients and non-diseased ('normal') subjects, let T denote a random variable for the outcome of a biomedical test. It is assumed that the identification of group membership here is made without error. Define a decision rule by t_0, a threshold value of T, such that if $T > t_0$ the person is classified as positive (diseased) and if $T \leqslant t_0$, the person is classified as negative ('normal'). For a given threshold, define specificity as the probability that a normal person is classified as normal (true negative) and sensitivity as the probability that a diseased person is classified as diseased (true positive). The theoretical receiver operating characteristic (ROC) curve is the continuous function of sensitivity versus (1-specificity) as the threshold t_0 ranges over all possible values. This curve is estimated using empirical (or sample) distribution functions, as illustrated in Section 2. The resulting ROC plots are useful in diagnostic and laboratory test evaluation, in imaging, and in the assessment of artificial neural network performance.

Of primary interest in this paper is statistical ROC methodology for the evaluation of the performance of continuous medical tests. Historically, the ROC methodology was first developed for so-called 'ratings' data, namely, discrete ordinal data with only a few categories.[1–4] As an example of such a scale, radiograms are often assessed on the scale 'definitely diseased', 'probably

CCC0277–6715/94/050499–10

diseased', 'possibly diseased', 'possibly non-diseased', 'probably non-diseased', and 'definitely non-diseased'. Albert and Harris use the continuous ROC methodology to evaluate logistic regression.[5] Zweig and Campbell[6] review the literature for continuous data with application to laboratory studies. It is important to note that in many instances laboratory investigators discretize continuous data into coarse categories in order to apply the parametric theory for ratings data. More recently, non-parametric ROC methods have been developed to analyse the underlying continuous data.

Examples to illustrate the usefulness of ROC methodology are plentiful. It has been utilized to compare tumour markers for breast cancer, to evaluate creatinine kinase for detecting acute myocardial infarction in patients presenting to the emergency room with chest pain, to compare laboratory blood tests for the screening of prostate cancer, to mention a few. These and other applications are referenced in two review papers.[6,7]

2. THE EMPIRICAL ROC PLOT

Let X denote the medical test T for the 'normal' population and Y the test for the diseased group, with cumulative distribution functions F and G, respectively. Sensitivity (SENS) at threshold t is $\bar{F}(t) = 1 - F(t)$ and 1 − specificity (SPEC) is $\bar{G}(t) = 1 - G(t)$. The theoretical ROC curve is a plot in the unit square of the function pair $(\bar{F}(t), \bar{G}(t))$ as t ranges over all possible values. For random samples of size m from the normals and n from the diseased groups, let F_m and G_n denote the empirical (or sample) distribution functions. Then the empirical ROC plot is merely a plot of the empirical pairs $(\bar{F}_m(t_i), \bar{G}_n(t_i))$ as t_i ranges over the observed test values. Adjacent points (with no ties between values in the two groups) are then connected by horizontal and vertical line segments to produce the staircase empirical ROC plot such as Figure 1. (The occasional diagonal line in the figure is the result of a tied value of total cholesterol between the individuals in the two groups.) Usually the thresholds are not indicated on this plot, although they could be. Each plotted point corresponds to the half-open interval of the form $[t_i, t_{i+1})$. Every line segment on the ROC plot can be associated with an observed t_i; if the observation is from a normal subject the line is horizontal and if from a diseased patient, vertical.

One virtue of the empirical ROC plot (and the true ROC curve) is that it does not depend on the scale of the measurement for the test. It is invariant with respect to monotone transformations such as the linear (with positive slope), logarithm and square root. In fact, the empirical ROC is non-parametric, depending only on the ranks of the observations in the combined sample. Of course, it is of value to retain the value of the threshold for each ROC point so that one can convert the chosen sensitivity and specificity into a rule on the original test scale or in the transformed one, but this is easily done.

The ROC plot is an easy visual way of conveying the test's behaviour over its entire range. In decision analysis, one usually applies known tradeoffs for misclassification and known prevalence to decide what rule to choose, that is, what point on the ROC plot minimizes expected loss. This rule may depend on the clinical setting. Hence there is an advantage in concisely conveying the entire performance of the test as a discriminator, especially if medical personnel choose different rules. Since the empirical ROC plot is a visual representation of all the data, analyses based on the entire ROC plot are used to assess the overall performance of diagnostic and laboratory tests. Because the empirical ROC plot depends only on the ranks, it is no surprise that non-parametric methods are inherently associated with this assessment. The area under the theoretical ROC curve is $P(X < Y)$. An unbiased estimate of $P(X < Y)$ is the area under the empirical ROC plot, which is the Mann–Whitney version of the two-sample rank-sum statistic of Wilcoxon.[8] Statistical inference based on the area is well known and straightforward.[9] The maximum horizontal

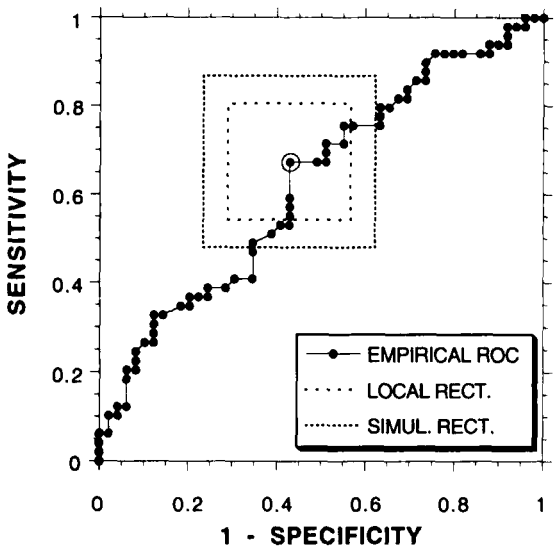

Figure 1. Graph of empirical ROC plot for total serum cholesterol in $m = 49$ normals and $n = 49$ individuals with coronary artery disease. For a threshold value for total cholesterol of 197, the smaller rectangle (- - -) corresponds to a local joint region for (1-specificity) and sensitivity and the larger rectangle (---) the simultaneous region at the same threshold at the identical confidence level $(0{\cdot}95)^2$

distance of the ROC plot from the diagonal line (SENS = 1 − SPEC) in the unit square is merely the value of the Kolmogorov–Smirnov (KS) two-sample test, a fact pointed out by Gail and Green[10] that will be exploited later.

3. LOCAL CONFIDENCE REGIONS ASSOCIATED WITH THE EMPIRICAL ROC PLOT

Confidence intervals can be constructed for sensitivity and for specificity at a particular threshold, t, using either the large-sample approximation to the binomial[7] or the exact binomial distribution.[11] (For ratings data, parametric models are used to produce confidence intervals.[12]) For a fixed threshold t, Hilgers[11] constructs a joint confidence region (the 'local' rectangle in Figure 1) centred about the observed ROC point $(\bar{F}_m(t), \bar{G}_n(t))$. This confidence interval is constructed from separate $(1 - \alpha)$ level confidence intervals for $\bar{F}(t)$ and for $\bar{G}(t)$. From the independence of the data on diseased and normal subjects, the joint region includes $(\bar{F}(t), \bar{G}(t))$ with desired confidence level $(1 - \alpha)^2$. For two adjacent test values t_i and t_{i+1}, note that $\bar{F}_m$ and $\bar{G}_n$ are constant on the half-open interval $[t_i, t_{i+1})$ so that the associated rectangle is a confidence region not just for the point t_i but for the interval. Greenhouse and Mantel[13] give a non-parametric large-sample confidence interval for sensitivity at the estimated threshold needed to obtain a desired specificity. It can be seen that their interval depends on the slope $\beta(t) = g(t)/f(t)$ of the theoretical ROC curve at estimated threshold t, where f and g denote the densities of F and G, respectively. The resulting confidence interval for $\bar{F}(t)$ reflects the fact that the threshold must also be estimated:

$$G(t)[1 - G(t)]/n + \beta^2(t)F(t)[1 - F(t)]/m .$$

For the usual case that $\beta(t)$ is unknown, Greenhouse and Mantel suggest estimation of $\beta(t)$ by smooth estimation of the densities at t.

4. SIMULTANEOUS JOINT CONFIDENCE REGIONS FOR SENSITIVITY AND (1-SPECIFICITY)

In this paper, simultaneous, rather than pointwise, confidence regions for the entire ROC curve are presented. Since the ROC plot depicts the entire performance of a test, and since many different uses can be made of the ROC plot, it is useful to develop simultaneous confidence regions associated with the entire ROC plot. In this section, a procedure is presented to obtain equal-sized rectangles in the unit square so that for each threshold, the rectangle centred at the observed (1-specificity, sensitivity) point contains the respective theoretical (1-specificity sensitivity) for that threshold, with confidence coefficient that is associated with all thresholds simultaneously. This is the global analogue of the local rectangle of Hilgers in the previous section.

The proposed region uses the separate confidence bands for the cumulative distribution functions F and G based on the distribution theory of Kolmogorov.[14] These bounds are based on the sup norm

$$\sup |F_m(t) - F(t)| ,$$

where the sup is over all t. Under the assumption that the distribution function is continuous, the exact and large-sample distributions of this sup norm are well known and tabulated.[14,9] To obtain a simultaneous confidence region for the entire ROC curve, form the Kolmogorov $(1 - \alpha)$ confidence band $(\bar{F}_m(t) - d,\ \bar{F}_m(t) + d))$ for $\bar{F}(t)$ and the $(1 - \alpha)$ confidence band $(\bar{G}_n(t) - e,\ \bar{G}_n(t) + e)$ for $\bar{G}(t)$. Then by independence of the two groups, $P\{\bar{F}_m(t) - d < \bar{F}(t) < \bar{F}_m(t) + d,\ \bar{G}_n(t) - e < \bar{G}(t) < \bar{G}_n(t) + e$ for all $t\} = (1 - \alpha)^2$. Thus, the collection of rectangles with width $2d$ and height $2e$ each centred at an observed point $(\bar{F}_m(t_i), \bar{G}_n(t_i))$ of the ROC plot has simultaneous coverage $(1 - \alpha)^2$ for the respective points $(\bar{F}(t_i), \bar{G}(t_i))$. Note that the coverage is valid not just at the observed t_i's but at all thresholds since the sensitivity and specificity are the same on the interval $[t_i, t_{i+1})$ as at t_i. Also, unlike the approach of Hilgers, all the rectangles are necessarily the same size.

As an example consider the data from a study of apolipoproteins for detecting coronary artery disease (CAD) in patients presenting in the emergency room with chest pain.[15] The samples consist of $m = 49$ individuals with no CAD and, for ease of presentation, only $n = 49$ of the original 255 patients with CAD. The rule that is employed is that large values of total cholesterol (TOT) indicate CAD. The ROC plot for TOT is presented in Figure 1. The 95 per cent Kolmogorov bounds for $\bar{F}$ and $\bar{G}$ give $d = e = 0{\cdot}194$. These are combined to form the simultaneous rectangles for all thresholds with confidence coefficient $(0{\cdot}95)^2 = 0{\cdot}9025$. In Figure 1, at the circled ROC point (3/7, 33/49), corresponding to the total cholesterol interval [197, 200), the rectangle is a square of half-width 0·194. By way of contrast, the size of the local rectangles of Hilgers varies according to the empirical ROC point, but, at this same point, the 90·25 per cent rectangle has half-width 0·138 and half-length 0·131. Thus the simultaneous confidence square has half-width roughly 40–50 per cent larger than the local rectangle at this point.

At first sight, one might think that this collection of rectangles is a confidence band for the entire ROC curve at the reported confidence level, but this is not the case. The band that would result from these overlapping rectangles for the total cholesterol example is depicted by the $d = 0{\cdot}194$ band in Figure 2. The confidence statement is merely that, for all thresholds simultaneously, the theoretical values of (1-specificity) and sensitivity are in their associated rectangles with coefficient $(0{\cdot}95)^2$. However, using a bootstrap mechanism explained in the next section, there is only one bootstrapped ROC plot out of 1000 that is not entirely within the $d = 0{\cdot}194$ band of Figure 2, for a bootstrapped coverage of 99·9 per cent. The important distinction is that in

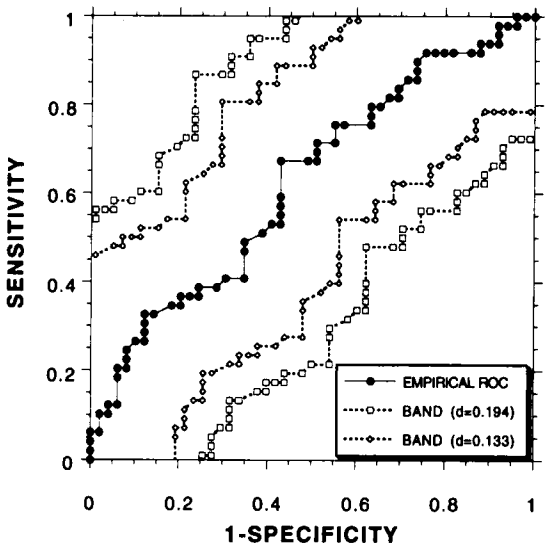

Figure 2. Fixed-width bands for the entire ROC curve. For the data in Figure 1, the wider band (open squares) is obtained from all the simultaneous rectangles as in Figure 1, corresponding to a fixed horizontal distance $d = 0{\cdot}194$ from the empirical ROC plot (solid circles). The narrower band (open diamonds) has fixed horizontal distance $d = 0{\cdot}133$

the case of the simultaneous rectangles of this section, it is not just that each theoretical ROC point is in some rectangular region and is hence covered, but that it is in the rectangle that corresponds to its threshold.

5. A FIXED-WIDTH CONFIDENCE BAND FOR THE ENTIRE ROC CURVE BASED ON THE BOOTSTRAP

In this section a fixed-width confidence band for the entire theoretical ROC curve is developed. Any type of region that covers the theoretical ROC curve with high probability could be used. One approach is to form vertical (or horizontal) bands about the empirical ROC plot. However, a desirable feature of the region is that it be invariant if the tags of group membership are interchanged. This suggests using a band at an angle that allows for the variability of the two groups. Since the variance of $F_m(t)$ is $F(t)[1 - F(t)]/m$ and that if $G_n(t)$ is $G(t)[1 - G(t)]/n$, the ratio of the standard deviations associated, respectively, with vertical and horizontal distance is $\sqrt{[G(t)[1 - G(t)]m/\{nF(t)[1 - F(t)]\}]}$. For $0{\cdot}3 < F(t), G(t), < 0{\cdot}7$, this is approximately $\sqrt{(m/n)}$. Therefore bands of fixed width are constructed by displacing the empirical ROC plot 'northwest' and 'southeast' along lines with slope $b = -\sqrt{(m/n)}$. To determine how much displacement there should be to obtain the desired coverage, bootstrap resampling[16,17] is used.

The bootstrap has been employed in ROC analysis by others. Moise *et al.* used it to investigate the crossing of two ROC curves[18] while Linnet has employed it to reduce bias.[19] Campbell *et al.*[20] used it for statistical inference for ROC plots with fuzzy or probabilistic assignment of membership in the two classes, normal and diseased. The bootstrap will be used here to estimate the coverage for a known band about the ROC plot as well as to generate the sampling distribution associated with a distance metric from the observed ROC plot.

Consider first the use of the bootstrap to estimate the coverage probability for a fixed region. Here the fundamental application of the bootstrap is to sample with replacement separately from

F_m and G_n and to form the empirical ROC plot for each bootstrap sample. More specifically, take a sample $X_1^*, \ldots, X_m^*$ with replacement from the observations $X_1, \ldots, X_m$, and calculate the empirical function for (1-specificity) given by $\bar{F}_m^*$. Generate an independent sample $Y_1^*, \ldots, Y_n^*$ with replacement from the observations $Y_1, \ldots, Y_n$ and calculate $\bar{G}_n^*$. Then the empirical ROC plot based on this single bootstrap is the plot of the points $(\bar{F}_m^*, \bar{G}_n^*)$. One estimates the coverage of a given region by counting the proportion of such bootstrap replications that are completely covered by the region. By adjusting the width of this region about the original ROC empirical plot, one can use this procedure to obtain desired coverage. The statistical device of the bootstrap works for large samples because, suitably normed, the bootstrap distribution of F_m^* and the sampling distribution of F_m converge to the same limiting distribution,[21] although for moderate samples the bootstrap confidence band for a continuous distribution function is slightly narrower than the exact theory predicts.[22] (Note that the bootstrap used here is different from that used for the null distribution, under F and G identical, of the Kolmogorov–Smirnov (KS) statistic. To bootstrap the null distribution of the KS test, the two samples are combined, the tags indicating group membership are ignored and the bootstrap is then used to approximate the null distribution of the test statistic.[23])

This procedure is used to construct a confidence region for the total cholesterol example. To simplify the calculations, equal numbers of diseased and normal subjects were used. In this case the ROC plot is displaced in either direction along the lines with slope -1. To determine the required magnitude of this displacement, the maximum distance along these lines between the bootstrapped ROC and the original empirical (non-bootstrapped) ROC plot is measured. Then the bootstrap sampling distribution of this distance measure can be reported and simultaneous bands chosen accordingly. A displacement of magnitude $d\sqrt{2}$ along lines with slope -1 is equivalent to a horizontal displacement of magnitude d. Based on the cholesterol data, the distribution of the maximal side length d is approximated using 1000 bootstraps, where all bootstraps are performed using IMSL random number generators on a mainframe computer:

d	1/49	2/49	3/49	4/49	5/49	6/49	7/49	8/49	9/49	10/49
bootstraps	0	26	199	318	252	108	68	25	3	1

Thus a fixed band width region with $d = e = 6{\cdot}5/49 = 0{\cdot}133$ indicated in Figure 2 has bootstrap coverage 0·903.

6. INFERENCE FOR PAIRED ROC PLOTS BASED ON THE SUP NORM

An important problem concerns the comparison of two (or more) diagnostic tests in the two defined groups. ROC methodology permits a comprehensive graphical and quantitative comparison. Note that such a comparison is inherently non-parametric if the two tests are measured on different scales. Although the two tests may be studied in independent samples, in this section the case in which both tests are studied in the same subjects is analysed. An example of ROC plots obtained in this way (Figure 3) gives rise to the question 'Is there evidence that the two tests have the same ROC curves, even though the scales of the two tests differ?'. In answering this question, it is important to recognize that diagnostic or laboratory tests on the same subjects are usually correlated, and that comparisons of these two ROC plots will need to take such correlations into account.

Comparison of correlated diagnostic tests using areas under the ROC plot has been examined independently by a number of authors who have considered non-parametric inference based on

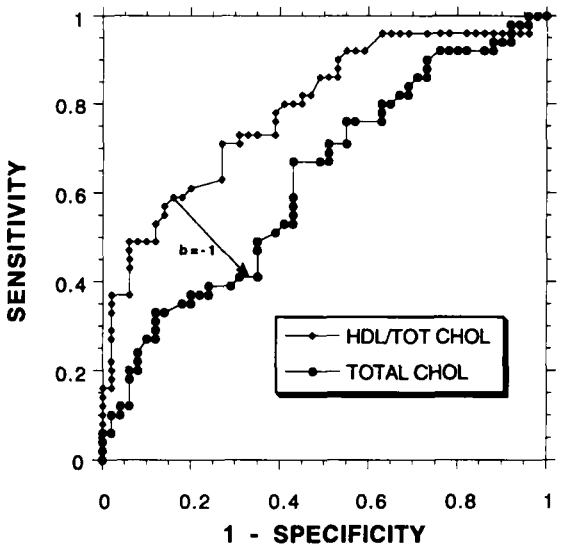

Figure 3. ROC plots for a pair of tests on the same subjects. For the same individuals as in Figure 1, ROC plots of total cholesterol (solid circles) and of the ratio of high density lipids to total cholesterol (solid triangles) are graphed. The arrow indicates the point of maximal separation between the two plots along a line with slope $b = -1$

the estimation of the correlation of the Mann–Whitney ROC areas.[24,25] In the case of ratings data, there are procedures based on maximum likelihood[26] and on an approximation,[27] but all these area approaches may be unsatisfactory for two ROC plots that are very different yet have similar areas. Wieand *et al.*[28] developed methods for focusing on only selected portions of the ROC curve, such as regions corresponding to high specificity. Here an alternative approach based on the sup norm is considered. The sup norm approach would be advantageous in circumstances where the ROC plots cross or where area does not seem to adequately reflect the situation, in much the same way that for two-sample problem that there are situations in which the Kolmogorov–Smirnov test is more powerful than the Wilcoxon.

The test is as follows. First measure how far the two correlated ROC plots are from each other using the maximal distances between the two ROC plots along lines with slope $b = -\sqrt{(m/n)}$. The empirical ROC plot is unchanged if the observations are replaced by their ranks in the combined samples. For the ith pair from the normal subjects, let R_{ki} denote the rank of the ith score on the kth test among the combined $(m + n)$ values of the kth test, for $i = 1$ to m and $k = 1, 2$; for the jth pair from the diseased patients, let S_{kj} denote the rank of the jth score on the kth test among all $(m + n)$ values of the kth test, for $j = 1$ to n and $k = 1, 2$. To test the null hypothesis that the two theoretical ROCs have the same functional representation in the unit square, the rank pairs are bootstrapped in such a way that preserves the correlation. This is done as follows. Randomly select with replacement a sample (R^*_{1i}, R^*_{2i}) of size m from the m normal test rank pairs (R_{1i}, R_{2i}). For each i of the m sample pairs (R^*_{1i}, R^*_{2i}), with probability 0·5 interchange the order to (R^*_{2i}, R^*_{1i}). Generate a sample (S^*_{1j}, S^*_{2j}) of size n with replacement from the n rank pairs (S_{1j}, S_{2j}) and for each sample pair with probability 0·5 interchange the order to (S^*_{2j}, S^*_{1j}). (This random interchange preserves the correlational structure and tends to equalize the ROC functions in the unit square. In particular, this procedure makes the areas under the ROC plots approximately equal, as has been verified in subsequent bootstrap computer simulations. Note that because the

ranks are exchanged with probability 0·5, rather than the original observations, it is not necessary to assume that $F_1 = F_2$ and $G_1 = G_2$ but only that the two theoretical ROC functions are equal.) For each bootstrap collection of m pairs and n pairs, calculate the two empirical ROC plots (one for each test) and measure how far apart they are in the sup norm described above. The inference is now simple: generate the approximate sampling distribution for this metric by bootstrapping a large number of times. To estimate the P-value merely record the percentage of bootstrap empirical ROC plots that have distance at least as large as that actually observed.

As an example, two tests for the $m = 49$, $n = 49$ patients in the CAD study are compared: total cholesterol (TOT) and the high density lipid ratio HDL/TOT. Here the decision rule is that small values of HDL/TOT indicate CAD. The empirical ROC plots are in Figure 3. The distance measure between the two empirical ROC plots using the side of the square along lines with slope of -1 (as indicated by the arrow in Figure 3) is 9/49. Out of 1000 bootstraps, the average areas of the ROC plots for the two tests were 0·705 and 0·703 and the average Spearman correlations of the two tests were 0·32 for the normals and 0·48 for the diseased patients. The bootstrap produced the desired effect, namely preserving the correlations but equalizing the ROC areas. Of the 1000 bootstraps, 24 had distances at least as large, for an estimated P-value of 0·024. (The non-parametric analysis based on areas[26] gives an estimated correlation of the areas of 0·43, ROC areas of 0·6256 and 0·7884 and associated estimated variances of 0·0030 and 0·0020 for total and HDL/TOT, respectively. The test statistic is $z = 3·04$, with P-value = 0·003.)

7. DISCUSSION AND CONCLUSIONS

Presented here are two simultaneous confidence procedures based on the ROC plot. Although the confidence regions presented are two-sided, it is straightforward to generate one-sided versions as well. For the rectangular regions using the Kolmogorov bounds of Section 4, the coverage probabilities for F and G were identical. It is a simple generalization to use coverages $(1 - \alpha_1)$ for F and $(1 - \alpha_2)$ for G to get joint rectangular regions that have simultaneous confidence $(1 - \alpha_1)(1 - \alpha_2)$. The bootstrap approach has a great deal of appeal but is computer intensive, especially for large sample sizes.

These approaches assume that the data in the two groups are continuous. If that is not the case and there are a few ties, this is not critical for either approach because it has been demonstrated by others in one dimension that the Kolmogorov bounds and the bootstrap bounds for the separate distribution functions are conservative.[14,22] However, if the number of ties between the two groups is large as usually occurs with ratings data, then caution must be exercised.

Consider the $(1 - \alpha)$ confidence bands for the entire ROC curve for the special case in which $F(t) = G(t)$ for all t (that is, $F \equiv G$). The maximal horizontal (or vertical) distance of the empirical ROC plot to the diagonal line $y = x$ is just the Kolmogorov–Smirnov two-sample statistic $\max |F_m(t_i) - G_n(t_i)|$. The large-sample distribution of this maximal horizontal distance under the assumption that $F \equiv G$ is approximated by $q_\alpha\sqrt{(2/m)}$, where q_α is the upper quantile of the limiting distribution of the supremum of $\sqrt{m}\,|F_m(t) - F(t)|$. This corresponds to a band of $d = e$ of half that value; that is, $d = e = q_\alpha/\sqrt{(2m)}$. For $m = n = 49$ and $\alpha = 0·1$, this yields $d = 0·123$, in close agreement with $d = 0·133$ for the band with bootstrapped coverage 90·2 per cent in the example. It is conjectured that such bands have approximate coverage $(1 - \alpha)$ even if F and G are not identical. The simulations below for $m = n = 10$ tend to bear out this out. ROC functions of the form $y = x^{1/k}$ are considered, for $k = 1, 3, 6, 10$. For each of 1000 different simulations for each

value of k, the number of bands of half-width $d = e = c/20$ that do not contain the true theoretical ROC curve are recorded:

	c				
k	5	6	7	8	9
1	164	58	16	3	0
3	206	62	16	6	1
6	138	55	16	7	2
10	101	55	16	10	2
KS2	167·8	52·4	12·3	2·1	0·2

The last line of the table (KS2) is the expected number (out of 1000 simulations) whose exact null two-sample Kolmogorov–Smirnov statistic exceeds $c/10$. Note that, as should be the case, the bootstraps from $k = 1$ in the table are in agreement with these values. The entries for other values of k in the table are reasonably close as well.

The bootstrap hypothesis test in Section 6 can be extended. For two independent tests, with sample sizes m_1 and n_1 for the first test and m_2 and n_2 for the second test, the distance between two ROCs would use the measure above along lines with slope $b = -[\sqrt{m_1} + \sqrt{m_2}]/[\sqrt{n_1} + \sqrt{n_2}]$. In the dependent case, as above for the same subjects for both tests, in order to adjust explicitly for the correlation, recall that the grade correlation ρ_s is estimated by Spearman's rank correlation coefficient r_s:[29]

$$\rho_s = \mathrm{corr}(F_1(X_1), F_2(X_2)); \quad r_s = r(F_{1m}(X_{1i}), F_{2m}(X_{2i})) ,$$

where (X_1, X_2) has a distribution with marginal F_1 for X_1 and F_2 for X_2 and r is Pearson's correlation. For a value s of the first test with t of the second test such that $F_1(s) = F_2(t)$ and $G_1(s) = G_2(t)$, the distances $[F_{1m}(s) - F_{2m}(t)]$ and $[G_{1n}(s) - G_{2n}(t)]$ have approximate variances $2F_2(t)[1 - F_2(t)]\,(1 - \rho_{sN})/m$ and $2G_2(t)[1 - G_2(t)](1 - \rho_{sD})/n$. Thus, away from the borders of the unit square, measure vertical and horizontal distances along lines with slope $b = -\sqrt{\{[m(1 - r_{sD})]/[n(1 - r_{sN})]\}}$, where r_{sD} and r_{sN} are the Spearman correlations for the pair of tests for the diseased and normal groups, respectively. If $r_{sD} = r_{sN}$ this reduces to the same slope as in the uncorrelated case. In the above example the observed Spearman correlations are 0·35 for the normals and 0·51 for those with coronary artery disease (CAD), resulting in a line with slope $-0{\cdot}87$ and a distance measure corresponding to a horizontal distance of 9/49 and a vertical of 8/49. The associated estimated P-value is 0·039; it is no surprise that this value is somewhat larger than that of Section 6 since all squares with half-width of 9/49 are counted in this latter procedure. The effect of using the metric that incorporates the unequal correlations can lead to a different inference in some cases.

ACKNOWLEDGEMENTS

The author gratefully acknowledges a referee and an associated editor whose careful reading led to an improved paper.

REFERENCES

1. Swets, J. A. 'Measuring the accuracy of diagnostic systems', *Science*, **240**, 1285–1293 (1988).
2. Green, D. M. and Swets, J. A. *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966.
3. Swets, J. A. and Pickett, R. M. *Evaluation of Diagnostic Systems*, Academic Press, New York, 1982.

4. Metz, C. E. 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, **8**, 283–298 (1978).
5. Albert, A. and Harris, E. K. *Multivariate Interpretation of Clinical Laboratory Data*, Marcel Dekker, New York, 1987.
6. Zweig, M. and Campbell, G. 'Receiver operating characteristic (ROC) curves: A fundamental evaluation tool in clinical medicine', *Clinical Chemistry*, **39**, 561–577 (1993).
7. Beck, J. R. and Shultz, E. K. 'The use of relative operating characteristic (ROC) curves in test performance evaluation', *Archives of Pathology and Laboratory Medicine*, **110**, 13–20 (1986).
8. Bamber, D. 'The area above the ordinal dominance graph and the area below the receiver operating characteristic curve', *Journal of Mathematical Psychology*, **12**, 387–415 (1975).
9. Hollander, M. and Wolfe, D. A. *Nonparametric Statistical Methods*, Wiley, New York, 1973.
10. Gail, M. H. and Green, S. B. 'A generalization of the one-sided two-sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests', *Biometrics*, **32**, 561–570 (1976).
11. Hilgers, R. A. 'Distribution-free confidence bounds for ROC curves', *Methods of Information in Medicine*, **30**, 96–101 (1991).
12. McNeil, B. J. and Hanley, J. A. 'Statistical approaches to the analysis of receiver operating characteristic (ROC) curves', *Medical Decision Making*, **2**, 137–150 (1984).
13. Greenhouse, S. W. and Mantel, N. 'The evaluation of diagnostic tests', *Biometrics*, **6**, 399–412 (1950).
14. Conover, W. J. *Practical Nonparametric Statistics*, Second Edition, Wiley, New York, 1980.
15. Kottke, B. A., Zinsmeister, A. R., Holmes, Jr. D. R., Kneller, R. W., Hallaway, B. J. and Mao, S. J. T. 'Apolipoproteins and coronary artery disease', *Mayo Clinic Proceedings*, **61**, 313–320 (1986).
16. Efron, B. 'Bootstrap methods: another look at the jackknife', *Annals of Statistics*, **7**, 1–26 (1979).
17. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
18. Moise, A., Clement, B., Ducimetiere, P. and Bourassa, M. G. 'Comparison of receiver operating curves derived from the same population: A bootstrapping approach', *Computers and Biomedical Research*, **18**, 125–131 (1985).
19. Linnet, K. 'Assessing diagnostic tests by a strictly proper scoring rule', *Statistics in Medicine*, **8**, 609–618 (1989).
20. Campbell, G., Levy, D. and Bailey, J. J. 'Bootstrap comparison of fuzzy R.O.C. curves for ECG-LVH algorithms using data from the Framingham heart study', *Journal of Electrocardiology*, **23**(suppl), 132–137 (1990).
21. Bickel, P. J. and Freidman, D. A. 'Some asymptotic theory for the bootstrap', *Annals of Statistics*, **9**, 1196–1217 (1981).
22. Bickel, P. J. and Krieger, A. M. 'Confidence bands for a distribution function using the bootstrap', *Journal of the American Statistical Association*, **84**, 95–100 (1989).
23. Romano, J. P. 'A bootstrap revival of some nonparametric distance tests', *Journal of the American Statistical Association*, **83**, 698–708 (1988).
24. DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. 'Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach', *Biometrics*, **44**, 837–845 (1988).
25. Campbell, G., Douglas, M. A. and Bailey, J. J. 'Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve', *in* Ripley, K. L. and Murray, A. (eds.), *Computers in Cardiology*, IEEE Computer Society, Washington, D.C., 1989, pp. 267–270.
26. Metz, E., Wang, P. -L., and Kronman, H. B. 'A new approach for testing the significance of differences between ROC curves measured from correlated data', *in* Deconinck, F. (ed.), *Information Processing in Medical Imaging: Proceedings of the Eighth Conference*, Martinus Nyhoff, The Hague, 1984, pp. 432–445.
27. Hanley, J. A. and McNeil, B. J. 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases', *Radiology*, **148**, 839–843 (1983).
28. Wieand, S., Gail, M. H., James, B. R. and James, K. L. 'A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data', *Biometrika*, **76**, 585–592 (1989).
29. Gibbons, J. D. and Chakraborti, S. *Nonparametric Statistical Inference*, Third Edition, Marcel Dekker, New York, 1992.