



A Permutation Test to Compare Receiver Operating Characteristic Curves

Author(s): E. S. Venkatraman

Source: *Biometrics*, Vol. 56, No. 4, (Dec., 2000), pp. 1134-1138

Published by: International Biometric Society

Stable URL: <http://www.jstor.org/stable/2677047>

Accessed: 07/07/2008 06:33

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=ibs>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

A Permutation Test to Compare Receiver Operating Characteristic Curves

E. S. Venkatraman

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center,
1275 York Avenue, New York, New York 10021, U.S.A.
e-mail: venkat@biosta.mskcc.org

SUMMARY. We developed a permutation test in our earlier paper (Venkatraman and Begg, 1996, *Biometrika* **83**, 835–848) to test the equality of receiver operating characteristic curves based on continuous paired data. Here we extend the underlying concepts to develop a permutation test for continuous unpaired data, and we study its properties through simulations.

KEY WORDS: Diagnostic test; Permutation test; Receiver operating characteristic curves.

1. Introduction

In clinical research, we are often faced with a classification problem where the outcome is binary. For example, we may be interested in a new continuously valued marker for diagnosing a disease or a model for predicting the response to a new treatment. The accuracy of a classification rule in such a problem is summarized by the two misclassification rates, which vary with the threshold level used for the rule. Receiver operating characteristic (ROC) curves, where the true-positive rates are plotted against the false-positive rate for all classification points, are a popular method for displaying the misclassification rates in these problems (Swets and Pickett, 1982). The ROC curves for two diagnostic tools or models for classification are identical if and only if, for every classification point from one method, there is an equivalent one from the other with the same misclassification rates. Thus, the problem of comparing alternate diagnostic markers for diagnosing a disease or alternate prediction models for response can be reduced to one of comparing ROC curves.

There are two commonly used approaches to comparing ROC curves. One is to fit a parametric model to the data, such as the binormal model of Dorfman and Alf (1969), and test the equality of the parameters (Metz, Wang, and Kronman, 1984). The second is to test the equality of a summary measure from the ROC curves, such as the area under the curve, obtained either from a parametric model or nonparametrically. The nonparametric version of the area test was developed by Hanley and McNeil (1982, 1983) for both unpaired and paired data. The test was refined by DeLong, DeLong, and Clarke-Pearson (1988) with their derivation of a jackknife estimate of the variance for the area under the ROC curve of one or more diagnostic markers from the same subjects.

In our earlier paper (Venkatraman and Begg, 1996), we developed a fully nonparametric test to compare two ROC

curves when the data are paired and continuous. This test evaluates the equality of the curves at all operating points with an appropriate test statistic, whose reference distribution is generated by permuting the pooled ranks of the test scores within a subject. We showed through Monte Carlo simulations that the new test is comparable to a nonparametric test of the equality of the two areas under the ROC curves when one marker is uniformly better than the other. More importantly, the new test is capable of distinguishing ROC curves when they cross but have equal areas. In this article, we extend the permutation test presented in Venkatraman and Begg (1996) to the case of continuous unpaired data, study its properties via simulations, and give an example illustrating its use.

2. Methods

Let D be a binary response variable of interest and let X and Y represent the two independent variables that are the predictors of D . For simplicity, in this section, we will let X and Y be the two markers used to diagnose the disease status D of a subject, where $D = 1$ represents diseased and $D = 0$ normal. Let the distributions of the two markers conditioned on the response D be as follows:

$$\begin{aligned} F_x(x) &= P(X \leq x \mid D = 1), \\ F_y(y) &= P(Y \leq y \mid D = 1), \\ G_x(x) &= P(X \leq x \mid D = 0), \\ G_y(y) &= P(Y \leq y \mid D = 0). \end{aligned}$$

These distribution functions give us the false-negative rates (F) and true negative rates (G) for the markers X and Y ; the ROC curve is a plot of $1 - F$ against $1 - G$ over all classification points. The ROC curves for X and Y are identical if and only if, for every classification point x of X , there exists a y for Y such that the true- and false-positive rates are equal, i.e., $F_x(x) = F_y(y)$ and $G_x(x) = G_y(y)$.

The calibration of the classification points for comparison of the two ROC curves under the null hypothesis can be attained as follows. Let the mixture distributions $M_{\kappa x}$ and $M_{\kappa y}$ be defined as

$$M_{\kappa x}(x) = \kappa F_x(x) + (1 - \kappa)G_x(x)$$

and

$$M_{\kappa y}(y) = \kappa F_y(y) + (1 - \kappa)G_y(y),$$

where $0 < \kappa < 1$ is the mixing proportion. If the ROC curves are equal, then $M_{\kappa x}(x) = M_{\kappa y}(y)$ for the calibrated classification points x and y . Thus, if we denote the common value by p , then the calibration is given by $M_{\kappa x}^{-1}(p)$ and $M_{\kappa y}^{-1}(p)$ for some $0 < p < 1$. If the κ used in the mixture is the true probability of disease, then these mixture distributions are the unconditional distributions of the markers in the population. For notational simplicity, we drop the κ from the subscripts.

The misclassification probability in classifying a randomly chosen subject is the weighted average of the false-positive and false-negative rates. Let $x_p = M_x^{-1}(p)$ and $y_p = M_y^{-1}(p)$ be the calibrated classification points. The corresponding misclassification probabilities $e_x(p)$ and $e_y(p)$ are given by

$$\begin{aligned} e_x(p) &= \kappa \times F_x(x_p) + (1 - \kappa) \times \{1 - G_x(x_p)\}, \\ e_y(p) &= \kappa \times F_y(y_p) + (1 - \kappa) \times \{1 - G_y(y_p)\}. \end{aligned} \quad (1)$$

Observe that $e_x(p) - e_y(p) = \kappa\{F_x(x_p) - F_y(y_p)\} + (1 - \kappa)\{G_y(y_p) - G_x(x_p)\}$ and is identically zero for all p and for any $0 < \kappa < 1$ if and only if the ROC curves are equal. Thus, testing the equality of the ROC curves can be reduced to testing the hypothesis that the parameter $\theta = \int |e_x(p) - e_y(p)|dp$ is zero. Observe also that a graph of $\kappa - e$ against $1 - M$, where e and M are as defined earlier, is a rotation of the ROC curve using the transformation $(u, v) \rightarrow \{(1 - \kappa)u + \kappa v, -(1 - \kappa)u + \kappa v\}$. Hence, the parameter θ is proportional to the unsigned area between the two ROC curves. We now derive an estimate of θ from the sample and its reference distribution under the null hypothesis.

2.1 Test Statistic

In this section, we describe how to estimate the parameter θ . Let $\{(X_i, D_{xi}); i = 1, \dots, n_0 + n_1\}$ and $\{(Y_j, D_{yj}); j = 1, \dots, m_0 + m_1\}$ be the data for the two ROC curves, where n_0 and m_0 are the numbers of normal subjects in the data set for X and Y and where n_1 and m_1 the numbers of diseased subjects. Let $\hat{F}_x, \hat{F}_y, \hat{G}_x$, and \hat{G}_y be empirical estimates of the conditional distributions of the markers X and Y .

We see that the misclassification rates e_x and e_y can be estimated by substituting the empirical distributions \hat{F} and \hat{G} for F and G in equation (1). However, in order obtain the classification points x_p and y_p , we need to know the distribution functions M_x and M_y . Since these functions are unknown, we approximate them using their empirical values given by

$$\hat{M}_x = \kappa \hat{F}_x + (1 - \kappa) \hat{G}_x \quad \text{and} \quad \hat{M}_y = \kappa \hat{F}_y + (1 - \kappa) \hat{G}_y. \quad (2)$$

Notice that the error rates e_x and e_y are weighted averages of the false-positive and false-negative rates and are not the same as the proportion of misclassified subjects in the sample. The two are equal only if the sample prevalence of disease

$n_1/(n_0 + n_1)$ and $m_1/(m_0 + m_1)$ are equal and if κ is set to be the common value. The exact computation of an estimate \hat{e}_x of e_x is given below.

Let $N = n_0 + n_1$ be the total sample size for the first ROC curve and let $x_1 < \dots < x_n$ be the distinct marker values in the sample. The empirical distributions at x_i are given by

$$\hat{F}_x(x_i) = n_1^{-1} \sum_{j=1}^N \mathcal{I}(X_j \leq x_i) \times \mathcal{I}(D_j = 1)$$

and

$$\hat{G}_x(x_i) = n_0^{-1} \sum_{j=1}^N \mathcal{I}(X_j \leq x_i) \times \mathcal{I}(D_j = 0)$$

for $i = 1, \dots, n$, where \mathcal{I} is an indicator function. Let us denote $M_x(x_i)$ by p_i . Then p_i and $e_x(p_i)$ are approximated by

$$\hat{p}_i = \kappa \hat{F}_x(x_i) + (1 - \kappa) \hat{G}_x(x_i)$$

and

$$\hat{e}_x(\hat{p}_i) = \kappa \hat{F}_x(x_i) + (1 - \kappa) \{1 - \hat{G}_x(x_i)\}.$$

Since $e_x(0) = 1 - \kappa$, we can set \hat{p}_0 and $\hat{e}_x(\hat{p}_0)$ to be zero and $1 - \kappa$, respectively. Observe also that $\hat{p}_n = 1$ and $\hat{e}_x(\hat{p}_n) = \kappa$. Finally, we estimate the function $e_x(\cdot)$ by joining the points $(\hat{p}_i, \hat{e}_x(\hat{p}_i)), i = 0, \dots, n$, by straight lines, giving us a continuous, piecewise linear curve. The misclassification rate function e_y for marker Y is estimated similarly. Finally, the estimate $\hat{\theta}$ of the parameter θ is calculated as $\int |\hat{e}_x(p) - \hat{e}_y(p)|dp$.

2.2 Permutation Reference Distribution

In the previous section, we demonstrated how we can obtain a statistic to test the equality of the ROC curves. We now describe the steps needed to obtain a reference distribution to conduct the test. Observe that the conditional distributions of the random variable $M_x(X)$ given $D = 1$ and $D = 0$ are $F_x M_x^{-1}(\cdot)$ and $G_x M_x^{-1}(\cdot)$, respectively, and those of $M_y(Y)$ are $F_y M_y^{-1}(\cdot)$ and $G_y M_y^{-1}(\cdot)$. Since G and F are the true- and false-negative rates, it follows that, under the null hypothesis of equal ROC curves, the conditional distributions of $M_x(X)$ and $M_y(Y)$ are equal. We can use this equality of distributions to generate a reference distribution by permuting the data as follows.

An ROC curve is invariant under any monotonically increasing transformation of the marker. Thus, the statistic $\hat{\theta}$ defined in the previous section would be unchanged if we transformed the markers X and Y using M_x and M_y , respectively. Let us denote the transformed data $M_x(X)$ for normal and diseased subjects by $\{U_{xi}, i = 1, \dots, n_0\}$ and $\{V_{xi}, i = 1, \dots, n_1\}$, respectively. Similarly, let $\{U_{yi}, i = 1, \dots, m_0\}$ and $\{V_{yi}, i = 1, \dots, m_1\}$ be the transformed marker $M_y(Y)$ for the normal and diseased subjects. Then the variables U and V are mutually independent and, under the null, $U_x \stackrel{L}{=} U_y$ and $V_x \stackrel{L}{=} V_y$. Let us denote by $\Theta\{(\mathbf{U}_x, \mathbf{V}_x), (\mathbf{U}_y, \mathbf{V}_y)\}$ the estimator described in Section 2.1 that gives θ for normal and diseased marker values given by \mathbf{U}_x and \mathbf{V}_x for marker X and \mathbf{U}_y and \mathbf{V}_y for marker Y .

The permuted marker data \mathbf{U}_x^* and \mathbf{U}_y^* for the normal subjects are obtained by pooling the U_{xi} and U_{yi} and assigning n_0 of them sampled without replacement to the x group. Similarly, let \mathbf{V}_x^* and \mathbf{V}_y^* denote the permuted marker data for the diseased subjects. Let $\hat{\theta}^* = \Theta\{(\mathbf{U}_x^*, \mathbf{V}_x^*), (\mathbf{U}_y^*, \mathbf{V}_y^*)\}$ be the parameter estimate for the permuted data. Since, under the null hypothesis, the distributions of the U and the V are identical, the distributions of $\hat{\theta}$ and $\hat{\theta}^*$ are identical. Thus, a reference distribution for $\hat{\theta}$ is given by the permutation distribution of $\hat{\theta}^*$, which is obtained either by complete enumeration of all permutations or by sampling a large number of permutations.

If the transformations M_x and M_y are the same, then the conditional distributions of the markers are identical in their original scale. Thus, the reference distribution can be obtained by permuting the marker data directly instead of requiring they be transformed. However, it is very unlikely that the conditional distributions of the markers are identical or that we know the transformations M_x and M_y . Thus, we need to approximate the transformations to obtain a reference distribution. Earlier, for the calculation of the test statistic, we used the empirical values of M_x and M_y given by equation (2). We use the same functions to approximate U and V to obtain the permutation distribution. For normal subjects, let $\hat{U}_x = \hat{M}_x(X)$ and $\hat{U}_y = \hat{M}_y(Y)$, and for diseased subjects, let $\hat{V}_x = \hat{M}_x(X)$ and $\hat{V}_y = \hat{M}_y(Y)$. Let $\hat{\mathbf{U}}_x, \hat{\mathbf{V}}_x, \hat{\mathbf{U}}_y$ and $\hat{\mathbf{V}}_y$ be the approximate transformed data and $\hat{\mathbf{U}}_x^*, \hat{\mathbf{V}}_x^*, \hat{\mathbf{U}}_y^*$ and $\hat{\mathbf{V}}_y^*$ be the permuted data. Then $\hat{\theta} = \Theta\{(\hat{\mathbf{U}}_x, \hat{\mathbf{V}}_x), (\hat{\mathbf{U}}_y, \hat{\mathbf{V}}_y)\}$ and $\hat{\theta}^* = \Theta\{(\hat{\mathbf{U}}_x^*, \hat{\mathbf{V}}_x^*), (\hat{\mathbf{U}}_y^*, \hat{\mathbf{V}}_y^*)\}$ are the observed and permuted data estimated of the parameter θ . The p -value of the test is given by the ratio of the number of times $\hat{\theta}^* \geq \hat{\theta}$ to the number of permutations.

Observe that both the test statistic and the reference distribution described above are functions of the mixing constant κ . While any choice of κ such that $0 < \kappa < 1$ provides a valid test, a natural question to ask is how the choice of κ affects the power of the test. In our earlier paper, because of the paired nature of the data, the sample prevalence of the disease for both the markers is the same (Venkatraman and Begg, 1996). Hence, we chose κ to be the sample prevalence rate of disease, i.e., $\kappa = n_1/(n_0 + n_1)$. This had the advantage of reducing \hat{M} to the empirical distributions of the sample and thus we could calculate the test statistic and its reference distributions using the ranks of the markers. Since κ is used primarily to facilitate the comparison of X and Y marker data, we expect the choice of κ to have negligible impact on the power of the test. We recommend that the pooled estimate of prevalence $(n_1 + m_1)/(n_0 + n_1 + m_0 + m_1)$ be chosen as the value of κ .

Observe also that the nature of the permutation for unpaired data is different from that in the paired data case. For paired data, while the marginal distributions of the markers are identical, they are not independent. Thus, in our earlier paper, we assumed that the marker pairs are exchangeable in the transformed scale to ensure that the permutation of the marker data within a subject is valid. In the unpaired data case, independence ensures the validity of the permutation of transformed data.

In the following section, we evaluate the performance of this test and compare it to the nonparametric area test (Hanley and McNeil, 1983) using Monte Carlo simulations. We also study the effect of the choice of κ on the performance of the test.

3. Monte Carlo Simulations

The permutation test described above is designed as an omnibus test to compare entire ROC curves, as opposed to the area test, which compares a summary measure of the curves. We conducted a series of Monte Carlo simulations in order to assess the performance of the proposed test in relation to the area test. In these simulations, the marker values of the nondiseased subjects were generated from a standard normal distribution and those of the diseased subjects from $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ for X and Y , respectively. The areas under the ROC curve A_x and A_y for our simulation framework are given by $\Phi(\mu_x/(1 + \sigma_x^2)^{1/2})$ and $\Phi(\mu_y/(1 + \sigma_y^2)^{1/2})$, where Φ is the standard normal cumulative distribution function. The relationship between the two ROC curves can fall into one of two categories. The uniform alternative (where one curve is uniformly above the other) occurs when $\sigma_x^2 = \sigma_y^2$ and the crossing alternative (where the two curves cross) when $\sigma_x^2 \neq \sigma_y^2$. We choose the values of the parameters μ_x, μ_y, σ_x^2 , and σ_y^2 to give us a variety of A_x and A_y that covers both the uniform and the crossing alternative cases. Finally, since both the number of subjects and the sample prevalence of disease can differ for the two markers, we considered a variety of sample size combinations (n_0, n_1, m_0, m_1) to assess their effect on the performance of the test.

The reference test in these simulations is the nonparametric area test, which compares the areas under the two ROC curves. The nonparametric area estimate is the Mann-Whitney statistic, and its variance is obtained by the jackknife method in DeLong et al. (1988). Since the two areas are calculated from mutually independent samples, the variance of their difference is the sum of their individual variances. The simulation results are based on 2000 replications, and the p -value of the permutation test was computed using 1000 permutations. In these simulations, we set the value of κ to be $(n_1 + m_1)/(n_0 + n_1 + m_0 + m_1)$, which is the pooled estimate of disease prevalence.

We present in Table 1 the proportion of times a nominal 5% test rejects the null hypothesis for various configurations of parameters and sample sizes. The top half of the table corresponds to uniform alternatives, which are obtained by setting $\sigma_x^2 = \sigma_y^2 = 1$, while the bottom half corresponds to crossing alternatives, which are obtained by setting $\sigma_x^2 = 1$ and $\sigma_y^2 = 4$. We chose the difference in areas under the two ROC curves Δ to be one of 0, 0.1, or 0.2, and for each choice of Δ , we considered two different values for the area A_x . The table has two pairs of columns that correspond to the two sample size combinations. The total sample size $N_{\text{tot}} = n_0 + n_1 + m_0 + m_1$ is assigned in the ratios (a) 1:1:1:1 for the first pair and (b) 2:1:1:1 for the second. The total sample size N_{tot} was chosen to give approximately 80% power when the curves are unequal.

The two ROC curves are identical when the curves fall in the uniform alternative category and Δ is zero. The first two rows of the table correspond to this case and show that the permutation test has the correct size. Further simulation

Table 1

The proportion of times the area test (AT) and the permutation test (PT) test reject the null hypothesis at a nominal 5% level. The top and bottom halves correspond to the uniform and crossing alternatives, respectively. The parameters μ_x , μ_y , σ_x^2 , and σ_y^2 were chosen to give a specified area A_x and difference $\Delta (= A_y - A_x)$. Cases a and b correspond to the two different ratios in which the total sample size N_{tot} is distributed to the four groups.

Δ	A_x	N_{tot}	Case a		Case b	
			AT	PT	AT	PT
0.0	0.7	200	0.061	0.061	0.047	0.043
	0.8	200	0.052	0.050	0.047	0.047
0.1	0.7	600	0.739	0.726	0.706	0.680
	0.8	600	0.902	0.896	0.887	0.868
0.2	0.6	200	0.790	0.780	0.761	0.735
	0.7	200	0.924	0.918	0.905	0.877
0.0	0.7	600	0.060	0.726	0.052	0.610
	0.8	600	0.047	0.626	0.040	0.539
0.1	0.7	400	0.524	0.805	0.491	0.720
	0.8	400	0.725	0.882	0.696	0.820
0.2	0.6	160	0.703	0.765	0.643	0.677
	0.7	160	0.826	0.860	0.806	0.827

results, not presented here, confirm the accuracy of the permutation test. The rest of the table corresponds to unequal ROC curves and thus gives us the power of the two tests. The power of the area test is a function of the difference in areas Δ (which is the signed area between the two ROC curves), and the power of the permutation test is a function of the parameter θ (which is proportional to the unsigned area between the two ROC curves). Note that the signed and unsigned areas are identical for uniform alternatives, whereas the unsigned area is larger when the curves cross.

The next four rows in the top half of the table correspond to the powers of the two tests when one ROC curve is uniformly better than the other. Since the permutation test is designed as an omnibus test to detect any difference in the ROC curves, we expect it to have lower power than the area test. However, for uniform alternatives, the signed and unsigned areas between the two ROC curves, which determine the power of the area and the permutation tests, are equal. Thus, we expect the power of the two tests to be comparable, and the simulation results confirm this.

The bottom half of the table corresponds to crossing alternatives. As we noted earlier, the power of the permutation and area tests depends on the unsigned and signed areas between the curves. The first two rows in this half of the table, where Δ is zero, show that the permutation test can detect differences in the curves with high power whereas the area test cannot. The next four lines, where curves cross and the areas are different, show that the permutation test continues to have larger power than the area test. However, the difference in the two powers, which is a function of the relative difference between the unsigned and signed areas between the curves, decreases as the difference in the areas increases. Finally, from the two pairs of columns that correspond to cases a and b, we see that, regardless of the sample sizes and sample prevalence

rates of disease, the overall pattern of behavior of the permutation test in relation to the area test holds.

In the simulations above, we chose κ to be the overall prevalence of disease. A natural question to ask is whether this choice of κ is optimal. We conducted a series of simulations to study the effect of κ on the power of the permutation test. As in the earlier simulations, we considered both uniform and crossing alternatives as well as several sample sizes. For each case, we calculated the power for each value of κ in 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99. Data were generated independently for each value of κ so that the variability in the power of the area test, which does not depend on κ , would serve as a benchmark for the natural variation in the simulations. Across several sets of parameters and sample sizes, the variation in the power of the permutation test was incrementally larger than that of the area test. Thus, we feel confident that the pooled sample prevalence rate would give us power comparable to any other choice of κ .

4. Example

We now present an example to illustrate this method. The data for this example comes from a study conducted at the Memorial Sloan-Kettering Cancer Center on the use of three-dimensional conformal radiation therapy (3DCRT) for the treatment of prostate cancer. Patients with prostate cancer are treated with either a low dose (7020 Gy) or a high dose (7560 Gy) 3DCRT regimen. A side effect of radiation therapy for prostate cancer is rectal bleeding, which can occur several months after treatment. The dose level is known to be a factor that influences rectal bleeding. Since a lower rectal volume implies that a larger proportion of rectal wall receives radiation and hence has a potential for more damage, rectal volume is considered a factor that affects the occurrence of rectal bleeding. The question we pose here is whether the effect of volume is the same for both dose levels.

Rectal bleeding typically occurs within 30 months of treatment, and thus the study population is restricted to patients with at least 30 months follow-up. Since the proportion of patients who had rectal bleeding is low and the calculation of volume is labor intensive, we undertook the following case-control study to assess the effect of rectal volume on bleeding. All the bleeders were selected from both dose levels and the nonbleeders were sampled randomly within dose level (no matching was done). The data consists of the rectal volumes of 53 patients (13 bleeders and 40 nonbleeders) at the 7020 Gy dose level and 123 patients (41 bleeders and 82 nonbleeders) at the 7560 Gy dose level.

The empirical ROC curves for the two dose levels are shown in Figure 1. Observe that the two curves cross and appear to have similar areas. The estimated areas under the ROC curves and their standard errors for the two dose levels are 0.692 (0.092 SE) and 0.708 (0.047 SE). The area test to compare the two ROC curves gives us a p -value of 0.877. The permutation test, which compares entire ROC curves, is designed to detect differences in markers where the ROC curves cross while having similar areas, like the one we observe in this example. This test gives us a p -value of 0.346, which was obtained from 2500 permutations. The permutation distribution of the test statistic for this test is also shown in Figure 1. Both the tests do not reject the hypothesis that the ROC curves are equal,

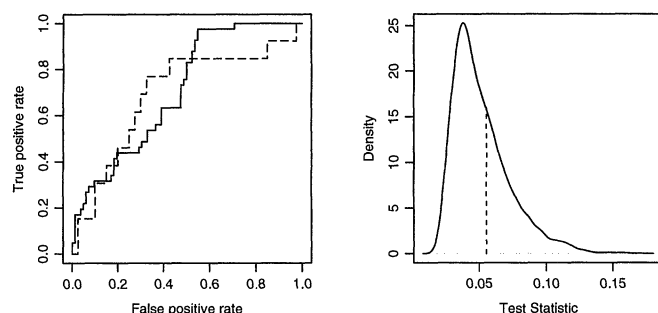


Figure 1. The ROC curves showing the effect of rectal volume on rectal bleeding by dose level are shown in the left panel. The solid and dashed lines correspond to 7020 Gy and 7560 Gy, respectively. The right panel shows the density of the square root of the test statistic obtained from the permutation distribution. The value of the statistic for the original data is shown as a dashed line.

suggesting that the effect of volume is the same for both dose levels. However, the failure to reject the null hypothesis could be due to the relatively small sample sizes. For these sample sizes, the area test has approximately 62% power to detect a difference in area of 0.2.

5. Discussion

The use of a diagnostic test in practice requires a classification point and a decision theoretic approach to decide between two diagnostic tools. Our earlier paper on the permutation method for paired data and the method herein for unpaired data give investigators fully nonparametric tests to compare entire ROC curves when the data are continuous. Since this method tests the equality of the two ROC curves at all operating points, it is helpful in detecting crossing ROC curves where one test could be genuinely superior despite having the same area under the curve. This method could easily be adapted to compare the curves over a range of interest, as in Wieand et al. (1989), by defining the parameter θ to be the integral on an interval (a, b) of the mixture distributions M instead of the entire $(0, 1)$ interval. The mixture distribution provides a calibration of the cutoffs. Since, under the null, sensitivities and specificities are equal, choosing a range of values of the mixture distributions is equivalent to specifying a range of specificities for the two markers. The permutation distribution for this test statistic is then generated exactly as before.

Another common problem in medical diagnostics is to adjust the comparison of diagnostic methods for covariates. When the covariate of interest is categorical with relatively few categories, a stratified analysis can be used where the permutation is done within stratum. However, if the covariate is continuous or has a large number of categories with few observations in each, stratification cannot be used. Alternately, one can use the semiparametric regression models proposed by Pepe (1998), where test scores, summary measures of accuracy, or the curves themselves are modeled, to adjust for covariates. However, they require assumptions that techniques such as stratification do not. Further research is needed to develop purely nonparametric approaches to address this problem.

ACKNOWLEDGEMENTS

The author is grateful to Drs Colin Begg and Glenn Heller for some helpful suggestions. The author would also like to thank Drs Michael Zelefsky, Andrew Jackson, and Mark Skwarchuk for the data used in the example. Finally, the author is grateful to the reviewer and the editors for their careful review and suggestions. This research is supported by the National Cancer Institute, award CA 73848.

RÉSUMÉ

Dans une publication précédente (Venkatraman and Begg, 1996, *Biometrika* **83**, 835–848), nous avons présenté un test de permutations de l'égalité de courbes ROC (Receiver Operating Characteristic Curves) dans le cas de données continues et appariées. Nous généralisons maintenant ce concept à un nouveau test de permutations pour données continues et non appariées, test dont nous étudions les propriétés au moyen de simulations.

REFERENCES

- DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–846.
- Dorfman, D. D. and Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals; rating method data. *Journal of Mathematical Psychology* **6**, 487–496.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.
- Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology* **148**, 839–843.
- Metz, D. E., Wang, P.-L., and Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging VIII*, F. Deconick (ed), 432–445. The Hague: Nijhoff.
- Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135.
- Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems. Methods from Signal Detection Theory*. New York: Academic.
- Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83**, 835–848.
- Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585–592.

Received October 1998. Revised April and December 1999.

Accepted June 2000.