Biometrika (1996), **83**, 4, *pp.* 835–848
Printed in Great Britain

A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment

BY E. S. VENKATRAMAN AND COLIN B. BEGG
Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, U.S.A.

SUMMARY

A distribution-free permutation test procedure is proposed for comparing receiver operating characteristic curves based on continuous data from a paired design. The method tests the hypothesis that the two curves are identical for all operating points, unlike previously proposed methods which test the equivalence of the areas under the curves. The new test is shown by simulation to have very similar operating characteristics to the standard method based on comparisons of the areas when the curves are parallel, but markedly superior power when the curves cross, that is when the curves are different but have similar areas. The prospects of generalising the approach to unpaired experiments and to comparisons of ordinal rating data are discussed.

Some key words: Diagnostic test; Exchangeability; Paired samples, Permutation test; Receiver operating characteristic curve.

1. INTRODUCTION

A common problem facing clinical researchers is the comparison of alternative diagnostic tests. Although diagnostic test accuracy is frequently characterised by two simple measures, the sensitivity and the specificity, reflecting the two types of error in diagnosis, false positive and false negative, comparisons of tests are hindered by the fact that these measures are arbitrarily determined by the selection of a classification point, or cut-off value. In recognition of this, it is necessary to calibrate this arbitrary classification rule to develop valid procedures for comparing tests. The most complete approach is to use receiver operating characteristic analysis, denoted by ROC. A ROC curve is a plot of the true positive ratio, the sensitivity, against the false positive ratio, the specificity subtracted from one. Since the ROC curves involve all possible classification points, the calibration of these points is embedded in ROC analysis.

This calibration problem is particularly pertinent when a binary test is being compared with a continuous test. Beam & Wieand (1991) approached this problem by comparing the test sensitivities after equating the specificities, by using the classification point for the continuous test that possesses a specificity that is the same as the specificity of the binary test. The idea of calibrating the comparison by fixing the specificities and comparing sensitivities was suggested originally by Greenhouse & Mantel (1950). It is, however, more common in practice to use methods that employ the entire ROC curves. These curves are often modelled using the assumption that the test results in diseased and nondiseased subjects are normally distributed, in which case a smooth curve can be estimated using maximum likelihood (Dorfman & Alf, 1969). In this case the curve is characterised by

two parameters, a slope and an intercept. Comparison of two ROC curves can be accomplished by developing a test of equivalence of the two slopes and the two intercepts. A likelihood ratio test for this purpose was developed by Metz, Wang & Kronman (1984) for paired data, where each patient receives both tests.

An alternative approach that is frequently used in practice is to consider the area under the curve as a measure of accuracy, and to compare the tests by evaluating the hypothesis that the area under the curve is the same for each test (Swets & Pickett, 1982). The trapezoidal area under the curve has been shown to be equivalent to the Mann–Whitney statistic for comparing diseased and nondiseased subjects (Bamber, 1975), permitting a nonparametric approach to the problem. For paired experiments, Hanley & McNeil (1983) built on this idea to develop a test that accounts for the correlations induced by the paired design. DeLong, DeLong & Clarke-Pearson (1988) developed a fully nonparametric approach in which all of the covariance terms are estimated nonparametrically, leading to an asymptotically normal test statistic.

A problem with using the area under the curve as the measure of accuracy is that the two tests may have different ROC curves which nonetheless have the same area. Receiver operating characteristic curves are typically asymmetric (Swets, 1986), and so two tests with different asymmetries could possess the same area. Since the use of a diagnostic test in practice involves an operating classification point, and since the utilities of these different classifications are, in principle, distinguishable using decision theory, one test may be genuinely superior to the other despite having the same area (Campbell, 1994). Therefore, it is preferable to construct a hypothesis test in which the null hypothesis represents equality of the entire ROC curves. In this paper we develop a simple permutation test for this purpose for paired data in § 2, study its properties via simulation in § 3, and illustrate its use by examples in § 4.

2. METHODS

Let X and Y represent the results of the two diagnostic tests, and let D be a binary indicator of true disease status, where 1 represents diseased and 0 represents normal subjects. Let the prevalence of disease in the study population be denoted by θ. In a retrospective study θ will be fixed by design, but regardless of the sampling scheme θ will be common to both tests in a paired experiment. A ROC is characterised by the distributions of the test result among diseased and nondiseased subjects.

Let the conditional distributions of the markers given the disease status be as follows:

$$F_x(x) = \text{pr}(X \leqslant x \mid D = 1), \quad G_x(x) = \text{pr}(X \leqslant x \mid D = 0),$$

$$F_y(y) = \text{pr}(Y \leqslant y \mid D = 1), \quad G_y(y) = \text{pr}(Y \leqslant y \mid D = 0).$$

The sensitivity and specificity of X at cut-off x are $1 - F_x(x)$ and $G_x(x)$ respectively, with corresponding definitions for Y. The two ROC curves are equivalent if, for every cut-off value x of X, there exists a corresponding cut-off value y of Y at which $F_x(x) = F_y(y)$ and $G_x(x) = G_y(y)$ and vice versa. Let the unconditional distributions of the markers be

$$\begin{aligned} M_x(x) &= \text{pr}(X \leqslant x) = \theta F_x(x) + (1 - \theta) G_x(x), \\ M_y(y) &= \text{pr}(Y \leqslant y) = \theta F_y(y) + (1 - \theta) G_y(y). \end{aligned} \tag{1}$$

The two curves are equivalent at all operating points if and only if there exists a transform-

ation $Y \to \mathscr{T}(Y)$ such that, for all x,

$$\begin{aligned} \text{pr}(X \leqslant x \mid D = 1) &= \text{pr}\{\mathscr{T}(Y) \leqslant x \mid D = 1\}, \\ \text{pr}(X \leqslant x \mid D = 0) &= \text{pr}\{\mathscr{T}(Y) \leqslant x \mid D = 0\}, \end{aligned} \tag{2}$$

and then $y \to M_x^{-1} M_y(y)$ is one such transformation. That is, if for every x there exists a corresponding y such that $F_x(x) = F_y(y)$ and $G_x(x) = G_y(y)$, then

$$\text{pr}(X \leqslant x) = \text{pr}(Y \leqslant y) \Rightarrow M_x(x) = M_y(y) \Rightarrow x = M_x^{-1} M_y(y).$$

Since ROC curves are invariant under monotone transformations, any comparison of the curves from the marker pair (X, Y) can be effected using the transformed data $\{X, \mathscr{T}(Y)\}$.

From (2) we see that the conditional distributions of X and $\mathscr{T}(Y)$ are identical, for both the normal and diseased populations, under the null hypothesis of equal ROC curves. If we further assume that the joint distributions are 'exchangeable', in the sense that the joint distributions of $\{X, \mathscr{T}(Y)\}$ and $\{\mathscr{T}(Y), X\}$ for both the normal and diseased subjects are the same, we can develop a permutation test by random permutations of $\{X, \mathscr{T}(Y)\}$ within subjects. This exchangeability assumption also means that the conditional distributions $\text{pr}\{X \mid \mathscr{T}(Y)\}$ and $\text{pr}\{\mathscr{T}(Y) \mid X\}$ are the same. Since the data are paired, this is a far less stringent assumption than independence of the two markers conditional on the disease status. Conditional independence is an unrealistic assumption in most studies of diagnostic tests (Begg, 1987).

Calibration of the ROC curves can be accomplished by evaluating the sensitivities and specificities of the two markers at the zth quantile of the unconditional distribution of the markers. Let $x_z = M_x^{-1}(z)$ and $y_z = M_y^{-1}(z)$. Consider the following transformation which maps marker pair (X, Y) and disease status D to a function from $[0, 1]$ to $\{-1, 0, 1\}$:

$$\mathscr{E}(z; X, Y, D) = \begin{cases} 1 & \text{if } (X \leqslant x_z, Y > y_z, D = 0) \text{ or } (X > x_z, Y \leqslant y_z, D = 1), \\ -1 & \text{if } (X > x_z, Y \leqslant y_z, D = 0) \text{ or } (X \leqslant x_z, Y > y_z, D = 1), \\ 0 & \text{otherwise.} \end{cases}$$

This function takes value 1 at a calibrated cut-off value of z when the result of test X is correct and the result of test Y is wrong, with the converse for -1. Values of 0 represent cut-offs at which either both tests are correct or both tests are wrong. It is shown in the Appendix that $E\{\mathscr{E}(z)\} = 0$ if and only if both the sensitivities and the specificities of the two tests are equal at z. Therefore the expectation of this function is identically zero if and only if the null hypothesis is true, that is if the two ROC curves are identical. The function

$$W(z) := \int \mathscr{E}(z; X, Y, D)\, d\hat{H}(X, Y), \tag{3}$$

which is a sample estimate of $E\{\mathscr{E}(z)\}$, is a measure of the 'closeness' of the two curves at the zth quantile where $\hat{H}$ is the empirical joint distribution of X and Y. In order to detect departures from the null hypothesis in all circumstances, including those where the ROC curves have similar overall area, but where the curves 'cross', an appropriate overall test statistic is

$$W_1 = \int_0^1 |W(z)|\, dz. \tag{4}$$

We show in the Appendix, with the assumption of exchangeability under the null hypothesis, that a permutation test using this statistic is consistent against the alternative of unequal ROC curves.

Direct computation of the statistic W is hindered by the absence of knowledge of the population quantiles of the unconditional distributions of the markers. Therefore we make use of the empirical quantiles, that is the rank statistics, in the following way. Denote the entire data set by $\{(X_i, Y_i, D_i); i = 1, \ldots, n\}$, where $D_i = 1$ if the case is diseased and $D_i = 0$ if the case is nondiseased. Furthermore let $\{R_i\}$ and $\{S_i\}$ denote the corresponding ranks of $\{X_i\}$ and $\{Y_i\}$, respectively. Then, setting $k = 1, \ldots, n-1$, we can define an empirical error matrix by

$$e_{ik} = \begin{cases} 1 & \text{if } (R_i \leqslant k, S_i > k, D_i = 0) \text{ or } (R_i > k, S_i \leqslant k, D_i = 1), \\ -1 & \text{if } (R_i > k, S_i \leqslant k, D_i = 0) \text{ or } (R_i \leqslant k, S_i > k, D_i = 1), \\ 0 & \text{otherwise.} \end{cases}$$

Analogous to $W(z)$, the statistic $e_{.k} := e_{1k} + \ldots + e_{nk}$ is a measure of the 'closeness' of the two ROC curves at the kth order statistic. The corresponding overall test statistic is

$$E := \sum_{k=1}^{n-1} |e_{.k}|.$$

This statistic is very easily computed by recognising that $e_{.k}$ is the difference in the total numbers of errors of each test when the kth classification point is used as is shown in § 4.

If the two tests are evaluated on the same metric, and there is no systematic measurement bias, then we can directly exchange the marker values for any subject to generate the permutation distribution as follows. Let $(q_1, \ldots, q_n)$ represent a sequence of 0's and 1's. Then a permuted data set $\{X_i^*, Y_i^*\}$ indexed by that sequence is given by

$$X_i^* = q_i X_i + (1 - q_i) Y_i, \quad Y_i^* = q_i Y_i + (1 - q_i) X_i \quad (i = 1, \ldots, n).$$

A new set of ranks $\{R_i^*, S_i^*\}$ is evaluated based on $\{X_i^*, Y_i^*\}$, and a corresponding statistic E^* is computed. The permutation distribution is the distribution which assigns a uniform mass to each value of E^* given by all the 2^n sequences of 0's and 1's. Since this may be a very large number, in practice we can use a sampling scheme where $(q_1, \ldots, q_n)$ is a random permutation generated by n fair coin tosses and the process is repeated a sufficiently large number of times to obtain a stable P-value. An example of a data set where the marker values can be directly exchanged is described in § 4.

If the direct exchangeability of X and Y is not considered to be an appropriate assumption, then it is necessary to rely on the ranked samples to evaluate the P-value, since in general the transformation $\mathcal{T}(Y)$ required for exchange of individual data values will be unknown. In this case each permuted set of ranks is generated by randomly exchanging pairs of ranks and reranking them. That is, we first generate $\{R_i^*, S_i^*\}$ using

$$R_i^* = q_i R_i + (1 - q_i) S_i, \quad S_i^* = q_i S_i + (1 - q_i) R_i \quad (i = 1, \ldots, n).$$

A similar idea was discussed by Campbell (1994). This process will invariably introduce numerous ties, so it is necessary to have a second randomisation step to break the ties. That is, we generate $\{R_i^{**}, S_i^{**}\}$, where

$$R_i^{**} = J(R_i^*), \quad S_i^{**} = J(S_i^*) \quad (i = 1, \ldots, n),$$

where $J(.)$ represents the process by which tied ranks are re-ranked by randomisation.

3. Operating characteristics of the test

The novelty of the proposed method is that it involves a test of the equality of the two ROC curves, as opposed to a test of an index of accuracy, such as the area under the curve. We have conducted a series of simulations to evaluate the operating characteristics of our permutation test, and to compare the method with the nonparametric method for comparing area indices (DeLong et al., 1988), called hereafter the area test.

The area test is based on the fact that the trapezoidal area under a single ROC curve is estimated by the Mann–Whitney two-sample U-statistic, the asymptotic variance of which is easily computed. Furthermore, formulae are available for the covariances between two correlated U-statistics, and DeLong et al. (1988) used these facts to develop the area test, using an asymptotic variance estimator. In our simulations we use this asymptotic formula for the area test. In our simulations of the permutation test we have used the permutation distribution as the reference distribution as we have been unable to develop a sufficiently accurate asymptotic approximation for the distribution of the statistic.

The simulation results are presented in Tables 1–3. These three tables distinguish the three distinctive scenarios of importance. In all cases the test results for the nondiseased subjects are generated from a standard normal $N(0, 1)$ distribution, while the test results for the diseased subjects are generated from $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ for tests X and Y respectively. The simulations distinguish the models in which either the test results are statistically independent, conditional on D, despite the paired design, $\rho = 0$, or the test results are correlated with correlation $\rho = 0{\cdot}5$. The ROC curves are identical if and only if $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$. This is the case in all simulations in Table 1. The results show clearly that the permutation test has a similar size to the area test in all configurations of area and sample sizes studied.

Table 2 contains configurations in which test Y is uniformly superior to test X, evidenced by the superior separation between diseased and normal subjects, that is $\mu_y > \mu_x$ with equivalent variances, and correspondingly superior area indices, $A_y > A_x$. These represent configurations in which it is appropriate to summarise diagnostic accuracy using the area index. An example of the ROC curves for one of these configurations is displayed in Fig. 1(a). Notice that the curve for test Y is uniformly higher than the curve for text X.

Table 1. *Comparison of test sizes*

Area		Sample sizes				Area test		Permutation test	
A_x	A_y	n_d	$n_{\bar{d}}$	$\mu_x = \mu_y$	$\sigma_x^2 = \sigma_y^2$	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$
0·6	0·6	20	20	0·36	1·0	0·053	0·051	0·049	0·042
		40	40			0·065	0·056	0·063	0·061
		80	80			0·048	0·043	0·045	0·047
0·7	0·7	20	20	0·74	1·0	0·058	0·047	0·055	0·062
		40	40			0·045	0·057	0·048	0·060
		80	80			0·049	0·041	0·049	0·043
0·8	0·8	20	20	1·19	1·0	0·059	0·048	0·052	0·055
		40	40			0·043	0·042	0·043	0·050
		80	80			0·056	0·038	0·057	0·038
0·9	0·9	20	20	1·81	1·0	0·040	0·028	0·044	0·041
		40	40			0·038	0·042	0·043	0·050
		80	80			0·045	0·035	0·048	0·043

Table 2. *Power against uniform alternatives*

Areas		Sample sizes						Test power: Area test		Test power: Permutation test	
A_x	A_y	n_1	n_2	μ_x	σ_x^2	μ_y	σ_y^2	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$
0·6	0·7	20	20	0·36	1·0	0·74	1·0	0·127	0·200	0·121	0·184
		40	40					0·210	0·349	0·199	0·321
		80	80					0·394	0·599	0·377	0·569
0·6	0·8	20	20	0·36	1·0	1·19	1·0	0·401	0·647	0·367	0·600
		40	40					0·686	0·913	0·671	0·897
0·6	0·9	20	20	0·36	1·0	1·81	1·0	0·846	0·971	0·829	0·959
		40	40					0·981	0·998	0·980	0·997
0·7	0·8	20	20	0·74	1·0	1·19	1·0	0·145	0·215	0·131	0·208
		40	40					0·253	0·455	0·245	0·430
		80	80					0·481	0·706	0·460	0·683
0·7	0·9	20	20	0·74	1·0	1·81	1·0	0·556	0·763	0·536	0·753
		40	40					0·857	0·976	0·842	0·970
0·8	0·9	20	20	1·19	1·0	1·81	1·0	0·182	0·292	0·176	0·277
		40	40					0·393	0·565	0·362	0·542
		80	80					0·685	0·889	0·673	0·865

The simulated power estimates show that the two test statistics have very similar power, with the area test being fractionally more powerful in all cases.

Table 3 contains configurations in which the use of the area test is inappropriate. These are configurations in which the diagnostic tests have quite different ROC curves, but the area indices are nevertheless identical. An example is provided in Fig. 1(b). The simulations show that the permutation test does possess limited power to detect differences of this

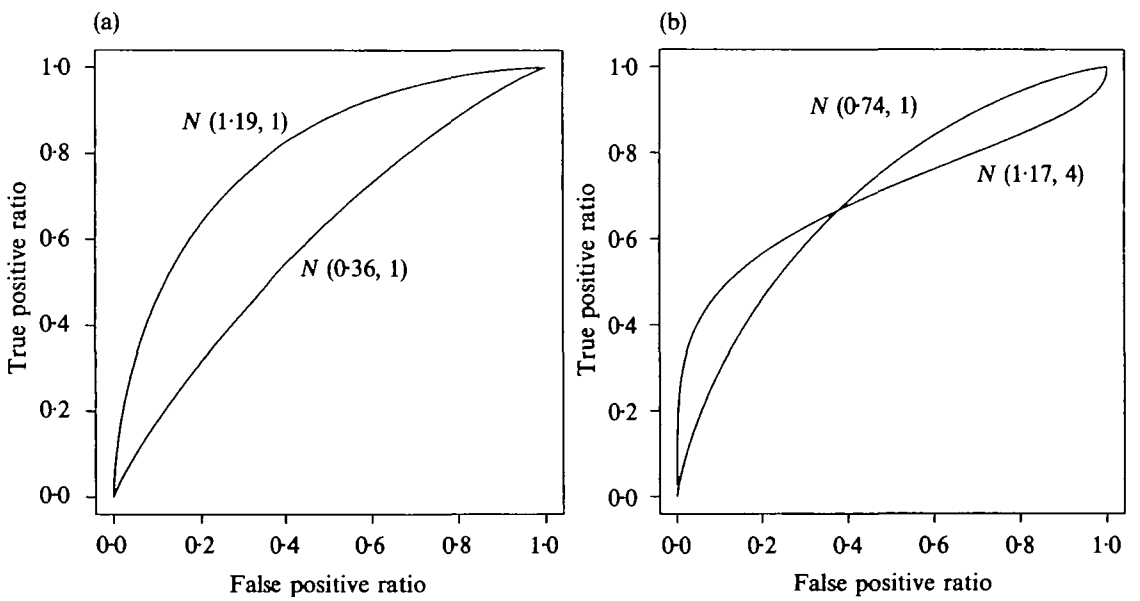

Fig. 1. Receiver operating characteristic curves: (a) generated from diseased populations $N(1{\cdot}19, 1)$ test Y, and $N(0{\cdot}36, 1)$, test X; (b) generated from diseased populations $N(0{\cdot}74, 1)$ and $N(1{\cdot}17, 4)$. In each case normal subjects are generated from $N(0, 1)$.

Table 3. *Power against crossing alternatives*

Areas		Sample sizes						Test power			
								Area test		Permutation test	
A_x	A_y	n_1	n_2	μ_x	σ_x^2	μ_y	σ_y^2	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$	$\rho = 0{\cdot}0$	$\rho = 0{\cdot}5$
0·6	0·6	20	20	0·36	1·0	0·56	4·0	0·063	0·060	0·077	0·148
		40	40					0·056	0·039	0·152	0·319
		80	80					0·054	0·052	0·361	0·669
0·7	0·7	20	20	0·74	1·0	1·17	4·0	0·049	0·051	0·074	0·141
		40	40					0·059	0·041	0·131	0·298
		80	80					0·049	0·036	0·316	0·652
0·8	0·8	20	20	1·19	1·0	1·88	4·0	0·046	0·054	0·060	0·129
		40	40					0·057	0·054	0·119	0·238
		80	80					0·052	0·051	0·259	0·515
0·9	0·9	20	20	1·81	1·0	2·87	4·0	0·051	0·041	0·062	0·092
		40	40					0·038	0·049	0·084	0·154
		80	80					0·044	0·039	0·182	0·348

nature, while the area test is extremely insensitive, as we would expect. Note that, in both Table 2 and Table 3, highly correlated test results lead to increased power.

In summary, the simulations show that the permutation test has very good operating characteristics. It has essentially equivalent power to the nonparametric area test in all circumstances in which the area test is appropriate, and clearly superior power in configurations in which one ROC curve is not uniformly superior to the other, that is where the curves differ but the area indices are similar.

4. Examples

We present two worked examples of the new method. In the first of these, a comparison of techniques for diagnosing melanoma, the two diagnostic test scores are derived from different data items. As a result they are not exchangeable, and so we must construct the test on the basis of the ranks. In the second example, a comparison of the accuracy of using computed tomography in different nodal sites for staging testicular cancer, the two diagnostic approaches involve measuring by eye in millimeters the size of the largest lymph node in two distinct anatomic regions, the null hypothesis being that each anatomic region is equally predictive of disease spread. Exchangeability of the measured test results is a reasonable premise in this case, and so we have employed the permutation test based on the measured test results.

Example 4·1. The definitive diagnosis of a pigmented lesion suspected of being a melanoma involves a biopsy. Dermatologists are frequently faced with the task of clinically evaluating suspicious lesions to determine which ones warrant a biopsy evaluation. They do this on the basis of observable features of the lesion such as asymmetry, border irregularity, colouration and size (Stolz et al., 1994). The use of a dermoscope helps to clarify these visible features. In the example presented in Table 4 investigators have examined 72 suspicious lesions using both a clinical scoring scheme without the dermoscope, and a dermoscopic scoring scheme. The features examined and the scoring system were somewhat different in each case and so the resulting scores are not regarded as exchangeable. The purpose of our analysis is to determine whether the dermoscope contributes diagnostic

Table 4. *Results from Example* 4·1

	Test X				Test Y				
Rank	Patient identifier	Test result	Disease status	Total errors	Patient identifier	Test result	Disease status	Total errors	$e_{.k}$
1	34	−5·881	0	50	34	−7·103	0	50	0
2	65	−5·164	0	49	6	−6·568	0	49	0
3	23	−4·952	0	48	5	−6·524	0	48	0
4	6	−4·788	0	47	65	−6·344	0	47	0
5	11	−4·764	0	46	28	−5·783	0	46	0
6	55	−4·717	0	45	40	−5·418	0	45	0
7	9	−4·576	0	44	10	−5·297	0	44	0
8	72	−4·412	0	43	72	−4·986	0	43	0
9	29	−4·376	0	42	2	−4·933	0	42	0
10	5	−4·363	0	41	62	−4·782	0	41	0
11	19	−4·060	0	40	46	−4·621	0	40	0
12	26	−4·023	0	39	19	−4·399	0	39	0
13	62	−3·835	0	38	55	−4·315	0	38	0
14	59	−3·776	0	37	67	−4·206	0	37	0
15	17	−3·503	0	36	9	−4·131	0	36	0
16	43	−3·471	0	35	59	−4·108	0	35	0
17	2	−3·412	0	34	45	−3·831	0	34	0
18	28	−3·221	0	33	23	−3·808	0	33	0
19	40	−3·221	0	32	29	−3·799	0	32	0
20	46	−3·060	0	31	18	−3·641	0	31	0
21	33	−2·871	0	30	48	−3·545	0	30	0
22	8	−2·763	1	31	43	−3·502	0	29	−2
23	10	−2·716	0	30	26	−3·493	0	28	−2
24	57	−2·457	0	29	11	−3·476	0	27	−2
25	45	−2·408	0	28	33	−3·244	0	26	−2
26	25	−2·389	0	27	38	−3·240	0	25	−2
27	15	−2·293	0	26	4	−2·857	0	24	−2
28	7	−2·254	0	25	57	−2·808	0	23	−2
29	18	−2·245	0	24	21	−2·730	0	22	−2
30	48	−2·102	0	23	44	−2·609	0	21	−2
31	21	−2·089	0	22	52	−2·220	0	20	−2
32	60	−1·924	0	21	69	−2·085	0	19	−2
33	67	−1·750	0	20	1	−1·941	1	20	0
34	38	−1·738	0	19	51	−1·932	0	19	0
35	69	−1·726	0	18	63	−1·847	0	18	0
36	52	−1·714	0	17	54	−1·728	0	17	0
37	13	−1·677	0	16	70	−1·472	1	18	2
38	51	−1·623	0	15	7	−1·443	0	17	2
39	4	−1·622	0	14	17	−1·121	0	16	2
40	54	−1·597	0	13	25	−1·118	0	15	2
41	1	−1·571	1	14	74	−0·998	0	14	0
42	27	−1·430	0	13	15	−0·939	0	13	0
43	74	−1·193	0	12	14	−0·805	0	12	0
44	64	−1·174	1	13	66	−0·770	1	13	0
45	36	−1·112	0	12	64	−0·744	1	14	2
46	44	−1·057	0	11	60	−0·676	0	13	2
47	68	−0·973	1	12	24	−0·656	0	12	0
48	70	−0·961	1	13	41	−0·476	1	13	0
49	63	−0·441	0	12	30	−0·435	1	14	2
50	76	−0·421	1	13	36	−0·395	0	13	0

Table 4. (*Continued*)

	Test X				Test Y				
Rank	Patient identifier	Test result	Disease status	Total errors	Patient identifier	Test result	Disease status	Total errors	$e_{.k}$
51	20	−0·091	0	12	37	−0·323	0	12	0
52	24	−0·079	0	11	16	−0·314	0	11	0
53	12	−0·055	1	12	49	−0·119	1	12	0
54	49	−0·052	1	13	8	0·032	1	13	0
55	16	0·183	0	12	68	0·241	1	14	2
56	37	0·298	0	11	56	0·247	1	15	4
57	32	0·314	1	12	39	0·258	1	16	4
58	39	0·359	1	13	77	0·399	1	17	4
59	30	0·406	1	14	20	0·483	0	16	2
60	31	0·786	1	15	58	0·593	0	15	0
61	58	1·007	0	14	32	0·887	1	16	2
62	14	1·183	0	13	3	0·916	1	17	4
63	66	1·195	1	14	50	1·262	0	16	2
64	41	1·407	1	15	13	1·279	0	15	0
65	56	1·491	1	16	27	1·548	0	14	−2
66	47	1·631	1	17	73	1·563	1	15	−2
67	50	1·751	0	16	47	1·695	1	16	0
68	53	1·939	1	17	76	1·873	1	17	0
69	77	1·939	1	18	31	2·111	1	18	0
70	3	2·398	1	19	12	2·146	1	19	0
71	73	2·762	1	20	53	2·407	1	20	0
72	35	3·032	1	21	35	3·258	1	21	0

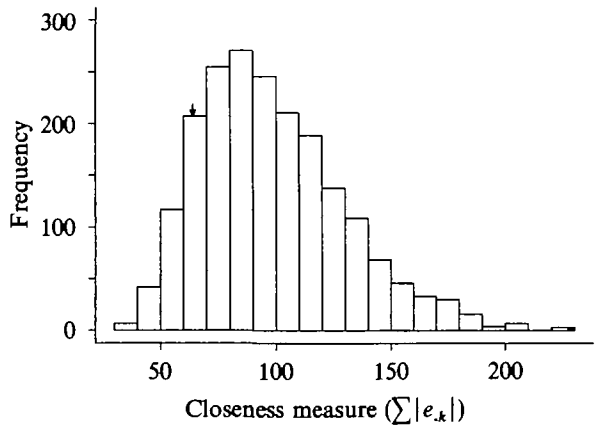

Fig. 2. Reference distribution for Example 4·1. Value of test statistic is indicated by the arrow.

information. The null hypothesis is that the dermoscope contributes no useful information, and this is represented by equivalence of the two ROC curves.

The data are organised in Table 4 to reveal the raw data and to illustrate computation of the test statistic. Each of the scoring systems, tests X and Y, is ranked, and is also labelled by a unique patient identifier to reveal the pairings. For example patient number 62 had a clinical score of −3·835 and a dermoscopic score of −4·782. This patient was negative for melanoma on biopsy. The clinical score was ranked 13, that is 13th from

least positive, and the dermoscopic score was ranked 10. This patient contributes to the test statistic E only for those values of k between 10 and 12 in that, if the classification point for diagnosis were in this range, the clinical score X would be a false positive and the dermoscopic score Y would be a true negative. For values of k less than 10 both scores would be false positives, and for values of k greater than 12 both would be true negatives. The kth specific contribution $e_{.k}$ to the overall test statistic is easily computed from the table by subtracting the total number of errors on test Y from the total number of errors on test X, where the errors are the sums of the false negative and false positive frequencies. These are displayed in the last column of the table. These differences are uniformly small, demonstrating that the two scoring systems are essentially equivalent. The overall test statistic is 64. The reference distribution, obtained by randomly permuting the pairs and recomputing the statistic, is displayed in Fig. 2. The two-sided P-value is 0·88. This was computed using 1000 permutations.

Example 4·2. In patients with clinically localised primary testicular cancer it is important to determine the necessity for an operation to remove any disease that may have spread to the retroperitoneal lymph nodes. These nodes can be evaluated by computed tomography to determine the necessity for this operation. In our data set the size of the largest node detected by computed tomography was used as the diagnostic criterion, and the goal of the study was to determine if the accuracy of this criterion is different for anterior versus posterior nodes. The 'gold standard' diagnosis is the presence of any nodal disease at surgery. In the data set in Table 5, from the, as yet unpublished report by S. Hilton et al. 'CT for detection of retroperitoneal lymph node metastases in patients with clinical stage 1 testicular nonseminomatous germ cell cancer: assessment of size criteria', the test result recorded is the average size in millimeters of the largest node detected by three independent readers. Since anything smaller than 4 millimeters is considered undetectable by the naked eye, many observations were designated as undetectable, recorded as 3·9 in the table but representing <4 mm. If the null hypothesis that the sizes of anterior and posterior nodes possess equivalent diagnostic information is true, then we can infer that the measured sizes are exchangeable, and so we have used the test based on the raw scores rather than the ranked scores as in Example 4·1.

This data set has the unusual feature of having a number of pairs with tied scores, even though the test results are continuous. There are many ties at the undetectable level 3·9 and several ties at higher levels. In fact there are only eleven values of the ranks at which both tests exhibit a jump, that is neither spans the rank with a tie. These occur at ranks 48, 49, 53, 58, 59, 70, 62, 63, 66, 67, 68. Instead of computing the test statistic at every possible value of k we have restricted it to these eleven values. This leads to a test statistic of 64 and a P-value of 0·033, based on 1000 random permutations.

5. DISCUSSION

The methodology derived in this paper is directly applicable only to continuous data from paired experiments. The paired design is commonly employed in comparisons of diagnostic tests. However, unpaired comparisons are frequently used also. If the tests are evaluated on the same metric, then an obvious generalisation would be to create an analogous error function based on the total number of errors for each diagnostic test at each rank of the combined sample, and to obtain a reference distribution by using random permutations of the combined sample of X and Y values within the diseased and non-diseased categories respectively. If the test results are not directly exchangeable then the

Table 5. *Data for Example* 4·2

Patient identifier	Test X	Test Y	Disease status	Patient identifier	Test X	Test Y	Disease status
1	3·9	3·9	0	36	3·9	3·9	0
2	5·3	5·0	0	37	3·9	3·9	0
3	4·6	3·9	0	38	3·9	3·9	0
4	3·9	3·9	0	39	5·3	4·9	0
5	7·7	4·0	0	40	3·9	3·9	0
6	3·9	4·9	0	41	9·6	3·9	1
7	3·9	3·9	0	42	3·9	3·9	1
8	5·3	3·9	0	43	4·6	4·9	1
9	3·9	3·9	0	44	3·9	4·9	1
10	4·3	4·3	0	45	17·6	4·9	1
11	5·0	4·3	0	46	5·6	3·9	1
12	9·0	3·9	0	47	21·3	3·9	1
13	4·6	3·9	0	48	5·6	3·9	1
14	5·3	3·9	0	49	8·7	3·9	1
15	3·9	3·9	0	50	14·0	3·9	1
16	3·9	3·9	0	51	18·6	3·9	1
17	3·9	3·9	0	52	3·9	4·3	1
18	4·0	3·9	0	53	4·6	3·9	1
19	3·9	3·9	0	54	7·6	9·3	1
20	3·9	7·3	0	55	9·3	5·6	1
21	5·5	4·6	0	56	15·0	7·6	1
22	3·9	3·9	0	57	3·9	5·9	1
23	3·9	3·9	0	58	6·3	3·9	1
24	5·7	3·9	0	59	3·9	3·9	1
25	3·9	3·9	0	60	4·3	3·9	1
26	3·9	3·9	0	61	5·3	6·6	1
27	3·9	3·9	0	62	6·6	3·9	1
28	3·9	5·6	0	63	15·0	9·3	1
29	3·9	3·9	0	64	5·0	3·9	1
30	6·6	5·5	0	65	3·9	3·9	1
31	5·0	3·9	0	66	19·0	3·9	1
32	3·9	6·3	0	67	15·0	3·9	1
33	3·9	3·9	0	68	6·0	6·3	1
34	7·6	3·9	0	69	7·3	3·9	1
35	3·9	3·9	0	70	14·3	3·9	1

ranks of the individual tests would have to be standardised first. This could lead to ties in the permuted samples depending on the sample sizes in the two test groups. Such an approach would be dependent on the assumptions that test results are independent and identically distributed within disease categories, and the asymptotic population prevalences of disease are the same in the two independent test samples, which is commonly not the case in retrospective comparisons of diagnostic tests. These extensions require further research to evaluate the properties of the methods.

Another very common application of ROC analysis is in the comparison of subjective radiological tests, where the test results are recorded as subjective ordinal classifications, often on a five-point scale. Since even in a paired experiment the ordinal classification points will not be calibrated between the two diagnostic tests, the ratings are, in general, not exchangeable. Furthermore, the ranks are equivalent to the ratings in this setting, so this highlights the problem crated for our method by tied test results between patients. In

fact the theory in § 2 indicates that in comparing ROC curves, the calibration points for comparisons are the equivalent rankings in the rank-order statistic. With ordinal data we can only obtain truly nonparametric estimates of the ROC curves at the limited number of pre-defined classification points defined by the rank order statistic, that is $p-1$ points for a rating scale with p categories. Moreover these rankings will invariably be different for the two diagnostic tests. In this setting any evaluation of equivalence of the ROC curves implicitly requires a probability model for the curves to facilitate extrapolation to the other potential classification points. This reasoning suggests that a fully nonparametric comparison of ROC curves is theoretically impossible for classified data. Nonetheless, adaptations of our permutation-based approach to this setting may still possess good statistical properties, and further research is needed in this area.

ACKNOLWEDGEMENT

We are grateful to Dr Susan Hilton and colleagues for permission to use their data on the staging of testicular cancer, to Dr Alfred Kopf and colleagues for permission to use the data concerning the diagnosis of melanoma, and to Professor Richard Olshen for helpful advice. The research is supported by the National Institutes of Health.

APPENDIX

Consistency of the permutation test

We will show that the permutation test is consistent when the ROC curves are unequal. For any $0 \leqslant z \leqslant 1$, the expected value of $\mathscr{E}(z; X, Y, D)$ is given by

$$\begin{aligned} E\{\mathscr{E}(z; X, Y, D)\} &= \theta \operatorname{pr}(X > x_z, Y \leqslant y_z | D = 1) + (1-\theta) \operatorname{pr}(X \leqslant x_z, Y > y_z | D = 0) \\ &\quad - \theta \operatorname{pr}(X \leqslant x_z, Y > y_z | D = 1) - (1-\theta) \operatorname{pr}(X > x_z, Y \leqslant y_z | D = 0) \\ &= \theta\{\operatorname{pr}(X > x_z | D = 1) - \operatorname{pr}(Y > y_z | D = 1)\} \\ &\quad + (1-\theta)\{\operatorname{pr}(Y > y_z | D = 0) - \operatorname{pr}(X > x_z | D = 0\} \\ &= \theta\{F_y(y_z) - F_x(x_z)\} + (1-\theta)\{G_x(x_z) - G_y(y_z)\}. \end{aligned} \tag{A1}$$

However, from (1) it is easily seen that

$$F_x(x_z) > F_y(y_z) \Leftrightarrow G_x(x_z) < G_y(y_z).$$

Therefore $E\{\mathscr{E}(z; X, Y, D)\} = 0$ if and only if $F_x(x_z) = F_y(y_z)$ and $G_x(x_z) = G_y(y_z)$. Thus the expectation is zero if and only if the ROC curves are identical.

Observe that the function $W(z)$ as defined in (3) is the average of independent and identically distributed random functions, $\mathscr{E}(z; X, Y, D)$, and so it converges to its expected value $E\{\mathscr{E}(z; X, Y, D)\}$ almost surely. Thus

$$W = \int_0^1 |W(z)|\, dz \to \int_0^1 |E\{\mathscr{E}(z; X, Y, D)\}|\, dz, \tag{A2}$$

almost surely as n goes to infinity. Since the distribution functions are continuous, we can see from (A1) that $E\{\mathscr{E}(z; X, Y, D)\}$ is continuous. Hence, from the previous paragraph, W converges to zero if the ROC curves are identical and to a positive number, say c, if they are not.

Since the ROC curves are invariant under monotone transformations, we assume, without loss of generality, that the marker values are between 0 and 1, and M_x and M_y are uniform. Let P represent the true probability distribution of the markers. The permutation of the two markers within a subject symmetrises the joint distributions of the markers because the permutation mechan-

ism ensures that the conditional probability of an observed marker pair is 0·5, that is

$$\text{pr}[X = x, Y = y | (X, Y) \in \{(x, y), (y, x)\}] = \text{pr}[X = y, Y = x | (X, Y) \in \{(x, y), (y, x)\}] = 0{\cdot}5.$$

Let Q denote the probability distribution of the markers derived from P by the above symmetrisation. Observe that P and Q are the same if the ROC curves are identical and the markers are exchangeable. Thus the permutation mechanism does not alter the distribution under the null hypothesis, and the test has the correct size. When the curves are unequal, the above symmetrisation renders the two ROC curves identical under Q because it enforces exchangeability of the joint distribution.

If the ROC curves from P are not identical, by (A2) W converges almost surely to 0 under Q and to a constant $c > 0$ under P. Observe that

$$\text{pr}_P(W > c') \to 1, \quad \text{pr}_Q(W < c') \to 1, \tag{A3}$$

for any $0 < c' < c$, as n goes to infinity. So we can make the probabilities arbitrarily close to 1, for all sample sizes larger than an appropriately chosen n.

For any given observed data set, let $W_{(1-\alpha)}$ denote the $1 - \alpha$ quantile of the statistic W derived from all the 2^n permutations of markers, where n is the sample size. Observe that the value of $W_{(1-\alpha)}$ remains the same regardless of which one of the 2^n permutations was actually observed. Thus the distribution of $W_{(1-\alpha)}$ depends on P only through Q since the two distributions differ only in the individual probabilities of observing each of the 2^n permutations, conditioned on observing one of the permutations. That is

$$\text{pr}_P(W_{(1-\alpha)} \leqslant c') = \text{pr}_Q(W_{(1-\alpha)} \leqslant c')$$

for all $c' > 0$. We need to show that, for any $c' > 0$, $\text{pr}_Q(W_{(1-\alpha)} \leqslant c')$ can be made as close to 1 as desired. Since

$$\begin{aligned} \text{pr}_Q(W > c') &\geqslant \text{pr}_Q(W_{(1-\alpha)} > c' \text{ and } W > W_{(1-\alpha)}) \\ &\geqslant (1 - \alpha)\, \text{pr}_Q(W_{(1-\alpha)} > c'), \end{aligned} \tag{A4}$$

from (A3) $\text{pr}_Q(W_{(1-\alpha)} > c') \to 0$. Finally, observe that

$$\begin{aligned} \text{pr}_P(W > W_{(1-\alpha)}) &\geqslant \text{pr}_P(W > W_{(1-\alpha)} \text{ and } W_{(1-\alpha)} \leqslant c') \\ &\geqslant \text{pr}_P(W > c' \text{ and } W_{(1-\alpha)} \leqslant c') \\ &> \text{pr}_P(W > c') + \text{pr}_P(W_{(1-\alpha)} \leqslant c') - 1. \end{aligned}$$

Since the power of the test is $\text{pr}_P(W > W_{(1-\alpha)})$, and since by (A3) and (A4) both $\text{pr}_P(W > c')$ and $\text{pr}_P(W_{(1-\alpha)} \leqslant c')$ converge to 1, it follows that the power of the test converges to 1 unless the null hypothesis is true.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**, 387–415.

Beam, C. A. & Wieand, H. S. (1991). A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* **47**, 907–19.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statist. Med.* **6**, 411–23.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statist. Med.* **13**, 499–508.

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–46.

Dorfman, D. D. & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals; rating method data. *J. Math. Psychol.* **6**, 487–96.

Greenhouse, S. W. & Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics* **6**, 399–412.

Hanley, J. A. & McNeil, B. J. (1983). A method of comparing the area under two ROC curves derived from the same cases. *Radiology* **148**, 839–43.

METZ, D. E., WANG, P-L. & KRONMAN, H. B. (1984). A new approach for testing the significance of differences between ROC curves for correlated data. In *Information Processing in Medical Imaging*, Ed. F. Deconick, pp. 432–45. The Hague: Nijhoff.

STOLZ, W., RIEMANN, A., COGNETTA, A.B., PILLET, L., ABMAYR, W., HÖLZEL, D., BILEK, P., NACHBAR, F., LANDTHALER, M. & BRAUN-FALCO, O. (1994). ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. *Europ. J. Dermat.* **7**, 521–7.

SWETS, J. A. (1986). Form of empirical ROC's in discrimination and diagnostic tasks: implications of theory and measurement of performance. *Psychol. Bull.* **99**, 181–98.

SWETS, J. A. & PICKETT, R. M. (1982). *Evaluation of Diagnostic Systems. Methods from Signal Detection Theory*. New York: Academic Press.

[*Received September* 1995. *Revised December* 1995]