STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1239–1250
Published online 12 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2338

Combining dependent tests to compare the diagnostic accuracies—a non-parametric approach

Yuqing Yang*,† and Zhezhen Jin

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, U.S.A.

SUMMARY

In this paper, we propose a non-parametric approach for comparing diagnostic accuracies in multi-reader receiver operating characteristic (ROC) studies. The approach constructs a test from each reader by extending the conventional non-parametric method and then combines all the individual test statistics to draw an overall conclusion on the relative accuracies of different diagnostic tests. The method can handle both continuous and ordinal data. Compared to the existing non-parametric methods, the method is robust and effectively deals with the possible heterogeneity among readers. It can also be applied to the analysis of correlated ROC studies. The method is applied to a real example and its finite sample performance is examined through simulation studies. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: diagnostic accuracy; repeated measurements; area under the curve; non-parametric approach; combining dependent tests

1. INTRODUCTION

In diagnostic medicine, it is very common that the diagnostic test result is read by a reader and thus depends on the reader's interpretation. To assess the intrinsic abilities of such diagnostic tests in discriminating the diseased subjects from the non-diseased ones, multi-reader ROC studies are often conducted, in which each patient is examined by several readers with multiple diagnostic tests. In such studies, the comparison of the accuracies of different diagnostic tests is challenging due to the complicated correlation structure among the test results, see Reference [1]. The test responses, which are from different patients, interpreted

*Correspondence to: Yuqing Yang, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, U.S.A.
†E-mail: yy2019@columbia.edu

Contract/grant sponsor: New York City Council
Contract/grant sponsor: National Science Foundation

Received 30 November 2004
Accepted 13 June 2005

by the same reader are correlated, so are the test responses from the same patient. Moreover, since the diagnostic test results depend on reader's subjective interpretation, they might vary among different readers.

Several approaches have been proposed in the literature for the analysis of such data. Dorfman *et al*. [2] proposed a mixed-effects ANOVA model in which pseudo-values of the summary measures of ROC curve for each subject are computed by jackknife method, then a mixed-effects ANOVA model is used for comparing the accuracies of two diagnostic tests. Obuchowski and Rockette [3] applied a mixed-effects ANOVA model directly to the estimated summary measures of the ROC curve for each combination of readers and tests, and used a corrected F test statistic to test the equality of the accuracies of diagnostic tests. Two types of regression models have also been studied for the analysis of the multi-reader ROC study data: Tosteson and Begg [4], and Toledano and Gatsonis [5] discussed an indirect ROC regression model in which all test responses are assumed from a location-scale family and a regression model is fitted for the test responses, the ROC curves are then derived from the estimated regression parameters. Pepe [6], and Cai and Pepe [7] described a direct regression model on ROC curves. The common theme of the two types of regression models is that they both include readers and test types as covariates. The accuracies of different diagnostic tests are compared by the estimated coefficients of those covariates. The validity of these approaches, however, depends critically on the correct model specification.

Recently, a non-parametric method is proposed by Lee and Rosner [8]. The approach constructs an average ROC curve over all readers and compares the areas under the average ROC curves. Specifically, if there are K readers who examine a sample of subjects with a diagnostic test t, then each subject would have K measured test results from the test t. Lee and Rosner [8] constructed a Mann–Whitney U-statistic for each pair of diseased and non-diseased subjects by comparing the K^2 pairs of measured test results from test t, the area under the average ROC curve is then estimated by the average of Mann–Whitney U-statistics from all pairs of diseased and non-diseased subjects. This method is easy to implement, but it fails to take the possible heterogeneity among different readers into account by comparing a reader's scores for each diseased subject with every other reader's scores for each non-diseased subject. Since the diagnostic test results depend on readers' subjective interpretations and the tests under evaluation are usually new-developed methods, it often happens that the scales of readers' interpretations on test results do not agree very well: some readers might tend to give higher scores to both diseased and non-diseased patients and some other readers tend to give lower scores to both diseased and non-diseased subjects. Throughout the paper, the 'heterogeneity among readers' refers to this possible poor agreement among readers. Although the diagnostic test has an inherent ability of discriminating subjects with disease from those without, and the AUC of test from each reader reflects this inherent ability of the test, the area under the average ROC curve might be biased due to the poor agreement of readers. Thus, the hypothesis tests based on this average ROC curve may have inappropriate size and power. Since the inherent accuracy of diagnostic test is of our interest, it is desired to evaluate the diagnostic tests while adjusting for the possible heterogeneity among readers. In this paper, we address this issue and propose a non-parametric method for comparing the diagnostic accuracies. The proposed method is robust and can effectively handle the possible heterogeneity among readers. It can also deal with diagnostic accuracy studies in which repeated measurements for the same characteristic are obtained under different conditions for each subject.

The paper is organized as the following. In Section 2, we present the method of combining dependent tests along the line of Wei and Johnson [9]. In Section 3, we illustrate the method by using data from the experiment conducted by Muller *et al.* [10] for optimizing the operating parameters for photographic images used in scintigraphy. Simulation studies are presented in Section 4. We conclude the paper with some discussions and remarks in Section 5.

2. METHOD

Let Y and X denote the test results from the truly diseased and the truly non-diseased subjects, respectively, such that higher value of test result indicates the higher possibility of the presence of disease. If a threshold value c is used to determine the test result as positive or negative, then the true positive rate (TPR) and the false positive rate (FPR) at the threshold c are $\mathrm{TPR}(c) = P(Y \geqslant c)$ and $\mathrm{FPR}(c) = P(X \geqslant c)$, respectively. The ROC curve of the test is then the plot of the true positive rates *versus* the false positive rates with all possible threshold value c. The area under the ROC curve (AUC) is the probability that a randomly selected subject with the disease has a higher value of test result than that of a randomly chosen subject without the disease, i.e. $\mathrm{AUC} = P(Y > X)$ [11]. The AUC is a very useful summary measure of diagnostic accuracy since it describes a test's inherent ability of discrimination between diseased and non-diseased subjects.

In multi-reader ROC studies, a subject is examined by several readers with multiple diagnostic tests. Without loss of generality, we assume that a subject is examined by two diagnostic tests and each test is interpreted by several readers. Let X_{it}^k denote the test result of non-diseased subject i obtained from reader k with test t, and Y_{jt}^k denote the test result of diseased subject j obtained from reader k with test t, where $i = 1, \ldots, m$, $j = 1, \ldots, n$, $k = 1, \ldots, K$ and $t = 1, 2$. Let $X_i^k = (X_{i1}^k, X_{i2}^k)'$, $Y_j^k = (Y_{j1}^k, Y_{j2}^k)'$.

Assume that the diagnostic test t has an intrinsic but unknown AUC, say θ_t, where $t = 1, 2$. Let Δ denote the difference between two AUCs, $\Delta = \theta_1 - \theta_2$. Further assume that the AUC of test t from each reader reflects the intrinsic AUC of test t. However, as discussed previously, the readers' scales of interpretations on test results might not agree very well. To adjust this heterogeneity among the readers, we propose to compare the AUCs of two diagnostic tests from each reader and then estimate Δ by combining all the individually estimated Δs. For the kth reader, we consider a U-statistic [12] with kernel ϕ by extending the method of DeLong *et al.* [13]

$$U^k = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i^k, Y_j^k) \quad (k = 1, \ldots, K)$$

where

$$\phi(X_i^k, Y_j^k) = \psi(X_{i1}^k, Y_{j1}^k) - \psi(X_{i2}^k, Y_{j2}^k)$$

and $\psi(X,Y)=1$ if $Y>X$, $\psi(X,Y)=1/2$ if $Y=X$, and 0 otherwise. Consequently,

$$\phi(X_i^k, Y_j^k) = \begin{cases} 1 & \text{if } Y_{j1}^k > X_{i1}^k \text{ and } Y_{j2}^k < X_{i2}^k \\ \frac{1}{2} & \text{if } (Y_{j1}^k > X_{i1}^k \text{ and } Y_{j2}^k = X_{i2}^k) \text{ or } (Y_{j1}^k = X_{i1}^k \text{ and } Y_{j2}^k < X_{i2}^k) \\ 0 & \text{if } (Y_{j1}^k > X_{i1}^k \text{ and } Y_{j2}^k > X_{i2}^k) \text{ or } (Y_{j1}^k = X_{i1}^k \text{ and } Y_{j2}^k = X_{i2}^k) \\ & \text{or } (Y_{j1}^k < X_{i1}^k \text{ and } Y_{j2}^k < X_{i2}^k) \\ -\frac{1}{2} & \text{if } (Y_{j1}^k = X_{i1}^k \text{ and } Y_{j2}^k > X_{i2}^k) \text{ or } (Y_{j1}^k < X_{i1}^k \text{ and } Y_{j2}^k = X_{i2}^k) \\ -1 & \text{if } Y_{j1}^k < X_{i1}^k \text{ and } Y_{j2}^k > X_{i2}^k \end{cases}$$

It is easy to see that $\phi(X_i^k, Y_j^k)$ is a function of two real variables and $E[\phi(X_i^k, Y_j^k)]=\Delta$. Then by definition, U^k is a U-statistic and from the theory of U-statistic, if $E[\phi^2(X_i^k, Y_j^k)]<\infty$ and $m/n\to\lambda>0$ as $m,n\to\infty$, we have that $\sqrt{m}(U^k-\Delta)$ is asymptotically normally distributed with mean zero and variance $\tau_{10}^{kk}+\lambda\tau_{01}^{kk}$ as $m/n\to\lambda>0, m,n\to\infty$, where $\tau_{10}^{kk}=E[\phi(X_i^k, Y_j^k)\phi(X_i^k, Y_h^k)]-\Delta^2$, $j\neq h$, and $\tau_{01}^{kk}=E[\phi(X_i^k, Y_j^k)\phi(X_h^k, Y_j^k)]-\Delta^2$, $i\neq h$.

Similarly, when K readers examine the same sample of subjects with two diagnostic tests, we have K test statistics denoted as $U=(U^1,\ldots,U^K)'$. Under the conditions of $E[\phi^2(X_i^k, Y_j^l)]<\infty$ and $m/n\to\lambda>0$ as $m,n\to\infty$ for all $k=1,\ldots,K$, $l=1,\ldots,K$, by applying multivariate central-limit theory, one can show that $\sqrt{m}(U-B)$ converges in distribution to a multivariate normal vector with mean zero vector and covariance matrix $\Lambda=((\gamma_{kl}))$, as $m/n\to\lambda$, $m,n\to\infty$, where B is a K-dimensional vector with each element being Δ and $\gamma_{kl}=\tau_{10}^{kl}+\lambda\tau_{01}^{kl}$ is the (k,l)th element of matrix Λ.

The overall comparison of the relative accuracies of two diagnostic tests can be obtained by combining all the individual $U^k s$ along the line of Wei and Johnson [9]. Specifically,

$$D=\left(\sum_{k=1}^{K} w_k\right)^{-1}\sum_{k=1}^{K} w_k U^k=\frac{W'U}{W'\mathbf{1}}$$

where $\mathbf{1}$ is a K-dimensional vector with each element being 1 and W is a K-dimensional vector of weights. It is easy to see that D has an asymptotic normal distribution with mean Δ and variance $W'\Lambda W/(W'\mathbf{1})^2$. If Λ is positive-definite, the statistic $(D-\Delta)(W'\Lambda W/(W'\mathbf{1})^2)^{-1/2}$ converges in distribution to a standard normal random variable, as $m/n\to\lambda$ and $m,n\to\infty$.

Two common choices of the weights $W=(w_1,\ldots,w_K)'$ are given by $w_k=1/K$, $k=1,\ldots,K$ to give each reader equal weight and by $W=\Lambda^{-1}\mathbf{1}$ to maximize the local power of the hypothesis test. For example, if the hypothesis we wish to test is $H_0:\Delta=0$ *versus* $H_a:\Delta>0$, test statistic $[D(W'\Lambda W/(W'\mathbf{1})^2)^{-1/2}]$ follows a standard normal distribution asymptotically under H_0. With the fixed type I error α, we show that the local power of the test can be maximized by choosing $W=c\Lambda^{-1}\mathbf{1}$ for any constant $c\neq 0$, see Appendix A.

In practice, the covariance matrix of U is unknown. A consistent estimate of the covariance matrix Λ, can be obtained with the method of structural components proposed by Sen [14].

In summary, τ_{10}^{kl} and τ_{01}^{kl} can be estimated as

$$\hat{\tau}_{10}^{kl} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^{k}(X_i) - U^{k}][V_{10}^{l}(X_i) - U^{l}]$$

$$\hat{\tau}_{01}^{kl} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^{k}(Y_j) - U^{k}][V_{01}^{l}(Y_j) - U^{l}]$$

where the X-component $V_{10}^{k}(X_i)$ and Y-component $V_{01}^{k}(Y_j)$ are defined as

$$V_{10}^{k}(X_i) = \frac{1}{n}\sum_{j=1}^{n}\phi(X_i^{k}, Y_j^{k}) \quad (i = 1, 2, \ldots, m)$$

$$V_{01}^{k}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\phi(X_i^{k}, Y_j^{k}) \quad (j = 1, 2, \ldots, n)$$

When the accuracy of a single diagnostic test in multi-reader ROC study is of interest, proposed method can be adapted to obtain the summary accuracy measure of test by combining the accuracy measures from different readers. The estimate of AUC of diagnostic test in this way remains unbiased while the AUC under the average ROC curve tends to be biased when readers have effects on the test results.

Let $\hat{A}_t^{k}$ denote the estimated AUC of the diagnostic test t from the kth reader and recall that θ_t denote the unknown intrinsic AUC of the diagnostic test t. By exploiting the non-parametric method of DeLong *et al.* [13], the $\hat{A}_t^{k}$ can be obtained by

$$\hat{A}_t^{k} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(X_{it}^{k}, Y_{jt}^{k}) \quad k = 1, \ldots, K$$

where $\psi(X, Y)$ is as defined previously. For K readers, it is easy to see that the vector $\hat{A}_t = (\hat{A}_t^{1}, \ldots, \hat{A}_t^{K})'$ is a vector of U-statistics which converges in distribution to a multivariate normal random vector with mean vector A_t and covariance matrix Σ, where $A_t = (\theta_t, \theta_t, \ldots, \theta_t)'$, Σ can be estimated by the method of structural components developed by Sen [14], as shown previously.

A summary accuracy measure for single diagnostic test t with multiple readers is obtained by constructing a linear combination Q of the statistics $\hat{A}_t^{k}s\,(k = 1, \ldots, K)$, where $Q = (\sum_{k=1}^{K} w_k)^{-1}\sum_{k=1}^{K} w_k\hat{A}_t^{k} = (W'\mathbf{1})^{-1}(W'\hat{A}_t)$. Again two common choices of weights W are given by $w_k = \frac{1}{K}, (k = 1, \ldots, K)$ which gives equal weight to each reader and by $\hat{\Sigma}^{-1}\mathbf{1}$ which minimize the variance of Q. If Σ is positive-definite, then statistic $(Q - \theta_t)(W'\hat{\Sigma}W/(W'\mathbf{1})^2)^{-1/2}$ converges in distribution to a standard normal random variable, as $m, n \to \infty$.

3. APPLICATION

We apply the proposed method to analyse the data of Muller *et al.* [10] who conducted an experiment to investigate the optimization of the operating parameters for photographic

Table I. Comparison of diagnostic accuracy of two modalities in radiology example.

Parameter	Proposed method with equal weights	Proposed method with optimal weights	Method of average ROC curves
$\hat{\theta}_1$ (se)	0.743(0.024)	0.745(0.024)	0.619(0.012)
$\hat{\theta}_2$ (se)	0.805(0.020)	0.813(0.020)	0.680(0.039)
$\hat{\Delta}$ (se)	−0.062(0.019)	−0.063(0.018)	−0.061(0.023)
p-value	0.0009	0.0005	0.0040

images used in scintigraphy. In the experiment, 50 plates were used and a copper disc was placed in random position on each plate representing a lesion. Then a lead strip was placed on the plate along the diameter such that the plate was divided in half and only one half has the lesion. Thus, 50 half plates with lesions and 50 half plates without lesions were obtained to represent 50 diseased subjects and 50 non-diseased subjects, respectively. Muller *et al.* constructed two images for each subject using different operating parameters called modalities and asked three readers to analyse the images. Each of three readers read all images from the 100 subjects and determined whether the lesion exists or not with a five-point confidence scale with 1 indicating definitely not exist and 5 indicating definitely exist. For the details of the experiment, see References [6, 10].

The primary interest of the experiment is to find a modality that allows users to discriminate the diseased subjects from the non-diseased ones with higher accuracy. We estimated the AUC of each modality and performed a one sided test using proposed method and the method of average ROC curve. Table I displays the estimates calculated from data, where $\hat{\theta}_1$ and $\hat{\theta}_2$ denote the estimated AUCs for modality 1 and 2, respectively, and $\hat{\Delta}$ denotes the estimated difference between two AUCs. From Table I, we can see that both the proposed method and the method of average ROC curve lead to the conclusion that modality 2 is more accurate than modality 1 significantly. Pepe [6] analysed this data set by using a logistic type regression model with modalities and readers as covariates and reached a similar conclusion with p-value $= 0.06$ for the coefficient of modality.

4. SIMULATION STUDY

Simulation studies were conducted to assess the performance of the proposed method. Results were compared with those based on the average ROC curve of Lee and Rosner [10]. We generated 1000 data sets with the scenario that there are three readers who examine each subject with two different diagnostic tests.

When assessing the performance of the proposed method for comparing the accuracies of two diagnostic tests, we studied three parts: (1) when the scales of readers' interpretations do not agree very well, type I error and power of the proposed approach as well as the coverage of the 95 per cent CI for the true difference in AUCs were assessed and compared with those obtained from the method of average ROC curve; (2) when the scales of readers' interpretations agree very well with each other and sample sizes are small, the efficiency of the two methods were assessed; (3) when the scales of readers' interpretations agree with

each other, the performance of the proposed method with equal weights and with optimal weights were studied.

(1) For each non-diseased subject $i(i=1,\ldots,100)$, a random vector

$$X_i^{(6)}=(X_i^{1(t1)},X_i^{2(t1)},X_i^{3(t1)},X_i^{1(t2)},X_i^{2(t2)},X_i^{3(t2)})'$$

was generated from a 6-variate normal distribution with a mean vector $\mu_0^{(6)}=(\mu_0^{1(t1)},\mu_0^{2(t1)},\mu_0^{3(t1)},\mu_0^{1(t2)},\mu_0^{2(t2)},\mu_0^{3(t2)})'$ and a 6×6 covariance matrix Σ^a having a compound symmetric structure. For each diseased subject $j(j=1,\ldots,100)$, a random vector $Y_j^{(6)}=(Y_j^{1(t1)},Y_j^{2(t1)},Y_j^{3(t1)},Y_j^{1(t2)},Y_j^{2(t2)},Y_j^{3(t2)})'$ was generated from a 6-variate normal distribution with a mean vector $\mu_1^{(6)}=(\mu_1^{1(t1)},\mu_1^{2(t1)},\ \mu_1^{3(t1)},\mu_1^{1(t2)},\mu_1^{2(t2)},\mu_1^{3(t2)})'$ and the same covariance matrix Σ^a. Note that the first three elements in the vectors $X_i^{(6)}$ or $Y_j^{(6)}$ are the observations obtained from three readers with test 1, while the last three elements in the vectors are the observations obtained from the same three readers with test 2. The covariance matrix was specified as follows: $\mathrm{diag}(\Sigma^a)=(100,100,100,4,4,4)$; the correlation between the test results of a subject from same reader but different tests was set to be 0.5; for results of a subject from same test but different readers, correlation was set to be 0.4; and correlation for results of a subject from different readers and different tests was set to be 0.2. When assessing type I error, we let $\mu_1^{(6)}-\mu_0^{(6)}=(11.9,11.9,11.9,2.38,2.38,2.38)$ so the true AUCs of two diagnostic tests are equal to 0.8. A two sided test was performed and the significance level was set to be 0.05. When assessing the powers of the tests and the coverages of 95 per cent CIs, we set $\mu_1^{(6)}-\mu_0^{(6)}=(11.9,11.9,11.9,1.483,1.483,1.483)$ which makes the true AUCs of two diagnostic tests being 0.8 and 0.7, respectively. The type I error, power and coverage of 95 per cent CIs of the estimates were also assessed when the test results are in ordinal scale. To generate the ordinal data, we first simulated data from normal distributions, then assigned value v as test result if the generated data was between pre-specified cutpoints c_{v-1} and c_v, where $v=1,\ldots,10$. The cutpoints c_v were chosen to yield equal probability for each category for a non-diseased subject and for diagnostic tests 1 and 2, the vectors of cutpoints were $C^1=(-\infty,33.55,37.15,41.55,44.80,50.00,55.20,58.45,62.85,66.45,\infty)$ and $C^2=(-\infty,1.71,2.43,3.31,3.96,5.00,6.04,6.69,7.57,8.29,\infty)$, respectively. Note that the choice of cutpoints does not affect the resulting AUC if the original normal distributions are fixed.

Hypothesis tests were performed based on the proposed method with equal weight for each reader. Table II summarizes the type I error and the bias of the estimated difference between the AUCs of two diagnostic tests. Table III summarizes the power, the bias of the estimated differences between two AUCs and the coverages of the 95 per cent CI for the true difference. The results show that when readers' scales agree very well, both approaches have appropriate type I errors, powers and coverages for true difference. However, when readers' scales do not agree very well, the approach based on the average ROC curve has larger type I error, smaller power and insufficient coverage, and when disagreements increase, the type I error increases and the power and coverage decrease. While the test based on the proposed approach always has appropriate size, power and coverage.

(2) We studied the performance of the proposed method and the method of average ROC curve when readers' scales agree with each other with small and moderate sample sizes. A one sided test was conducted and the significance level was still set as 0.05. Tables IV and V

Table II. Comparing two diagnostic tests: type I error.

Data structure μ_0	Combining dependent tests Size (%)	Combining dependent tests Bias ($\times 10^2$)	Average ROC curves Size (%)	Average ROC curves Bias ($\times 10^2$)
Continuous test results				
(50, 50, 50, 5, 5, 5)	6.1	0.13	3.9	−0.11
(45, 50, 55, 5, 5, 5)	4.5	−0.13	10.9	−0.18
(43, 50, 57, 5, 5, 5)	5.2	0.10	27.9	−3.46
Ordinal test results				
(50, 50, 50, 5, 5, 5)*	4.1	0.03	2.7	0.05
(45, 50, 55, 5, 5, 5)*	4.1	0.29	9.0	−1.85
(43, 50, 57, 5, 5, 5)*	5.6	0.54	25.7	−3.51

*Distribution mean of continuous latent variable for ordinal test result.

Table III. Comparing two diagnostic tests: power (true difference = 0.1).

Data structure μ_0	Combining dependent tests Power (%)	Combining dependent tests Bias ($\times 10^2$)	Combining dependent tests Coverage (%)	Average ROC curves Power (%)	Average ROC curves Bias ($\times 10^2$)	Average ROC curves Coverage (%)
Continuous test results						
(50, 50, 50, 5, 5, 5)	96.0	−0.03	95.6	96.3	0.06	97.8
(45, 50, 55, 5, 5, 5)	95.6	−0.09	94.1	86.3	−1.84	90.8
(43, 50, 57, 5, 5, 5)	96.9	0.07	94.8	69.2	−3.45	75.5
Ordinal test results						
(50, 50, 50, 5, 5, 5)*	96.1	−1.11	91.6	95.9	−0.12	97.4
(45, 50, 55, 5, 5, 5)*	97.0	−0.84	93.8	85.7	−1.24	94.2
(43, 50, 57, 5, 5, 5)*	97.6	−0.62	93.8	66.1	−3.83	33.9

*Distribution mean of continuous latent variable for ordinal test result.

Table IV. Readers have no effect (sample size per group = 25).

True diff.	Combining dependent tests Bias ($\times 10^2$)	Combining dependent tests SE ($\times 10^2$)	Combining dependent tests Coverage (%)	Combining dependent tests Power (%)	Average ROC curves Bias ($\times 10^2$)	Average ROC curves SE ($\times 10^2$)	Average ROC curves Coverage (%)	Average ROC curves Power (%)
0.00	0.045	4.85	93.7	5.6(size)	0.049	4.82	97.9	3.5(size)
0.05	−0.139	5.06	93.9	25.2	−0.163	5.01	97.6	16.8
0.10	−0.215	5.49	93.6	58.8	−0.228	5.48	96.6	50.6
0.15	−0.011	5.43	96.4	85.1	−0.048	5.40	98.2	79.5
0.20	0.148	6.00	93.3	96.5	0.107	5.96	97.5	94.6

summarize the type I error, power, bias and coverage from two methods when sample sizes are 25, 50 subjects per group, respectively. The results show that the difference between these two methods are very small when the scales of readers agree: both approaches are unbiased;

Table V. Readers have no effect (sample size per group = 50).

	Combining dependent tests				Average ROC curves			
True Diff.	Bias ($\times 10^2$)	SE ($\times 10^2$)	Coverage (%)	Power (%)	Bias ($\times 10^2$)	SE ($\times 10^2$)	Coverage (%)	Power (%)
0.00	−0.055	3.39	94.3	4.9(size)	−0.054	3.38	96.4	4.2(size)
0.05	−0.180	3.74	94.2	40.6	−0.168	3.72	96.3	36.5
0.10	0.173	3.90	94.4	85.6	0.167	3.88	97.1	83.3
0.15	0.204	3.92	94.6	99.0	0.204	3.90	97.6	98.5
0.20	0.206	4.06	94.7	100.0	0.189	4.04	99.1	99.9

Table VI. Performance of proposed method with equal weights and optimal weights (sample size per group = 25).

	Equal weights				Optimal weights			
True Diff.	Bias ($\times 10^2$)	SE ($\times 10^2$)	Coverage (%)	Power (%)	Bias ($\times 10^2$)	SE ($\times 10^2$)	Coverage (%)	Power (%)
0.00	0.175	5.91	94.3	5.7(size)	0.141	5.69	95.2	4.8(size)
0.05	−0.150	5.66	95.6	15.9	−0.190	5.46	95.3	16.3
0.10	−0.149	5.35	94.9	44.7	−0.109	5.13	94.1	49.5
0.15	0.177	5.15	94.4	83.8	0.085	4.87	95.4	85.1
0.20	−0.270	4.90	93.6	98.1	−0.287	4.57	93.7	99.5

hypothesis tests from two methods have similar type I error and power; the standard errors of two estimators are very close. This implies that two methods perform similarly when the scales of readers' interpretations agree with each other.

(3) The performance of the proposed method with equal weights and with optimal weights were also assessed with sample size being 25 subjects per group. A two sided test was conducted with 0.05 significance level. Table VI summarizes the type I error, power, bias and coverage from the proposed method with equal weights and with optimal weights. It shows that the two different choices of weights lead to similar bias, coverage and size. It also shows that the method with optimal weights performs better in terms of power. However, the improvement is very moderate.

For a single diagnostic test, the estimate of AUC and the coverage of the 95 per cent CI for the true AUC were assessed. For each non-diseased subject i ($i = 1, \ldots, 100$), a random vector $X_i^{(3)} = (X_i^1, X_i^2, X_i^3)'$ was generated from a 3-variate normal distribution with a mean vector $\mu_0^{(3)} = (\mu_0^1, \mu_0^2, \mu_0^3)'$ and a covariance matrix Σ^a having a compound symmetric structure with $\text{diag}(\Sigma^a) = (100, 100, 100)$ and a common correlation 0.4. Here X_i^k can be considered as an observation from the kth reader ($k = 1, 2, 3$). When readers' scales agree with each other, $\mu_0^k s$ were set to be equal, otherwise, they were set to be different to indicate that test responses depend on both true diseased status and readers. Similarly, for each diseased subject j ($j = 1, \ldots, 100$), a random vector $Y_j^{(3)} = (Y_j^1, Y_j^2, Y_j^3)'$ was generated from a 3-variate normal

Table VII. Single diagnostic test; true AUC = 0.8.

Data structure μ_0	Combining dependent AUCs			Average ROC curves		
	Bias ($\times 10^2$)	Se ($\times 10^2$)	Coverage (%)	Bias ($\times 10^2$)	Se ($\times 10^2$)	Coverage (%)
(50, 50, 50)	0.026	2.27	95.1	0.014	2.39	94.6
(45, 50, 55)	0.014	2.27	95.5	−1.811	2.19	89.2
(43, 50, 57)	0.047	2.35	93.9	−3.400	2.18	67.6
(40, 50, 60)	0.020	2.33	94.3	−5.993	2.03	14.5

distribution with a mean vector $\mu_1^{(3)} = \mu_0^{(3)} + \delta$ and same covariance matrix Σ^a, where the constant δ was chosen to have a true AUC of 0.8.

The AUC was estimated from the proposed method with equal weight to each reader. Table VII summarizes the bias of the estimated AUCs, the empirical standard errors of the estimates, and the coverages of 95 per cent CIs for the true AUC. The results were compared to those obtained from the average ROC curve. The results show that: (1) when the scales of readers' interpretation agree, both estimators are unbiased and have appropriate coverage for the true AUC; (2) when the scales of readers' interpretations do not agree very well, the estimator from the proposed method remains unbiased and has appropriate coverage, however, the estimator based on the average ROC curve has a increasing bias and decreasing coverage when disagreements are increased (Table VII).

5. DISCUSSION

The proposed method accommodates the question of comparing diagnostic accuracies while adjusting the possible heterogeneity among readers. Numerical studies show that when there is heterogeneity among readers, the proposed test yields appropriate size and power while the test based on the existing non-parametric method have inappropriate size and insufficient power. When readers' interpretation scales are homogeneous, the proposed method has the similar efficiency as the existing method. Although we are using this method in multi-reader studies, it is applicable generally for the diagnostic accuracy study in which repeated measurements for the same characteristic are obtained under different conditions for each subject.

When some values of sensitivity or specificity are not acceptable in practice, partial AUC is more appropriate to use for evaluating the performance of diagnostic tests. The proposed approach can be easily extended to compare partial AUCs by extending the method of Zhang *et al.* [15]. Suppose that we are interested in comparing the partial AUCs over a range of specificity which is no less than r, the expectation of statistic U^K in Section 2 will be the difference between two partial AUCs over the pre-specified range of specificity if we let $\psi(X_{it}^k, Y_{jt}^k) = 1$ if $Y_{jt}^k > X_{it}^k$ and $X_{it}^k > c_t^k$, $\psi(X_{it}^k, Y_{jt}^k) = 1/2$ if $Y_{jt}^k = X_{it}^k$ and $X_{it}^k > c_t^k$, and 0 otherwise, where the c_t^k is determined by $P(X_t^k \leqslant c_t^k) = r$.

In addition, if it is desired to compare the ROC curves instead of AUCs with covariate adjustment, then the ROC regression approach is useful and we believe that more research is needed to improve the ROC regression approach.

APPENDIX A: OPTIMAL WEIGHT

The optimal weight in the comparison of two diagnostic tests can be obtained to maximize the local power. Under the null hypothesis $H_0 : \Delta = 0$, test statistic $D(W'\Lambda W/(W'\mathbf{1})^2)^{-1/2}$ is an asymptotic standard normal random variable. With the fixed type I error, say α, the power of the test under the alternative $H_a : \Delta > 0$ can be expressed as

$$\begin{aligned}\text{Power} &= P\left(D\left(\frac{W'\Lambda W}{(W'\mathbf{1})^2}\right)^{-1/2} \geqslant Z_\alpha | \Delta > 0\right) = P\left((D-\Delta+\Delta)\left(\frac{W'\Lambda W}{(W'\mathbf{1})^2}\right)^{-1/2} \geqslant Z_\alpha\right) \\ &= P\left(Z + \Delta\left(\frac{(W'\mathbf{1})^2}{W'\Lambda W}\right)^{1/2} \geqslant Z_\alpha\right)\end{aligned}$$

where Z follows a standard normal distribution and Z_α satisfy $P(Z \geqslant Z_\alpha) = \alpha$. From Cauchy–Schwarz inequality, if Λ is positive definite, then

$$\frac{(W'\mathbf{1})^2}{W'\Lambda W} \leqslant \mathbf{1}'\Lambda^{-1}\mathbf{1}$$

with equality if and only if $W = c\Lambda^{-1}\mathbf{1}$ for any constant c. Therefore, the weights $W = c\Lambda^{-1}\mathbf{1}$ will give the maximum value of $\Delta((W'\mathbf{1})^2/W'\Lambda W)^{1/2}$ and maximize the local power of the test.

ACKNOWLEDGEMENTS

We thank the editor and two referees for the helpful comments. This research is partially supported by New York City Council Speaker's Fund and a National Science Foundation Career Award.

REFERENCES

1. Zhou X, Obuchowski N, McClish S. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
2. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the Jackknife method. *Statistics in Radiology* 1992; **27**:723–731.
3. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Communications in Statistics-Simulation and Computation* 1995; **24**(2):285–308.
4. Tosteson A, Begg C. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988; **8**:204–215.
5. Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996; **15**:1807–1826.
6. Pepe MS. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* 1997; **84**(3):595–608.
7. Cai T, Pepe MS. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* 2002; **97**(460):1099–1107.
8. Lee MT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample wilcoxon statistics. *Applied Statistics* 2001; **50**(Part 3):337–344.
9. Wei LJ, Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985; **72**(2):359–364.
10. Muller C, Wasserman HJ, Erlank P, Klopper JK, Morkel HR, Ellmann A. Optimisation of density and contrast yielded by multiformat photographic imagers used for scintigraphy. *Physics in Medicine and Biology* 1989; **34**:473–481.
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:68–78.

12. Serfling RJ. *Approximation Theorems of Mathematical Statistics*. Wiley: New York, 1980.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
14. Sen PK. On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* 1960; **10**:1–18.
15. Zhang D, Zhou X, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 2002; **21**:701–715.

