Biometrics 70, 1033–1041
December 2014
DOI: 10.1111/biom.12240

Power Calculation for Comparing Diagnostic Accuracies in a Multi-Reader, Multi-Test Design

Eunhee Kim,[1,*] Zheng Zhang,[1] Youdan Wang,[1] and Donglin Zeng[2]

[1]Department of Biostatistics and Center for Statistical Sciences, Brown University, Providence, Rhode Island, U.S.A.

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, U.S.A.

[*]*email:* ekim@stat.brown.edu

Summary. Receiver operating characteristic (ROC) analysis is widely used to evaluate the performance of diagnostic tests with continuous or ordinal responses. A popular study design for assessing the accuracy of diagnostic tests involves multiple readers interpreting multiple diagnostic test results, called the multi-reader, multi-test design. Although several different approaches to analyzing data from this design exist, few methods have discussed the sample size and power issues. In this article, we develop a power formula to compare the correlated areas under the ROC curves (AUC) in a multi-reader, multi-test design. We present a nonparametric approach to estimate and compare the correlated AUCs by extending DeLong et al.'s (1988, *Biometrics* **44**, 837–845) approach. A power formula is derived based on the asymptotic distribution of the nonparametric AUCs. Simulation studies are conducted to demonstrate the performance of the proposed power formula and an example is provided to illustrate the proposed procedure.

Key words: Multi-reader, multi-test design; Power; Receiver operating characteristic curve; Sample size; *U*-statistics.

1. Introduction

The receiver operating characteristic (ROC) curve is a standard tool used to evaluate the performance of a diagnostic test when test results are continuous or ordinal (Metz, 1978; Hanly and McNeil, 1982; Swets and Pickett, 1982). In an ROC curve, the true positive rate is plotted as a function of the false positive rate across all possible cut-points. The area under the ROC curve (AUC) is a commonly used summary measure of diagnostic accuracy. Values of AUC close to 1.0 indicate that the test result has high diagnostic accuracy, and relative accuracies of diagnostic tests can be compared through their corresponding AUCs.

One important objective in many diagnostic studies is to examine whether new diagnostic tests provide performance that is superior to that of conventional tests for a certain condition or disease. The comparison of diagnostic accuracies often depends on the subjective interpretation of readers in diagnostic evaluation, so studies of such diagnostic tests usually involve multiple readers. The multi-reader, multi-test design is most likely to be employed in such settings in which multiple readers interpret all test results from a sample of patients who undergo multiple diagnostic tests. This design is efficient for comparing tests because it requires the smallest patient population; hence, it needs fewer interpretations per reader versus other study designs (Zhou, Obuchowski, and McClish, 2002). Sample size and power calculations are crucial to develop a multi-reader, multi-test ROC study. Although several statistical procedures to analyze data from this design have been developed, few methods have discussed the power considerations.

Most of the existing literature for analyzing multi-reader, multi-test studies has applied mixed-effects analysis of variance (ANOVA) models (e.g., Dorfman, Berbaum, and Metz, 1992; Obuchowski and Rockette, 1995; Beiden, Wagner, and Campbell, 2000; Obuchowski et al., 2004). In particular, the methods proposed by Dorfman et al. (1992) and Obuchowski and Rockette (1995) are widely used, often referred to as the Dorfman–Berbaum–Metz (DBM) and Obuchowski–Rockette (OR) methods, respectively. The DBM method uses a mixed-effects ANOVA model on the jackknife pseudo-values of the summary measures of the ROC curve. It assumes that readers and patients are random factors and tests are a fixed factor and that the random effects and error term in the model follow independent normal distributions. The DBM approach carries some concerns (Zhou et al., 2002; Hillis et al., 2005; Song and Zhou, 2005). One weakness of this approach is that the ANOVA model for pseudo-values does not have a straightforward interpretation as the jackknife pseudo-values in this model are treated as observed data. Second, the DBM method does not generally satisfy the regular assumptions for standard mixed-effects ANOVA models because the variance of the response variable may vary across tests and subjects and thus might lead to erroneous inferences. Furthermore, it is substantially conservative and not based on a satisfactory conceptual or theoretical model. Recently, solutions to various drawbacks of the original DBM method have been discussed in the literature (Hillis et al., 2005; Hillis, 2007; Hillis, Berbaum, and Metz, 2008).

On the other hand, the OR method applies a mixed-effects ANOVA model to the estimated summary measures of the ROC curve for each combination of readers and tests, where tests are considered fixed and readers are considered random. For hypothesis testing, an adjusted ANOVA F-test is used to correct for the correlations between and within readers. The

OR approach also makes strong assumptions as follows. First, the validity of the method depends on the assumptions about the underlying distributions of the random variables. Second, the complicated correlation structure arising from having the same patient sample evaluated by several readers in a set of tests is overly simplified by the three different correlations. Furthermore, it is not clear how well the adjusted F statistic follows an F distribution, especially in small samples.

Hillis et al. (2005) demonstrated that the DBM and OR approaches yield the identical test statistic when the same accuracy measure and covariance estimation methods are used, but inferences depend on the denominator degrees of freedom (ddf) method, DBM, or OR used. Hillis (2007) later pointed out problems with the OR and DBM ddf methods: The original OR method is very conservative with significance levels considerably below the nominal level while the DBM method can result in extremely wide confidence intervals because the ddf can be close to zero; he proposed using a new ddf estimator that overcomes these problems and described how the new ddf can be used with either the DBM or the OR procedure.

While the OR and DBM methods make use of the mixed-effects ANOVA model, several nonparametric approaches to the multi-reader ROC analysis have been developed (e.g., DeLong, DeLong, and Clarke-Pearson, 1988; Song, 1997; Gallas, 2006). Above all, DeLong et al. (1988) proposed a nonparametric approach to compare the correlated ROC areas using the theory of the U-statistics. Their method, however, applies only to cases in which each patient is interpreted using multiple tests or repeatedly examined using a single test; thus, it is not appropriate for the data from a multi-reader, multi-test study design. Song (1997) generalized DeLong et al.'s method for analysis of multi-reader, multi-test ROC data and proposed the jackknife method to estimate the variance of the U-statistics. Song's variance estimation using the jackknife methods can be computationally demanding and the article does not present how it performs with an unequal number of normal and diseased cases. It should be noted that when DBM and OR methods are used by treating readers as fixed effects and by applying Delong et al.'s variance estimation methods, their test statistics are equivalent to those by Delong at al. and Song's 2-sample jackknife approaches. More recently, Li and Zhou (2008) proposed a nonparametric approach to compare ROC curves for a paired design with repeated or clustered data. They treated nonparametric ROC curves as stochastic processes and derived their asymptotic distribution theory. A Monte Carlo resampling method was used to approximate the empirical ROC processes and compare correlated AUCs. Although their method was not specifically developed for multi-reader diagnostic accuracy studies, it can be applied to such studies when repeated marker measurements from each subject stem from interpretation by multiple readers.

With regard to the power calculation for multi-reader diagnostic accuracy studies, a formula based on the OR method is among the most widely used approaches (Obuchowski, 1995a, 1995b, 1998; Zhou et al., 2002). Obuchowski (1995a) used the adjusted ANOVA F statistic of the OR method to determine the power to test the equality of the diagnostic accuracies of multiple tests. Possible ranges for the parameter estimates of the variance components and correlations required for the sample size calculation were introduced in Obuchowski (1995b). The author's nonparametric power calculation was also described in Obuchowski (1998), but this article discussed several other methods for determining the same sizes that differ by study designs (other than multi-reader ROC studies) and diagnostic accuracy measures. Recently, Hillis, Obuchowski, and Berbaum (2011) described the procedure for estimating power that can be analyzed using either the DBM or OR method by applying the Hillis's (2007) recommended ddf for the F-statistic. In this approach, the ROC summary measure is estimated in advance and this estimate is used as a response in the second step of fitting a mixed-effects ANOVA model. This two-stage approach can be misleading because it may depend on the estimated values of the response accounted for in the ANOVA model and is sensitive to the normality assumptions about the underlying distributions of the random variables. Additionally, the method assumes that the complex correlation structure arising from having the same patient sample evaluated by several readers in a set of tests can be described by only three correlations; correlation of error terms in diagnostic accuracies of the same reader in different tests, the correlation of error terms in diagnostic accuracies of different readers in the same test, and the correlation of error terms in diagnostic accuracies of different readers in different tests.

In this article, we propose a new power formula to compare the correlated AUCs in a multi-reader, multi-test design. Specifically, we present a nonparametric approach to estimate and compare the correlated AUCs in a multi-reader, multi-test design by extending DeLong et al.'s (1988) approach. A power formula is derived based on the asymptotic normality of the nonparametric AUCs. This article is organized as follows: Inference procedures for correlated AUCs are described in Section 2 and a formula for power calculation is presented in Section 3. In Section 4, we describe simulation studies to compare the powers of our proposed tests. We apply our proposed method to the example from the American College of Radiology Imaging Network (ACRIN) Digital Mammographic Imaging Screening Trial (DMIST) in Section 5 and a discussion is followed in Section 6.

2. Inference for Correlated AUCs

Suppose h tests are performed on a sample of N patients (m diseased, n non-diseased, $N = m + n$) where r readers independently examine the test results from all patients. Let X_{ik}^{l} be the test result for diseased subject i from test l by reader k ($i = 1, \ldots, m; k = 1, \ldots, r; l = 1, \ldots, h$). Similarly, let Y_{jk}^{l} be the test result for non-diseased subject j from test l by reader k ($j = 1, \ldots, n; k = 1, \ldots, r; l = 1, \ldots, h$). The test results X_{ik}^{l} and Y_{jk}^{l} can be either continuous or ordinal. Without loss of generality, we assume that higher values of test results are more indicative of disease. Our primary goal is to estimate and compare the correlated AUCs of h diagnostic tests.

Let θ_k^l denote the AUC of diagnostic test l by reader k. We assume that the ratings from diseased and non-diseased subjects are independent and have the same distribution in diseased or non-diseased subjects for a fixed reader and a test. A nonparametric AUC of θ_k^l is then calculated by the

Mann–Whitney U statistic

$$\hat{\theta}_k^l = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_{ik}^l, Y_{jk}^l), \tag{1}$$

with

$$\phi(X,Y) = \begin{cases} 1 & \text{if } X > Y, \\ 1/2 & \text{if } X = Y, \\ 0 & \text{if } X < Y. \end{cases}$$

To evaluate the diagnostic accuracy of test l, we use the average AUC across r readers from a fixed test l. Thus, the AUC for diagnostic test l is defined as $\theta^l = \frac{1}{r}\sum_{k=1}^{r}\theta_k^l$ and is estimated by

$$\hat{\theta}^l = \frac{1}{r}\sum_{k=1}^{r}\hat{\theta}_k^l = \frac{1}{mnr}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{r}\phi(X_{ik}^l, Y_{jk}^l). \tag{2}$$

In our nonparametric approach, readers are considered fixed effects as indicated in equation (2).

Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1^1, \hat{\theta}_2^1 \ldots, \hat{\theta}_r^h) = (\hat{\theta}_k^l)_{k=1,\ldots,r;l=1,\ldots,h}$ be a $rh \times 1$ vector of U-statistics in which each element represents the nonparametric AUC of a diagnostic test examined by a specific reader. We use the notation (k,l) to denote the value corresponding to the lth diagnostic test interpreted by the kth reader. Asymptotic normality and variance expression for $\hat{\boldsymbol{\theta}}$ can be derived using the theory of generalized U-statistics. Define

$$\begin{aligned} \xi_{10}^{((k,l),(k',l'))} &= \mathrm{Cov}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{ik'}^{l'}, Y_{j'k'}^{l'})], \ \ j \neq j' \\ \xi_{01}^{((k,l),(k',l'))} &= \mathrm{Cov}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{i'k'}^{l'}, Y_{jk'}^{l'})], \ \ i \neq i' \\ \xi_{11}^{((k,l),(k',l'))} &= \mathrm{Cov}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{ik'}^{l'}, Y_{jk'}^{l'})]. \end{aligned} \tag{3}$$

The covariance of the (k,l)th and (k',l')th statistic can be written as

$$\mathrm{Cov}(\hat{\theta}_k^l, \hat{\theta}_{k'}^{l'}) = \frac{(n-1)\xi_{10}^{((k,l),(k',l'))} + (m-1)\xi_{01}^{((k,l),(k',l'))}}{mn} + \frac{\xi_{11}^{((k,l),(k',l'))}}{mn}. \tag{4}$$

We extend DeLong et al.'s (1988) nonparametric approach to analyze multi-reader, multi-test ROC data. Specifically, we use a method of structural components to provide consistent estimates of the elements of the variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ proposed by Sen (1960). For the (k,l)th statistics, $\hat{\theta}_k^l$, the X-components and Y-components are defined, respectively, as

$$V_{10}^{(k,l)}(X_{ik}^l) = \frac{1}{n}\sum_{j=1}^{n}\phi(X_{ik}^l, Y_{jk}^l) \quad (i = 1, 2, \ldots, m)$$

and

$$V_{01}^{(k,l)}(Y_{jk}^l) = \frac{1}{m}\sum_{i=1}^{m}\phi(X_{ik}^l, Y_{jk}^l) \quad (j = 1, 2, \ldots, n).$$

Additionally, we define the $(rh \times rh)$ matrix S_{10} such that $((k,l),(k',l'))$th element is

$$S_{10}^{((k,l),(k',l'))} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^{(k,l)}(X_{ik}^l) - \hat{\theta}_k^l][V_{10}^{(k',l')}(X_{ik'}^{l'}) - \hat{\theta}_{k'}^{l'}]$$

and define $(rh \times rh)$ matrix S_{01} such that $((k,l),(k',l'))$th element is

$$S_{01}^{((k,l),(k',l'))} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^{(k,l)}(Y_{jk}^l) - \hat{\theta}_k^l][V_{01}^{(k',l')}(Y_{jk'}^{l'}) - \hat{\theta}_{k'}^{l'}].$$

The covariance matrix for $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1^1, \hat{\theta}_2^1 \ldots, \hat{\theta}_r^h)$ is then estimated by

$$S = \frac{1}{m}S_{10} + \frac{1}{n}S_{01}. \tag{5}$$

$S_{10}^{((k,l),(k',l'))}$ and $S_{01}^{((k,l),(k',l'))}$ are asymptotically unbiased estimates of $\xi_{10}^{((k,l),(k',l'))}$ and $\xi_{01}^{((k,l),(k',l'))}$, respectively, and the last term in equation (4) is negligible so that it is not considered in the covariance estimation (Serfling, 1980, Chapter 5).

THEOREM 1. *Let* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_k^l)_{k=1,\ldots,r;l=1,\ldots,h}$ *and* $\boldsymbol{\theta} = (\theta_k^l)_{k=1,\ldots,r;l=1,\ldots,h}$. *If* $\lim_{N\to\infty} m/N = \lambda$ *and* $\lim_{N\to\infty} n/N = 1-\lambda$ *with* $0 < \lambda < 1$, *then under model (1),* $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ *is asymptotically normal with zero mean vector and covariance matrix* $\boldsymbol{\Sigma} = (\sigma^{((k,l),(k',l'))})$, *where*

$$\sigma^{((k,l),(k',l'))} = \left[\frac{1}{\lambda}\xi_{10}^{((k,l),(k',l'))} + \frac{1}{(1-\lambda)}\xi_{01}^{((k,l),(k',l'))}\right].$$

The proof of Theorem 1 uses the central limit theorem for generalized two-sample U-statistics (Lee and Dehling, 2005). More details can be found in the appendix. The covariance matrix $\boldsymbol{\Sigma}$ in Theorem 1 can be estimated using its empirical counterpart in (5).

Next, we make an inference for the accuracy of each diagnostic test. The following theorem implies that $\hat{\theta}^l$ is a consistent estimator for θ^l and follows an asymptotically normal distribution.

COROLLARY 1. *Under the above assumptions of Theorem* 1, $\sqrt{N}(\hat{\theta}^l - \theta^l)$ *converges in distribution to a normal distribution with mean zero and variance*

$$\frac{1}{r^2}\sum_{k,k'=1}^{r}\sigma^{((k,l),(k',l))} = \frac{1}{r^2}\sum_{k,k'=1}^{r}\left[\frac{1}{\lambda}\xi_{10}^{((k,l),(k',l))} + \frac{1}{(1-\lambda)}\xi_{01}^{((k,l),(k',l))}\right].$$

See Web Appendix for the proof.

Let g be a real-valued function that is continuously differentiable with Jacobian matrix $G \equiv \partial g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}$ that has full rank. If $\lim_{N\to\infty} m/N = \lambda$ is bounded and nonzero, $\sqrt{N}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}))$ is asymptotically normal with zero mean vector and covariance matrix $G\boldsymbol{\Sigma}G^{\mathrm{T}}$. When g is a linear function and $\boldsymbol{C}$ is a

$(1 \times rh)$ row vector of coefficients, for any contrast $\boldsymbol{C\theta}$, the test statistic

$$\frac{\boldsymbol{C}\hat{\boldsymbol{\theta}} - \boldsymbol{C\theta}}{\left[\boldsymbol{C}\left(\frac{1}{m}S_{10} + \frac{1}{n}S_{01}\right)\boldsymbol{C}^{\mathrm{T}}\right]^{1/2}} \tag{6}$$

is asymptotically standard normal. For example, in order to evaluate the differences in the AUCs of the two tests $l = 1$ and $l' = 2$, we can conduct a Z-test by setting $\boldsymbol{C} = (\frac{1}{r}\mathbf{1}_{r\times 1}, -\frac{1}{r}\mathbf{1}_{r\times 1}, \mathbf{0}_{(h-2)r\times 1})^{\mathrm{T}}$ and $\boldsymbol{C\theta} = 0$ to the test statistic (6). To compare the AUCs of two or more diagnostic tests, we define $\boldsymbol{C}$ is a $(c \times rh)$ matrix of coefficients with full rank $c \le rh$. Then, the Wald statistic

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\left[\boldsymbol{C}\left(\frac{1}{m}S_{10} + \frac{1}{n}S_{01}\right)\boldsymbol{C}^{\mathrm{T}}\right]^{-1}\boldsymbol{C}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \tag{7}$$

can be used which has approximately a chi-squared distribution with c degrees of freedom. A confidence interval or region for $\boldsymbol{C}\hat{\boldsymbol{\theta}}$ can be easily calculated by using (6) or (7).

3. Power Calculations

In this section, we propose a formula for the power calculation of a multi-reader, multi-test study design. Specifically, we aim to determine the power to detect the difference in AUC between two diagnostic tests given a specific number of readers and subjects.

In order to test the AUC difference between any two diagnostic tests, we state the null and alternative hypotheses as

$$\mathrm{H}_0:\ \delta_0 = \theta^l - \theta^{l'} = 0 \quad \mathrm{H}_1:\ \delta_1 = \theta^l - \theta^{l'} \neq 0.$$

According to (6), the test statistic $((\hat{\theta}^l - \hat{\theta}^{l'}) - (\theta^l - \theta^{l'}))/\sqrt{\mathrm{Var}(\hat{\theta}^l - \hat{\theta}^{l'})}$ possesses approximately a standard normal distribution. Define

$$\begin{aligned}
\rho_{11} &= \rho_1^{((k,l),(k,l))} \quad \rho_{12} = \rho_1^{((k,l),(k',l))} \quad \rho_{13} = \rho_1^{((k,l),(k,l'))} \quad \rho_{14} = \rho_1^{((k',l'),(k',l'))} \\
\rho_{21} &= \rho_2^{((k,l),(k,l))} \quad \rho_{22} = \rho_2^{((k,l),(k',l))} \quad \rho_{23} = \rho_2^{((k,l),(k,l'))} \quad \rho_{24} = \rho_2^{((k',l'),(k',l'))} \\
&\qquad\quad \rho_{32} = \rho_3^{((k,l),(k',l))} \quad \rho_{33} = \rho_3^{((k,l),(k,l'))} \quad \rho_{34} = \rho_3^{((k',l'),(k',l'))},
\end{aligned} \tag{9}$$

$$\begin{aligned}
\rho_1^{((k,l),(k',l'))} &= \mathrm{Corr}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{ik'}^{l'}, Y_{j'k'}^{l'})],\ j \neq j' \\
\rho_2^{((k,l),(k',l'))} &= \mathrm{Corr}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{i'k'}^{l'}, Y_{jk'}^{l'})],\ i \neq i' \\
\rho_3^{((k,l),(k',l'))} &= \mathrm{Corr}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{ik'}^{l'}, Y_{jk'}^{l'})].
\end{aligned} \tag{8}$$

$\rho_1^{((k,l),(k',l'))}$, $\rho_2^{((k,l),(k',l'))}$, and $\rho_3^{((k,l),(k',l'))}$ $(k, k' = 1, \ldots, r; l, l' = 1, \ldots, h)$ represent the correlations between ϕ's when the test results from the two test modalities are obtained, respectively, from the same diseased and different non-diseased subjects, from different diseased and the same non-diseased subjects, and from the same diseased and non-diseased subjects. We derive the explicit expression for the variance of the difference between two nonparametric AUCs in the following theorem that depicts all correlation structures that arise from having the same readers or the same tests specified in (8).

THEOREM 2. *The variance of difference between two correlated nonparametric AUCs* $Var(\hat{\theta}^l - \hat{\theta}^{l'})$ *is*

$$\begin{aligned}
&\frac{1}{mnr^2}\sum_{t=l,l'}\left[\sum_{k=1}^{r} V_k^t(1 + (n-1)\rho_1^{((k,t),(k,t))} + (m-1)\rho_2^{((k,t),(k,t))})\right. \\
&+ \sum_{k\neq k'}\sqrt{V_k^t}\sqrt{V_{k'}^t}\big((n-1)\rho_1^{((k,t),(k',t))} + (m-1)\rho_2^{((k,t),(k',t))} \\
&\left. + \rho_3^{((k,t),(k',t))}\big)\right] - \frac{2}{mnr^2}\left[\sum_{k=1}^{r}\sqrt{V_k^l}\sqrt{V_k^{l'}}\big((n-1)\rho_1^{((k,l),(k,l'))}\right. \\
&+ (m-1)\rho_2^{((k,l),(k,l'))} + \rho_3^{((k,l),(k,l'))}\big) \\
&+ \sum_{k\neq k'}^{r}\sqrt{V_k^l}\sqrt{V_{k'}^{l'}}\big((n-1)\rho_1^{((k,l),(k',l'))} \\
&\left. + (m-1)\rho_2^{((k,l),(k',l'))} + \rho_3^{((k,l),(k',l'))}\big)\right],
\end{aligned}$$

where $V_k^l = \mathrm{Var}(\phi(X_{ik}^l, Y_{jk}^l)) = \theta_k^l - \frac{1}{4}\mathrm{Pr}(X_{ik}^l = Y_{jk}^l) - (\theta_k^l)^2$. *(See Web Appendix for details.)*

For power determination, the followings are assumed:

(a) $V_k^l = \mathrm{Var}(\phi(X_{ik}^l, Y_{jk}^l)) = (\theta_k^l) - \frac{1}{4}\mathrm{Pr}(X_{ik}^l = Y_{jk}^l) - (\theta_k^l)^2$ is simplified as $V = \bar{\theta} - \bar{\theta}^2$ assuming that the variance V_k^l is same across readers and modalities. Here $\bar{\theta}$ denotes average of two comparing AUCs and $\mathrm{Pr}(X_{ik}^l = Y_{jk}^l)$ is assumed to be zero.

(b) The correlations specified in (8) are the same either across readers or across tests or both. Thus, the correlations are simplified as the following 11 representative correlations.

where $k \neq k'$ and $l \neq l'$. ρ_{11} and ρ_{21} are the correlations between ϕ's when the test results are evaluated by the same reader using the same test; ρ_{12}, ρ_{22}, and ρ_{32} are the correlations between ϕ's when the test results are evaluated by different readers using the same test; ρ_{13}, ρ_{23}, and ρ_{33} are the correlations between ϕ's when the test results are evaluated by the same reader using different tests; and ρ_{14}, ρ_{24}, and ρ_{34} are the correlations between ϕ's when the test results are evaluated by different readers using different tests. Note that $\rho_{3((k,l),(k,l))} = \mathrm{Corr}[\phi(X_{ik}^l, Y_{jk}^l), \phi(X_{ik}^l, Y_{jk}^l)]$ represents the correlation between ϕ's when the test results are obtained from the same diseased and non-diseased subjects, read by the

same reader from the same test. Therefore, $\rho_{31} = \rho_{3((k,l),(k,l))}$ is always 1 so is not considered in the estimation. Under the above assumptions (a) and (b), the variance in Theorem 2 is simplified as

$$\begin{aligned}\mathrm{Var}(\hat{\theta}^l - \hat{\theta}^{l'}) &\approx \mathrm{Var}^*(\hat{\theta}^l - \hat{\theta}^{l'}) \\ &= \frac{2V}{mnr}\Bigg[\big(1 + (n-1)\rho_{11} + (m-1)\rho_{21}\big) \\ &\quad + (r-1)\big((n-1)\rho_{12} + (m-1)\rho_{22} + \rho_{32}\big) \\ &\quad - \big((n-1)\rho_{13} + (m-1)\rho_{23} + \rho_{33}\big) \\ &\quad - (r-1)\big((n-1)\rho_{14} + (m-1)\rho_{24} + \rho_{34}\big)\Bigg] \end{aligned} \tag{10}$$

with $V = \bar{\theta} - \bar{\theta}^2$, $\bar{\theta}$ = average of the two comparing AUCs. We suggest determining V and 11 different types of correlations from a pilot study or a similar study. In a multi-reader, multi-test design, $\hat{\rho}_{2s} \le \hat{\rho}_{1s} < \hat{\rho}_{3s}$ for a fixed s ($s = 1, \ldots 4$), and $\hat{\rho}_{t4} \le \hat{\rho}_{t2} \le \hat{\rho}_{t3} < \hat{\rho}_{t1}$ for a fixed t ($t = 1, 2, 3$). The proposed power at the significance level α is then

$$\begin{aligned}\text{Power} &= \Phi\left(\frac{-z_{1-\alpha/2}\sqrt{\mathrm{Var}^*(\hat{\theta}^l - \hat{\theta}^{l'})} + \delta_1}{\sqrt{\mathrm{Var}^*(\hat{\theta}^l - \hat{\theta}^{l'})}}\right) \\ &\quad + \Phi\left(\frac{-z_{1-\alpha/2}\sqrt{\mathrm{Var}^*(\hat{\theta}^l - \hat{\theta}^{l'})} - \delta_1}{\sqrt{\mathrm{Var}^*(\hat{\theta}^l - \hat{\theta}^{l'})}}\right). \end{aligned} \tag{11}$$

Details are shown in Web Appendix.

4. Simulation Studies

We considered the situation where r readers examine the test results of N (m diseased, n non-diseased) subjects who undergo two test modalities ($h = 2$). We varied the total sample size N from 100 to 200 and set the number of readers to $r = 4$, 8, 12. We calculated the theoretical power based on the power formula in (11) under different scenarios for the number of readers and subjects.

First, the data is generated as follows. Let the test results for diseased subject i and non-diseased subject j be $\mathbf{X}_i = (X_{i1}^1, \ldots, X_{ir}^1, X_{i1}^2, \ldots, X_{ir}^2)$, $i = 1, \ldots, m$ and $\mathbf{Y}_j = (Y_{j1}^1, \ldots, Y_{jr}^1, Y_{j1}^2, \ldots, Y_{jr}^2)$, $j = 1, \ldots, n$, respectively. To generate $\mathbf{X}_i$ and $\mathbf{Y}_j$, we expressed their elements into the two components

$$X_{ik}^l = r_k^l + \epsilon_{ik}^l, \quad Y_{jk}^l = \widetilde{r}_k^l + \widetilde{\epsilon}_{jk}^l, \ k = 1, \ldots, r; l = 1, 2,$$

where r_k^l and $\widetilde{r}_k^l$ account for variability due to reader k of test l in the diseased and non-diseased groups, respectively. We assumed that $r_k^l \sim N(0, \sigma_R^2)$ and $\widetilde{r}_k^l \sim N(0, \widetilde{\sigma}_R^2)$ for a fixed l. On the other hand, ϵ_{ik}^l and $\widetilde{\epsilon}_{jk}^l$ account for variability due to subject i, j from test l in the two groups, independently with r_k^l and $\widetilde{r}_k^l$. For diseased subject i, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}^1, \ldots, \epsilon_{ir}^1, \epsilon_{i1}^2, \ldots, \epsilon_{ir}^2)$ is a $N((\mu_{d1}\mathbf{1}, \mu_{d2}\mathbf{1})', \Sigma_1)$ random variable with $\mathbf{1} = (1, \ldots, 1)'_{r\times 1}$ and $\Sigma_1 = \begin{pmatrix} \Sigma_{11(r\times r)}^1 & \Sigma_{12(r\times r)}^1 \\ \Sigma_{12(r\times r)}^1 & \Sigma_{11(r\times r)}^1 \end{pmatrix}$. For non-diseased subject j, $\boldsymbol{\epsilon}_j = (\epsilon_{j1}^1, \ldots, \epsilon_{jr}^1, \epsilon_{j1}^2, \ldots, \epsilon_{jr}^2)$ is a $N(\mathbf{0}, \Sigma_0)$ random variable with $\Sigma_0 = \begin{pmatrix} \Sigma_{11(r\times r)}^0 & \Sigma_{12(r\times r)}^0 \\ \Sigma_{12(r\times r)}^0 & \Sigma_{11(r\times r)}^0 \end{pmatrix}$. The diagonal elements of Σ_{11}^1 (or Σ_{11}^0) are the variance of ϵ_{ik}^l's (or ϵ_{jk}^l's), and their off-diagonal elements are the covariance between the test results when they are read by the different readers using the same test. On the other hand, the diagonal and off-diagonal elements of Σ_{12}^1 (or Σ_{12}^0) are the covariance between the ϵ_{ik}^l's (or ϵ_{jk}^l's) when they are read by the same reader using the different modalities, and those when they are read by the different readers using the different modalities, respectively.

We assumed $\sigma_R^2 = 0.02$ and $\widetilde{\sigma}_R^2 = 0.03$ for the reader variability and assumed $\Sigma_{11}^1(i,i) = 0.98$, $\Sigma_{11}^1(i,j) = 0.3$, $\Sigma_{12}^1(i,i) = 0.8$, and $\Sigma_{12}^1(i,j) = 0.25$ ($i \ne j$) for diseased subject i; and $\Sigma_{11}^0(i,i) = 0.72$, $\Sigma_{11}^0(i,j) = 0.225$, $\Sigma_{12}^0(i,i) = 0.6$, and $\Sigma_{12}^0(i,j) = 0.1875$ ($i \ne j$) for non-diseased subject j. The assumed values for σ_R^2 (or $\widetilde{\sigma}_R^2$) combined with Σ_1 (or Σ_0) imply that the correlations between the test results when they are evaluated, respectively, by the different readers using the same test, by the same reader using the different tests, and by the different readers using the different tests are 0.3, 0.8, and 0.25 for diseased subjects and are 0.225, 0.6, and 0.1875 for non-diseased subjects. We set μ_{d1}=μ_{d2}=1.12 to obtain the true AUCs θ^1=θ^2=0.8; and $\mu_{d1} = 1.37$ and $\mu_{d2} = 1.12$ to obtain θ^1=0.85 and θ^2=0.8.

Tables 1 and 2 summarize the results based on 1000 simulations. The difference of the two AUCs was estimated nonparametrically by equation (2), and its variance was derived using the structural components in equation (5). For testing the equality of the two AUCs, the Z-statistic in equation (6) was used.

In Table 1, we present the results of the theoretical powers and compare it with the empirical powers when the true AUC difference is 0.05 ($\theta^1 = 0.85$, $\theta^2 = 0.8$). The empirical AUC difference is very close to 0.05, and the asymptotic standard errors and the standard deviations are nearly identical. The last column in Table 1 presents the theoretical powers calculated using a power formula (11) in which it was assumed that $\bar{\theta} = 0.825$ and the 11 correlations are $\rho_{11} = 0.31$, $\rho_{12} = 0.08$, $\rho_{13} = 0.24$, $\rho_{14} = 0.06$, $\rho_{21} = 0.22$, $\rho_{22} = 0.06$, $\rho_{23} = 0.17$, $\rho_{24} = 0.05$, $\rho_{32} = 0.15$, $\rho_{33} = 0.55$, $\rho_{34} = 0.12$. These correlations were estimated by the corresponding sample correlations from the simulated data illustrated at the beginning of this section. The theoretical powers are quite close to the empirical powers. Overall, we can see that the power increases with the increasing number of readers, and a balanced design when $m = n$ has the most power when the total sample size N is fixed.

Next, we conducted simulation studies to examine the type I error rates of the proposed Wald test. The results when the difference in AUC is 0 ($\theta^1 = \theta^2 = 0.8$) are shown in Table 2. As expected, the Wald test showed negligible bias, and the standard errors calculated from the asymptotic theory closely mimic the empirical standard deviations. The empirical type I error rates are close to the 5% level for all scenarios.

Table 1
$\theta^1 - \theta^2 = 0$ $(\theta^1 = 0.85,\ \theta^2 = 0.8)$

N	(m, n)	k	Est[1]	ASE[2]	SE[3]	Emp.power[4]	Theory.power[5]
100	(50, 50)	4	0.048	0.018	0.018	0.788	0.807
		8	0.049	0.016	0.016	0.895	0.894
		12	0.048	0.015	0.015	0.913	0.921
	(33, 67)	4	0.049	0.019	0.020	0.749	0.726
		8	0.048	0.017	0.017	0.828	0.823
		12	0.049	0.016	0.016	0.865	0.857
	(25, 75)	4	0.049	0.021	0.022	0.634	0.640
		8	0.049	0.019	0.019	0.760	0.742
		12	0.048	0.017	0.018	0.802	0.780
200	(100, 100)	4	0.048	0.012	0.014	0.969	0.981
		8	0.049	0.011	0.012	0.992	0.995
		12	0.048	0.010	0.011	0.998	0.998
	(67, 133)	4	0.048	0.014	0.014	0.942	0.956
		8	0.048	0.012	0.012	0.980	0.985
		12	0.049	0.011	0.011	0.993	0.991
	(50, 150)	4	0.048	0.015	0.016	0.893	0.910
		8	0.048	0.013	0.013	0.961	0.960
		12	0.049	0.012	0.012	0.980	0.972

[1]Mean estimated AUC differences by (2).
[2]Mean estimated standard errors in (5).
[3]Empirical standard deviation of the estimated AUC differences.
[4]Empirical power to detect the difference in the two AUCs at $\alpha = 0.05$.
[5]Theoretical power calculated by power formula (11).

Table 2
$\theta^1 - \theta^2 = 0$ $(\theta^1 = \theta^2 = 0.8)$

N	(m, n)	k	Est[1]	ASE[2]	SE[3]	Emp.power[4]
100	(50, 50)	4	0.0002	0.019	0.019	0.049
		8	0.0008	0.016	0.016	0.050
		12	0.0010	0.015	0.015	0.053
	(33, 67)	4	0.0004	0.020	0.021	0.055
		8	−0.0002	0.018	0.018	0.055
		12	−0.0005	0.017	0.016	0.053
	(25, 75)	4	−0.0007	0.022	0.023	0.059
		8	−0.0003	0.019	0.020	0.059
		12	−0.0004	0.018	0.018	0.053
200	(100, 100)	4	0.0009	0.013	0.013	0.051
		8	−0.0004	0.011	0.012	0.053
		12	−0.0003	0.011	0.011	0.055
	(67, 133)	4	0.0007	0.014	0.015	0.059
		8	−0.0004	0.012	0.013	0.055
		12	−0.0002	0.012	0.011	0.047
	(50, 150)	4	−0.0004	0.016	0.017	0.059
		8	$< 0.1 \times 10^{-3}$	0.014	0.014	0.057
		12	−0.0003	0.013	0.013	0.058

[1]Mean estimated AUC differences by (2).
[2]Mean estimated standard errors in (5).
[3]Empirical standard deviation of the estimated AUC differences.
[4]Empirical power to detect the difference in the two AUCs at $\alpha = 0.05$.

5. Practical Implementation

We illustrate the proposed power formula using data from the ACRIN DMIST retrospective multi-reader study (Hendrick et al., 2008). The goal of this study was to compare the accuracy of soft-copy digital mammography with that of screen-film mammography for breast cancer diagnosis. Three digital mammography manufacturers (Fischer, Fuji, and GE) participated in the study; each had 6–12 readers and 98–120 women screened. We selected the data from the digital mammography machine from Fuji. For the Fuji study, each of the 12 radiologists read 98 cases (27 cancer cases and 71 benign or negative cases) for both soft-copy digital and screen-film mammograms. Each radiologist identified suspicious findings and rate suspicion of breast cancer in identified lesions by using a 7-point scale (from 1 = definitely not malignant to 7 = definitely malignant).

We calculated nonparametric AUCs for each reader and test combination separately. The values were then averaged across readers for each test. Using the nonparametric method, the average AUCs were 0.756 (SE 0.054) for the screen-film mammography and 0.715 (SE 0.065) for the soft-copy digital mammography. The estimated AUC difference between the two modalities was 0.041 (SE 0.032) with p-value = 0.20, indicating no significant difference in the AUCs between Fuji soft-copy digital and screen-film mammography.

In order to compute the power, we suggest obtaining 11 correlations in (9) using related prior studies or literature. In the DMIST data example, the estimated correlations between the two tests were $\hat{\rho}_{11}=0.528$, $\hat{\rho}_{12}=0.243$, $\hat{\rho}_{13}=0.267$, $\hat{\rho}_{14}=0.213$, $\hat{\rho}_{21}=0.24$, $\hat{\rho}_{22}=0.102$, $\hat{\rho}_{23}=0.118$, $\hat{\rho}_{24}=0.096$,

Table 3
Sample sizes and powers for different effect sizes and correlations

			$\delta_1 = 0.05$		$\delta_1 = 0.06$	
N	(m, n)	k	**Case I**	**Case II**	**Case I**	**Case II**
100	(50, 50)	4	0.452	0.345	0.598	0.465
		6	0.615	0.419	0.771	0.558
		8	0.739	0.470	0.877	0.618
		10	0.828	0.507	0.937	0.660
		12	0.890	0.534	0.969	0.690
	(33, 67)	4	0.380	0.275	0.509	0.373
		6	0.526	0.328	0.681	0.443
		8	0.647	0.364	0.801	0.490
		10	0.743	0.390	0.880	0.523
		12	0.817	0.410	0.930	0.547
200	(100, 100)	4	0.741	0.601	0.879	0.757
		6	0.891	0.702	0.969	0.848
		8	0.958	0.763	0.993	0.894
		10	0.985	0.801	0.999	0.920
		12	0.995	0.828	1.000	0.937
	(67, 133)	4	0.654	0.493	0.807	0.645
		6	0.822	0.580	0.933	0.737
		8	0.915	0.634	0.979	0.789
		10	0.961	0.670	0.994	0.821
		12	0.983	0.696	0.998	0.843

Case I: $\rho_{11} = 0.5, \rho_{12} = 0.25, \rho_{13} = 0.25, \rho_{14} = 0.25, \rho_{21} = 0.24, \rho_{22} = 0.1, \rho_{23} = 0.1, \rho_{24} = 0.1, \rho_{32} = 0.4, \rho_{33} = 0.4, \rho_{34} = 0.4$.
Case II: $\rho_{11} = 0.5, \rho_{12} = 0.25, \rho_{13} = 0.25, \rho_{14} = 0.2, \rho_{21} = 0.24, \rho_{22} = 0.1, \rho_{23} = 0.1, \rho_{24} = 0.1, \rho_{32} = 0.4, \rho_{33} = 0.4, \rho_{34} = 0.3$.

$\hat{\rho}_{32} = 0.382, \hat{\rho}_{33} = 0.431, \hat{\rho}_{34} = 0.341$. In addition to the 11 correlation values, if we assume $\bar{\theta} = 0.74$ and the difference between two AUCs under the alternative hypothesis δ_1 is 0.05, the estimated power is 0.452 at $\alpha = 0.05$ according to the power formula (11) for 27 diseased subjects, 71 non-diseased subjects and 12 readers.

Table 3 presents the power calculation for different numbers of diseased and non-diseased subjects with a varying number of readers. The total sample size was set to 100 or 200, letting the proportion of diseased subjects over non-diseased subjects be 1 or 0.5. The number of readers was set to 4, 6, 8, 10, or 12. The assumed AUCs for the two modalities (θ^1, θ^2) were either (0.75, 0.7) or (0.75, 0.69). Thus, the effect size δ_1 is 0.05 or 0.06. We assumed that the correlation coefficients $(\rho_{11}, \rho_{12}, \rho_{13}, \rho_{14}, \rho_{21}, \rho_{22}, \rho_{23}, \rho_{24}, \rho_{32}, \rho_{33}, \rho_{34})$ are (0.5,0.25,0.25,0.25,0.24,0.1,0.1,0.1,0.4,0.4,0.4) (Case I) or (0.5,0.25,0.25,0.2,0.24,0.1,0.1,0.1,0.4,0.4,0.3) (Case II). As indicated, the power increases as the number of total sample sizes, effect sizes, or readers increase. When the total sample size is fixed, having equal numbers of diseased and non-diseased subjects is the most powerful. It is shown that when the correlations ρ_{14} and ρ_{34} decreased, the power decreased substantially. Here, ρ_{14} and ρ_{34} indicate the correlations between ϕ's when the test results from the two test modalities are obtained from the same diseased and different non-diseased subjects and from the same diseased and non-diseased subjects, respectively, in which the test results are evaluated by different readers using different tests. Let's assume we need to design a study with 100 or 200 participants and up to 12 readers. We want a minimum power of 80% assuming that the effect size is 0.05. In Case I with an equal number of diseased and non-diseased subjects, we need 10 readers for $N = 100$ but only 6 readers for $N = 200$. If the ratio of the diseased to non-diseased is 1:2, then we need 12 readers for $N = 100$. However, In Case II, we need at least 200 participants (100 in each group) and 10 readers to achieve a power greater than 80%.

6. Discussion

The multi-reader, multi-test design is commonly used in radiological studies to compare different diagnostic techniques because it requires the smallest number of subjects. We developed a novel power formula to detect the AUC difference for any two diagnostic tests in a multi-reader, multi-test design based on the theory of generalized two-sample U-statistics. We showed the asymptotic normality of the nonparametric AUC differences and constructed the power formula using the Wald test.

DeLong et al.'s (1988) approach is the special case of the presented nonparametric approach because the former can be applied to situations where multiple readers interpret a single test, or multiple tests are read by a single reader. Our estimation method and hypothesis testing retain the spirit of Song's (1997) approach, in that the AUC estimates are the same as the U-statistics in which readers are treated as fixed effects, and both methods use the Wald-test for comparing correlated AUCs. However, the proposed method differs from Song's in the variance estimation for the nonparametric AUCs. In contrast to Song's jackknife method, we applied Sen's (1960) method of structural components similar to Delong's method. We derived the explicit expression of the variance of nonparametric AUCs accounting for the correlated data structure from the multi-reader, multi-test design and next introduced 11 representative correlations to simplify the complicated variance structure for the power calculation. In addition, we are not aware of any nonparametric methods that deal with the sample size and power issues including Song's method.

In practice, a power formula based on the OR method is one of the most widely used approaches for multi-reader diagnostic accuracy studies. Obuchowski and Rockette (1995) pointed out that the type I error rate based on the OR method is at the correct level for eight or more readers. This is likely due to the incorrect F-statistic approximation used in the mixed effects ANOVA model design. Note that the number of pseudo-observations used as the response in the mixed model is equal to the number of readers times the number of tests, which is very small in usual applications. In this case, the asymptotic distribution approximation may not be ideal. Instead, in our approach, the asymptotic approximation is based on the total number of subjects so that the asymptotic results are more reliable. Our simulation results also demonstrated that the proposed method performs well even with a small number of readers. The major strength of the proposed power formula is that it is easy to implement as a useful alternative to the OR method, especially when a study is expected to have a relatively small number of readers.

A major difference between the proposed and OR (or DBM) methods is that our method treats the reader as a fixed effect whereas the latter treats it as a random factor. As Obuchowski et al. (2004) explained in detail, in phase II studies in which readers are selected from a specific institution, the selected sample of readers is often not generalizable to a broad population of readers. In this case, the conclusion of the study should pertain to the particular readers only and treating readers as fixed effects makes sense. Our method is thus suitable for use in this setup. However, in phase III studies, the readers should present a general population of radiologists, and it is reasonable to assume that random readers can account for variability across readers. The proposed approach of assuming fixed readers in this situation is still applicable but may result in power loss when reader effects are actually random.

Several articles have discussed the comparison of nonparametric partial AUCs (pAUCs) by extending DeLong et al.'s (1988) approach (e.g., Zhang et al., 2002; Dodd and Pepe, 2003; He and Escobar, 2008). Although we focused on using the entire AUC as a measure of accuracy, which is most widely used in multi-reader, multi-test studies, the proposed method can be easily extended to compare pAUCs. The only change to our test statistic is that the AUC of a diagnostic test by a specific reader in equation (1) is replaced with the pAUC. Its variance can be estimated using Sen's (1960) method, as previously used, and the inference will be based on the asymptotic normality using trimmed U-statistics theory, as presented by He and Escobar (2008). The corresponding power calculation will be also based on the correlations of the pAUC estimators.

As a final remark, we would like to note that our power calculation for analyzing the multi-reader, multi-test ROC data is based on a complete (reader by test) factorial study design. In many applications, however, test result data might be missing. As our statistic is based on pooling AUC estimators across all readers, our method can allow different readers to examine a different number of diseased or non-diseased cases. We can make a simple correction for the variance estimation using the available data by taking an approach similar to that of Zhou and Gatsonis (1996). The power calculation will be modified to reflect different missing proportions from each reader.

7. Supplementary Materials

Web Appendix referenced in Sections 2 and 3 and R code to calculate the power are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgement

This research was supported in part by NIH/NCI grant U01-CA 079778. The authors thank Constantine Gatsonis for his support and helpful comments.

References

Beiden S. V., Wagner R. F., and Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments. *Academic Radiology* **7**, 341–349.

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**, 837–845.

Dodd, L. D. and Pepe, M. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614–623.

Dorfman D. D., Berbaum K. S., and Metz C. E. (1992). ROC rating analysis: Generalization to the population of readers and cases with the jackknife method. *Investigative Radiology* **27**, 723–731.

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.

He, Y. and Escobar, M. (2008). Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. *Statistics in Medicine* **27**, 5291–5308.

Hendrick R. E., Cole E. B., Pisano E. D., Acharyya S., Marques H., Cohen M. A., Jong R. A., Mawdsley G. E., Kanal K. M., D'Orsi C. J., Rebner M., and Gatsonis C. (2008). Accuracy of soft-copy digital mammography versus that of screen-film mammography according to digital manufacturer: ACRIN DMIST retrospective multireader study. *Radiology* **247**, 38–48.

Hillis S. L. (2007). A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Statistics in Medicine* **26**, 596–619.

Hillis S. L., Berbaum K. S., and Metz C. E. (2008). Recent developments in the Dorfman–Berbaum–Metz procedure for multireader ROC study analysis. *Academic Radiology* **15**, 647–661.

Hillis S. L., Obuchowski N. A., and Berbaum K. S. (2011). Power estimation for multireader ROC methods: An updated and a unified approach. *Academic Radiology* **18**, 129–142.

Hillis S. L., Obuchowski N. A., Schartz K. M., and Berbaum K. S. (2005). A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine* **24**, 1579–1607.

Gallas B. (2006). One-shot estimate of MRMC variance: AUC. *Academic Radiology* **13**, 353–362.

Lee M. T. and Dehling H. G. (2005). Generalized two-sample U-statistics for clustered data. *Statistica Neerlandica* **59**, 313–323.

Li, G. and Zhou, K. (2008). A unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *Journal of the American Statistical Association* **103**, 705–713.

Metz C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298.

Obuchowski, N. A. (1995a). Multi-reader multi-modality ROC studies: Hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. *Academic Radiology* **2**, 522–529.

Obuchowski, N. A. (1995b). Multireader receiver operating characteristic studies: A comparison of study designs. *Academic Radiology* **2**, 709–716.

Obuchowski N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research* **7**, 371–392.

Obuchowski N. A., Beiden S. V., Berbaum K. S., Hillis S. L., Ishwaran H., Song H. H., and Wagner R. F. (2004). Multireader, multicase receiver operating characteristic analysis: An empirical comparison of five methods. *Academic Radiology* **11**(9), 980–995.

Obuchowski N. A. and Rockette H. E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. *Communications in Statistics—Simulation and Computation* **24**, 285–308.

Sen P. K. (1960). On some convergence properties of *U*-statistics. *Calcutta Statistical Association Bulletin* **10**, 1–18.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Song H. H. (1997). Analysis of correlated ROC areas in diagnostic testing. *Biometrics* **53**, 370–382.

Song X. and Zhou X. H. (2005). A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. *Biostatistics* **6**, 303–312.

Swets J. A. and Pickett R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection theory*. New York: Academic.

Zhang, D. Z., Zhou, X. H., Freeman, D. H., and Freeman J. L. (2002). 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* **21**, 701–715.

Zhou, X. H. and Gatsonis, C. A. (1996). A simple method for comparing correlated ROC curves using incomplete data. *Statistics in Medicine* **15**, 1687–1693.

Zhou X. H., Obuchowski N. A., and McClish D. K. (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley.

Received December 2013. *Revised June* 2014.
Accepted June 2014.

Appendix

Proof of Theorem 1

We use the Wald's device so consider any linear combination of $\hat{\boldsymbol{\theta}}$, say $\sum_{k,l} q_{kl}\hat{\theta}_k^l$, where q_{kl} is any constant. The latter can be expressed as

$$\frac{1}{mn}\sum_{k=1}^{r}\sum_{l=1}^{h}\sum_{i}^{m}\sum_{j}^{n} q_{kl}\phi(X_{ik}^l, Y_{jk}^l).$$

This is one generalized *U*-statistics considered in Lee and Dehling (2005), where the kernel function is given by

$$\sum_{k=1}^{r}\sum_{l=1}^{h} q_{kl}\phi(X_{ik}^l, Y_{jk}^l).$$

Therefore, the asymptotical normality holds following the result in Lee and Dehling (2005). Particularly, we conclude that for some $\tilde{\boldsymbol{\theta}}$, $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a multivariate normal with zero mean vector and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{((k,l),(k',l'))})$ as given in Theorem 1. Clearly, the asymptotic limit $\hat{\theta}_k^l$ is given by

$$\Pr(X_{ik}^l > Y_{jk}^l) + \frac{1}{2}\Pr(X_{ik}^l = Y_{jk}^l) = \theta_k^l$$

where $\boldsymbol{\theta}=(\theta_k^l)_{k=1,\ldots,r;l=1,\ldots,h}$.