Published in final edited form as:

Stat Med. 2007 February 10; 26(3): 596–619.

A Comparison of Denominator Degrees of Freedom Methods for Multiple Observer ROC Analysis

Stephen L. Hillis^{1,2}

Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP), Iowa City VA Medical Center, Iowa City, IA, U.S.A.

Abstract

There are several methods available for analyzing multireader ROC studies that generalize results to both the reader and case populations. Two of these methods - the Dorfman-Berbaum-Metz (DBM) method and the Obuchowski-Rockette (OR) method - appear to be quite different in their original formulations. However, recently it has been shown that the DBM and OR procedures yield the same test statistic when based on the same accuracy measure and covariance estimation method, but inferences can vary depending on which denominator degrees of freedom (ddf) method, DBM or OR, is used. I show in simulations that there are problems with both ddf methods: OR is ultraconservative with significance levels considerably below the nominal level, and DBM can result in extremely wide confidence intervals because the ddf can be close to zero. I propose a new ddf method that overcomes both of these problems and can be used with either the OR or DBM procedure.

Keywords

Receiver operating characteristic (ROC) curve; corrected F ; diagnostic radiology; degrees of freedom

1 INTRODUCTION

Receiver operating characteristic (ROC) curve analysis is a well established method for evaluating and comparing the performance of diagnostic tests [1-5]. There are several methods available for analyzing multireader ROC studies that generalize results to both the reader and case populations. Five such methods have recently been described and compared in an analysis of three data sets by Obuchowski et al [6]. The study shows that the methods can lead to different conclusions, illustrating the need for more theoretical and empirical investigation of the methods to provide insight and guidance regarding the appropriateness of each procedure in a given situation.

Two of these methods - the Dorfman-Berbaum-Metz (DBM) method [7,8] and the Obuchowski-Rockette (OR) method [9,10] - appear to be quite different in their original formulations. However, recently it has been shown by Hillis et al [11] that the DBM and OR procedures yield the same test statistic when based on the same accuracy measure and covariance estimation method, but inferences depend on which denominator degrees of freedom (ddf) method, DBM or OR, is used.

In simulations I find that there are problems with both ddf methods: OR is ultraconservative with significance levels considerably below the nominal level while DBM, though much closer

¹Correspondence to: Stephen L. Hillis, Ph.D., Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP), Iowa City VA Medical Center (152), 601 Highway 6 West, Iowa City, IA 52246-2208, U.S.A.

²E-mail: steve-hillis@uiowa.edu

to the nominal significance level, sometimes results in extremely wide confidence intervals because the ddf can be close to zero. I propose using a new ddf estimator that overcomes these problems and can be used with either the DBM or OR procedure.

The outline of the paper is as follows. I describe the OR method in Section 2. In Section 3 I derive the new ddf from the OR model, discuss the rationale for the original OR ddf, and compare the new ddf with the DBM ddf. In Section 4 I show how the new ddf can be derived from the DBM model and discuss the possibility that the DBM ddf may be based on an inadequate Satterthwaite approximation. In Section 5 I discuss confidence intervals for test differences, inferences for single tests, and fixed readers analysis. In Section 6 I compare the performance of the new ddf method with the OR and DBM ddf methods in a simulation study. I illustrate and compare the three ddf methods using a previously published data set in Section 7 and make concluding remarks in the final section.

2 THE OBUCHOWSKI-ROCKETTE (OR) METHOD

2.1 Design and notation

A commonly used study design in radiology that allows generalization to both the reader and case populations is the test reader case factorial design, where each case (i.e., patient) undergoes each diagnostic test (e.g., CT scan, MRI, or X-ray) and the resulting images are evaluated once by each reader (usually a radiologist). Typically the number of cases is 25-200 while the number of readers is 3-15. Throughout I assume that the data have been collected using this factorial design.

Let Z_{ijk} denote the rating assigned to the k th case by the j th reader using the i th modality. For example, a five-level ordinal integer scale is commonly used with higher rating values indicating a higher level of confidence that the case is diseased; thus a rating value of 1 might correspond to “definitely not diseased,” 5 to “definitely diseased,” with other values corresponding to intermediate assessments. Alternatively, a quasi continuous 0% to 100% confidence scale is often used. The observed data consists of the Z_{ijk} , with $i = 1, \dots, t$, $j = 1, \dots, r$, $k = 1, \dots, c$, where t is the number of tests (or modalities), r the number of readers, and c the number of cases. In addition, each case is classified as diseased or nondiseased according to an available reference standard.

2.2 Model and test statistic

The corrected F test was proposed by Obuchowski and Rockette [9]. Let $\hat{\theta}_{ij}$ denote the AUC estimate (or other accuracy estimate) for the i th test and j th reader. Their approach is to use a test reader ANOVA model for the AUCs, but unlike a conventional ANOVA model they allow the errors to be correlated to account for correlation in the AUCs due to each reader evaluating the same cases. Thus their two-way ANOVA model corresponds to the three-way study design. For the factorial design with only one replication, the corrected F test model, which I refer as the OR model, can be written as

$$\hat{\theta}_{ij} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij} \quad (1)$$

$i = 1, \dots, t$, $j = 1, \dots, r$, where τ_i denotes the fixed effect of test i ; R_j denotes the random effect of reader j , $(\tau R)_{ij}$ denotes the random test reader interaction, and ϵ_{ij} is the error term. The R_j and $(\tau R)_{ij}$ are assumed to be mutually independent and normally distributed with zero means and variances σ_R^2 , reflecting differences in reader ability, and σ_{TR}^2 , reflecting test-by-reader interaction. The ϵ_{ij} are assumed to be normally distributed with zero mean and constant variance σ_ϵ^2 , which represents variability attributable to cases and within-reader variability that describes how a reader interprets the same image in different ways on different occasions. The

ϵ_{ij} are independent of the R_j and $(\tau R)_{ij}$. However, since the same cases are read by each reader using each test, the ϵ_{ij} are not assumed to be independent. Instead, equi-covariance of the errors between readers and tests is assumed, resulting in three possible covariances given by

$$\text{Cov}\left(\epsilon_{ij}, \epsilon_{i'j'}\right) = \begin{cases} \text{Cov}_1 & i \neq i', j = j' \text{ (different test, same reader)} \\ \text{Cov}_2 & i = i', j \neq j' \text{ (same test, different reader)} \\ \text{Cov}_3 & i \neq i', j \neq j' \text{ (different test, different reader)} \end{cases}$$

It follows from model (1) that σ_ϵ^2 , Cov_1 , Cov_2 , and Cov_3 are also the variance and corresponding covariances of the AUC estimates, conditional on the reader and test $\times$ reader effects. Obuchowski and Rockette [9] suggest the following ordering for the covariances:

$$\text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \geq 0. \quad (2)$$

I only consider the factorial design with one replication; however, my results will also apply when there are multiple replications.

The corrected F statistic for testing the null hypothesis of no test effect ($H_0: \tau_1 = \tau_2 = \dots = \tau_t$) is given by

$$F^* = \frac{\text{MS(T)}}{\text{MS(T}^*\text{R)} + r(\text{Cov}_2 - \text{Cov}_3)} \quad (3)$$

where $\text{MS(T)} = \frac{r}{t-1} \sum_{i=1}^t (\hat{\theta}_{i.} - \hat{\theta}_{..})^2$ and $\text{MS(T}^*\text{R)} = \frac{1}{(t-1)(r-1)} \sum_{i=1}^t \sum_{j=1}^r (\hat{\theta}_{ij} - \hat{\theta}_{i.} - \hat{\theta}_{.j} + \hat{\theta}_{..})^2$. A subscript replaced by a dot indicates that values are averaged across the missing subscript; for example, $\hat{\theta}_{i.} = \frac{1}{r} \sum_{j=1}^r \hat{\theta}_{ij}$ and $\hat{\theta}_{..} = \frac{1}{tr} \sum_{i=1}^t \sum_{j=1}^r \hat{\theta}_{ij}$. Obuchowski and Rockette [9] state that F^* has an approximate null $F_{t-1; \text{ddf}_0}$ distribution, where

$$\text{ddf}_0 = (t-1)(r-1) \quad (4)$$

Throughout, the ddf subscript denotes the first author of the paper that proposed the ddf method: O = Obuchowski, D = Dorfman, H = Hillis.

In practice we do not know Cov_2 or Cov_3 and have to estimate them from the data or use estimates from previous studies. Thus the statistic actually used is

$$F_{\text{OR}} = \frac{\text{MS(T)}}{\text{MS(T}^*\text{R)} + \max[r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3), 0]} \quad (5)$$

where $\widehat{\text{Cov}}_2$ and $\widehat{\text{Cov}}_3$ denote estimates for Cov_2 and Cov_3 , respectively. Note that equation (5) incorporates constraint (2) by setting $\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3$ to zero if it is negative. Since Cov_2 and Cov_3 are also the corresponding covariances of the AUC estimates conditional on the reader and test $\times$ reader effects, they can be estimated using ROC analysis methods that treat cases as random but readers as fixed. For example, for trapezoidal-rule AUC estimates [5], Obuchowski and Rockette [9] suggest estimating σ_ϵ^2 , Cov_1 , Cov_2 and Cov_3 by averaging corresponding variances and covariances computed using the method of DeLong et al [12], which provides covariance and variance estimates treating only cases as random. More generally, any acceptable method for estimating the AUC variance and covariances that treats readers as fixed can be used, such as jackknifing or bootstrapping. The OR estimates obtained from averaging

the corresponding fixed-reader AUC variances and covariances are denoted by $\hat{\sigma}_\epsilon^2$, $\widehat{\text{Cov}}_1$, $\widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$.

3 THE PROPOSED DDF

In this section I show that F^* has an approximate $F_{(t-1), df_2}$ null distribution, where

$$df_2 = \frac{\left[E\left(MS(T^*R) \right) + r\left(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3 \right) \right]^2}{\left[\frac{E\left(MS(T^*R) \right)}{(t-1)(r-1)} \right]^2} \quad (6)$$

In deriving (6) I treat $\widehat{\text{Cov}}_2$ and $\widehat{\text{Cov}}_3$ as known. An estimate of df_2 that incorporates constraint (2) is given by

$$ddf_H = \frac{\left\{ MS(T^*R) + \max\left[r\left(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3 \right), 0 \right] \right\}^2}{\left[\frac{MS(T^*R)}{(t-1)(r-1)} \right]^2} \quad (7)$$

Thus I propose using ddf_H as the ddf for F_{OR} , since F_{OR} approximates F^* . I also discuss why Obuchowski and Rockette suggested $(t-1)(r-1)$ as the ddf.

3.1 Derivation of ddf_H

The expected values for $MS(T)$, $MS(R)$, and $MS(T^*R)$ are derived in the Appendix and given in Table 1, where $MS(R) = \frac{t}{r-1} \sum_{j=1}^r (\hat{\theta}_{.j} - \hat{\theta}_{..})^2$. It follows from Table 1 that

$$E[MS(T)] = \begin{cases} E[MS(T^*R)] + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3), & \text{if } H_0 \\ \frac{r}{t-1} \sum_{i=1}^t (\tau_i - \tau_{..})^2 + E[MS(T^*R)] + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3), & \text{if } H_1 \end{cases} \quad (8)$$

where H_0 is the null hypothesis of no test effect ($H_0: \tau_1 = \dots = \tau_t = 0$). Thus the numerator and denominator of F^* in equation (3) have the same expected value under H_0 , while the numerator has a larger expected value if H_0 does not hold; hence F^* is an appropriate ANOVA test statistic. [Table 1]

To show that F^* has an approximate $F_{(t-1), df_2}$ null distribution with df_2 defined by equation (6), I first consider an approximation for the distribution of its denominator, $MS(T^*R) + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3)$. Let X_m denote a random variable with a χ_m^2 / m distribution that is independent of $MS(T^*R)$. Then $E(X_m) = 1$ and X_m converges to 1 in probability as $m \rightarrow \infty$. By Slutsky's theorem the limiting distribution of $MS(T^*R) + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3) X_m$ as $m \rightarrow \infty$ is equal to the distribution of $MS(T^*R) + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3)$. My approach is to approximate the distribution of $MS(T^*R) + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3)$ by the limiting distribution of an approximate distribution for $MS(T^*R) + r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3) X_m$ as $m \rightarrow \infty$.

I show in the Appendix that $MS(T)$ and $MS(T^*R)$ are independently distributed,

$MS(T^*R) \sim \chi_{(t-1)(r-1)}^2 \frac{E[MS(T^*R)]}{(t-1)(r-1)}$, and under H_0 $MS(T) \sim \chi_{t-1}^2 \frac{E[MS(T)]}{t-1}$. Satterthwaite [13, 14] has shown that a linear function of independent random variables $MS = \sum_i a_i MS_i$, where $MS_i \sim \chi_{df_i}^2 \frac{E[MS_i]}{df_i}$, is approximately distributed as $\chi_{df}^2 \frac{E[MS]}{df}$, where

$$df = \frac{\left[\sum_i a_i E[MS_i] \right]^2}{\sum_i \left[a_i E[MS_i] \right]^2 / df_i} \quad (9)$$

Using Satterthwaite's procedure with $a_1 = 1$; $MS_1 = MS(T^*R)$, $a_2 = r(Cov_2 - Cov_3)$, and $MS_2 = X_m$, it follows that $MS(T^*R) + r(Cov_2 - Cov_3) X_m$ is approximately distributed as

$$\chi_{df_S(m)}^2 \frac{E[MS(T^*R)] + r(Cov_2 - Cov_3)}{df_S(m)} \quad (10)$$

where $df_S(m)$ is given by

$$df_S(m) = \frac{\left[\frac{E[MS(T^*R)] + r(Cov_2 - Cov_3)}{(t-1)(r-1)} \right]^2}{\frac{E[MS(T^*R)]^2}{(t-1)(r-1)} + \frac{r^2(Cov_2 - Cov_3)^2}{m}}$$

Since $\lim_{m \rightarrow \infty} df_S(m) = df_2$, where df_2 is defined by equation (6), it follows that the limiting distribution of distribution (10) as $m \rightarrow \infty$ is

$$\chi_{df_2}^2 \frac{E[MS(T^*R)] + r(Cov_2 - Cov_3)}{df_2} \quad (11)$$

Thus the distribution of the denominator of F^* , $MS(T^*R) + r(Cov_2 - Cov_3)$, is approximated by distribution (11).

Since $MS(T)$ and $MS(T^*R) + r(Cov_2 - Cov_3)$ are independent, $MS(T) \sim \chi_{t-1}^2 E[MS(T)] / (t-1)$ under H_0 , $MS(T^*R) + r(Cov_2 - Cov_3)$ is approximately $\chi_{df_2}^2 \{E[MS(T^*R)] + r(Cov_2 - Cov_3)\} / df_2$, and $E[MS(T)] = E[MS(T^*R)] + r(Cov_2 - Cov_3)$ under H_0 , it follows that the null distribution of $F^* = MS(T) / [MS(T^*R) + r(Cov_2 - Cov_3)]$ is approximated by

$$\frac{U / (t-1)}{V / df_2}$$

where $U \sim \chi_{t-1}^2$, $V \sim \chi_{df_2}^2$, and U and V are independent; that is, F^* has an approximate

$F_{(t-1), df_2}$ null distribution. Since F_{OR} approximates F^* and $ddf_H(7)$ estimates df_2 , I propose using ddf_H as the ddf for F_{OR} .

In summary, approximating the distribution of F_{OR} involves three distinct approximations. First, F_{OR} is approximated by F^* . The adequacy of this approximation depends on how well $\widehat{Cov}_2 - \widehat{Cov}_3$ approximates $Cov_2 - Cov_3$. Throughout, I assume that the number of cases is moderate or large, as is typical for ROC studies; thus it is reasonable to expect that this approximation will be adequate since the precision of the covariance estimates increases with the number of cases. Second, the distribution of $MS(T^*R) + r(Cov_2 - Cov_3)$ is approximated by the distribution of $MS(T^*R) + r(Cov_2 - Cov_3) X_m$ for large m , where $X_m \sim \chi_m^2 / m$, so that the Satterthwaite approximation can be applied to the denominator of F_{OR} ; this approximation is justified by Slutsky's theorem and the law of large numbers. Finally, the distribution of $MS(T^*R) + r(Cov_2 - Cov_3) X_m$ for large m is approximated using the Satterthwaite procedure to give distribution (11), with df_2 then estimated by ddf_H . This approximation depends on the adequacy of the Satterthwaite approximation, with $Cov_2 - Cov_3$ approximated by $\widehat{Cov}_2 - \widehat{Cov}_3$ and $E[MS(T^*R)]$ approximated by $MS(T^*R)$. I expect this approximation to be adequate since the Satterthwaite approximation has been shown to perform acceptably for sums of mean squares, with expected mean squares estimated by observed mean squares when estimating the degrees of freedom, even when the mean square degrees of freedom are low [13,14]. In the simulation study I empirically investigate the performance of ddf_H .

3.2 Rationale for ddf_0

To understand why Obuchowski and Rockette [9] suggested using $(t-1)(r-1)$ as the ddf, I now reexamine their argument. Based upon the work of Pavur and Nath [15], they note that under H_0

$$\frac{1-\rho}{1+(r-1)\rho} F \quad (12)$$

is distributed as $F_{(t-1)(r-1)}$, where $F = MS(T)/MS(T^*R)$ and

$$\rho = \frac{2(\widehat{Cov}_2 - \widehat{Cov}_3)}{\widehat{Var}(\hat{\theta}_{ij} - \hat{\theta}_{i'j'})}, \quad i \neq i'$$

They then replace ρ in equation (12) by the estimate

$$\hat{\rho} = \frac{2(\widehat{Cov}_2 - \widehat{Cov}_3)}{\widehat{Var}(\hat{\theta}_{ij} - \hat{\theta}_{i'j'})}$$

where

$$\widehat{Var}(\hat{\theta}_{ij} - \hat{\theta}_{i'j'}) = 2(MS(T^*R) + \widehat{Cov}_2 - \widehat{Cov}_3) \quad (13)$$

to derive the OR F statistic (5).

Although replacing ρ by a high precision estimate should not appreciably alter the distribution of (12), the variance estimate (13) lacks precision for the typical study with only a few tests and readers, since $MS(T^*R)$ has degrees of freedom $(t-1)(r-1)$. Furthermore, the estimate of ρ is correlated with F , and this correlation needs to be taken into account because of the lack of precision in the estimate. Thus ddf_0 does not adequately describe the distribution of the F_{OR} statistic (5) for typical studies because it does not account for either the lack of precision in estimating ρ or the correlation between the estimate of ρ and F .

4 DDF_H AND THE DBM PROCEDURE

4.1 The DBM procedure

The DBM method proposed by Dorfman et al [7] for analyzing multireader ROC studies also generalizes results to both the reader and case populations. For this method AUC pseudovalues are computed using the Quenouille-Tukey jackknife [16-18] separately for each reader-test combination. Let Y_{ijk} denote the AUC pseudovalue for test i , reader j , and case k ; by definition $Y_{ijk} = c\hat{\theta}_{ij} - (c-1)\hat{\theta}_{ij(k)}$, where $\hat{\theta}_{ij}$ denotes the AUC estimate based on all of the data for the i th test and j th reader and $\hat{\theta}_{ij(k)}$ denotes the AUC estimate when data for the k th case are omitted. Using the Y_{ijk} as the responses, the DBM procedure specifies testing for a test effect using a fully crossed three-factor ANOVA, with test treated as a fixed factor and reader and case as random factors. The DBM estimate of $\hat{\theta}_{ij}$ is $Y_{ij\cdot}$, which is the jackknife accuracy estimate corresponding to $\hat{\theta}_{ij}$.

Recently Hillis et al [11] generalize the DBM method by showing how it can be used with normalized pseudovalues and quasi pseudovalues: normalized pseudovalues Y_{ijk}^* are defined by $Y_{ijk}^* = Y_{ijk} + (\hat{\theta}_{ij} - Y_{ij\cdot})$, and quasi pseudovalues are defined as any values such that the resulting test-reader sample means, variances, and covariances are identical to $\hat{\theta}_{ij}$ and $c\hat{\Sigma}$, where $\hat{\Sigma}$ is the estimated fixed-reader covariance matrix used to compute the OR procedure quantities $\hat{\sigma}^2$, $\widehat{Cov}_1$, $\widehat{Cov}_2$, and $\widehat{Cov}_3$. For normalized and quasi pseudovalues the DBM

accuracy estimate, given by Y_{dij} , is equal to $\hat{\theta}_{ij}$. They show that using DBM with normalized pseudovalues yields the same test statistic b for testing for a modality effect as using OR with jackknife covariance estimates, and more generally, using DBM with quasi pseudovalues yields the same test statistic as OR for other covariance estimation methods. However, even when DBM and OR yield the same test statistic, inferences depend on which ddf method, DBM or OR, is used. From hereon I assume, when comparing the DBM and OR procedures, that the pseudovalues are the normalized pseudovalues if $\widehat{\text{Cov}}_2$ and $\widehat{\text{Cov}}_3$ are based on jackknife covariance estimates, and are quasi pseudovalues if $\widehat{\text{Cov}}_2$ and $\widehat{\text{Cov}}_3$ are based on another fixed-reader covariance estimation method such as DeLong et al [12].

Let $\text{MS}(\text{T})_{\text{pseudo}}$, $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}$, $\text{MS}(\text{T}^*\text{C})_{\text{pseudo}}$, and $\text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}$ denote the test, test×reader, test×case, and test×reader case mean squares for the DBM three-way ANOVA of the pseudovalues. The DBM F statistic for testing the null hypothesis of no test effect is

$$F_{\text{DBM}} = \frac{\text{MS}(\text{T})_{\text{pseudo}}}{\text{MS}(\text{T}^*\text{R})_{\text{pseudo}} + \max[\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}, 0]} \quad (14)$$

Hillis et al [11] show that $F_{\text{DBM}} = F_{\text{OR}}$ by showing that

$$\text{MS}(\text{R}) = \frac{1}{c} \text{MS}(\text{R})_{\text{pseudo}} \quad \text{MS}(\text{T}^*\text{R}) = \frac{1}{c} \text{MS}(\text{T}^*\text{R})_{\text{pseudo}} \quad (15)$$

and

$$\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3 = \frac{1}{rc} [\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}] \quad (16)$$

The result follows by substitution in (14). They also show that the DBM model implies the OR model, and they give the one-to-one mapping between the OR and DBM model parameters.

This form (14) of the DBM F statistic utilizes the same constraints as F_{OR} (5), since equation (16) implies that the constraint $\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}} \geq 0$, utilized in the denominator of F_{DBM} (14), is equivalent to the constraint $\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3 \geq 0$ employed in

F_{OR} . Although Dorfman et al [7,8] suggest also constraining $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}$ to be nonnegative, References [11,19] discuss conceptual reasons for not using this constraint; furthermore, simulations [20] show that performance is better when only $\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}$ is constrained to be nonnegative. Throughout this paper I use F_{DBM} as specified by equation (14).

4.2 Comparison of ddf_H , ddf_O , and ddf_D

The DBM numerator and denominator degrees of freedom for the null distribution of F_{DBM} (14) are $t-1$ and ddf_D , respectively, where

$$\text{ddf}_D = \begin{cases} \frac{[\text{MS}(\text{T}^*\text{R})_{\text{pseudo}} + \text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}]^2}{\frac{\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}^2}{(t-1)(r-1)} + \frac{\text{MS}(\text{T}^*\text{C})_{\text{pseudo}}^2}{(t-1)(c-1)} + \frac{\text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}^2}{(t-1)(r-1)(c-1)}}, & \text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}} > 0 \\ (t-1)(r-1), & \text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}} \leq 0 \end{cases} \quad (17)$$

Hillis et al [11] show that ddf_D has the following form when expressed in terms of the OR mean squares and covariance estimates:

$$ddf_D = \begin{cases} \frac{[MS(T^*R) + r(\widehat{Cov}_2 - \widehat{Cov}_3)]^2}{\frac{MS(T^*R)^2}{(t-1)(r-1)} + \frac{[\hat{\sigma}_\epsilon^2 - \widehat{Cov}_1 + (r-1)(\widehat{Cov}_2 - \widehat{Cov}_3)]^2}{(t-1)(c-1)} + \frac{[\hat{\sigma}_\epsilon^2 - \widehat{Cov}_1 - \widehat{Cov}_2 + \widehat{Cov}_3]^2}{(t-1)(r-1)(c-1)}} & \widehat{Cov}_2 - \widehat{Cov}_3 > 0 \\ (t-1)(r-1), & \widehat{Cov}_2 - \widehat{Cov}_3 \leq 0 \end{cases} \quad (18)$$

Equations (17) and (18) yield the same value for ddf_D .

Comparing equations (7) and (18) we see that

$$ddf_D \leq ddf_H \quad (19)$$

This relationship is intuitive, since in deriving ddf_H I treat σ^2 , Cov_1 , Cov_2 , and Cov_3 as known while DBM treats them (or equivalently, the corresponding DBM model variance components) as unknown; thus the uncertainty from estimating them is manifested in the lower DBM degrees of freedom.

It follows from equation (7) that

$$ddf_O = (t-1)(r-1) \leq ddf_H$$

with equality attained if and only if $\widehat{Cov}_2 - \widehat{Cov}_3 < 0$, showing that the ddf_O method is more conservative than the ddf_H method. There is not a similar relationship between ddf_O and ddf_D : ddf_D can be larger or smaller than ddf_O . Table 2 presents the ddf formulas in terms of the OR model in part (a) and the ddf relationships in part (b). [Table 2]

4.3 Ddf_H in terms of the DBM mean squares

I can also express ddf_H in terms of the DBM analysis mean squares. It follows from equations (7,15-16) that

$$ddf_H = \frac{\{MS(T^*R)_{pseudo} + \max[MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo}, 0]\}^2}{MS(T^*R)_{pseudo}^2 / [(t-1)(r-1)]} \quad (20)$$

Part (c) of Table 2 presents the ddf formulas expressed in terms of the DBM model. Note that the relationships in part (b), which we previously derived from the part (a) formulas, can also be derived from the equivalent part (c) formulas.

Alternatively, I can derive ddf_H in form (20) directly from the DBM model in the following way. It is shown by Hillis et al [11] that, under the assumptions of the DBM model,

$$E[MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo}] = r\sigma_{\tau C}^2 \quad (21)$$

where $\sigma_{\tau C}^2$ is the test case variance component for the DBM model. Define

$$F_{DBM}^* = \frac{MS(T)_{pseudo}}{MS(T^*R)_{pseudo} + r\sigma_{\tau C}^2}$$

For typical ROC studies ($c > 25$) the degrees of freedom will be at least 25 for each mean square in equation (21); thus $MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo}$ should approximate $r\sigma_{\tau C}^2$ reasonably well, implying

$$F_{DBM} \approx F_{DBM}^*$$

My approach is to show that an approximate ddf for F_{DBM}^* , and hence for F_{DBM} , is given by equation (20).

Under the DBM model assumptions (conventional three-way ANOVA model assumptions for the pseudovalues) the numerator and denominator of F_{DBM}^* have the same null expected values, $\text{MS}(\text{T})_{\text{pseudo}}$ and $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}$ are independently distributed,

$$\text{MS}(\text{T}^*\text{R})_{\text{pseudo}} \sim \chi_{(t-1)(r-1)}^2 E[\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}] / [(t-1)(r-1)]$$

and

$$\text{MS}(\text{T})_{\text{pseudo}} \sim \chi_{t-1}^2 E[\text{MS}(\text{T})_{\text{pseudo}}] / (t-1) \text{ if } H_0 \text{ is true.}$$

Using the same argument given in Section 3.1 but with $\text{MS}(\text{T})$ replaced by $\text{MS}(\text{T})_{\text{pseudo}}$, $\text{MS}(\text{T}^*\text{R})$ replaced by $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}$, and r ($\text{Cov}_2\text{-Cov}_3$) replaced by $r\sigma_{\tau C}^2$, it follows that

F_{DBM}^* has an approximate F_{t-1, df_2} null distribution, with

$$\text{df}_2 = \frac{[E(\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}) + r\sigma_{\tau C}^2]^2}{\frac{[E(\text{MS}(\text{T}^*\text{R})_{\text{pseudo}})]^2}{(t-1)(r-1)}}$$

Replacing $E[\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}]$ and $r\sigma_{\tau C}^2$ by their estimates, $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}$ and $\max[\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}, 0]$, respectively, yields equation (20).

4.4 Inadequate DBM Satterthwaite approximation

When $\text{MS}(\text{T}^*\text{C})_{\text{pseudo}} - \text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}} > 0$, ddf_D in equation (17) is the Satterthwaite approximation (9) for $\text{MS}(\text{T}^*\text{R}) + \text{MS}(\text{T}^*\text{C}) - \text{MS}(\text{T}^*\text{R}^*\text{C})$. Satterthwaite [13] remarks that caution should be used when applying formula (9) to a linear function of mean squares when some of the coefficients are negative, and Gaylor and Hopper [21] provide guidelines for determining if the Satterthwaite approximate is valid for this situation. We will see in the simulation study that an inadequate Satterthwaite approximation can cause ddf_D to approach zero, resulting in an extremely wide confidence interval.

5 SPECIAL CASES

5.1 Confidence interval for the difference of two tests

Define θ_i to be the expected accuracy estimate for test i ; that is, $\theta_i = E(\hat{\theta}_{ij})$. I have shown that F_{OR} , and hence also F_{DBM} , has an approximate $F_{(t-1); \text{ddf}_H}$ distribution; it follows that for $t = 2$ an approximate $(1-\alpha)100\%$ confidence interval for $\theta_i - \theta_j$ is given by

$$(\hat{\theta}_{i.} - \hat{\theta}_{j.}) \pm t_{\frac{\alpha}{2}; \text{ddf}_H} \sqrt{\frac{2}{r} \text{MSden}_{\text{OR}}} \quad (22)$$

where MSden_{OR} is the denominator of F_{OR} (5), or equivalently by

$$(\hat{\theta}_{i.} - \hat{\theta}_{j.}) \pm t_{\frac{\alpha}{2}; \text{ddf}_H} \sqrt{\frac{2}{rc} \text{MSden}_{\text{DBM}}} \quad (23)$$

where $\text{MSden}_{\text{DBM}}$ is the denominator of F_{DBM} (14). It can be shown, using that approach of Section 3.1, that equations (22-23) are also valid when $t > 2$.

5.2 Single test inference using all of the data

The results in this and the following section are presented without proof, but they can be derived using an approach similar to that used in Section 3.1. Define

$$MS_{denOR_single} = \frac{1}{t} \{MS(R) + (t-1)MS(T^*R) + tr \max(\widehat{Cov}_2, 0)\} ddf_{H_single} = \frac{\{MS(R) + (t-1)MS(T^*R) + tr \max(\widehat{Cov}_2, 0)\}^2}{\frac{\{MS(R)\}^2}{r-1} + \frac{\{(t-1)MS(T^*R)\}^2}{(t-1)(r-1)}}$$

An approximate $(1-\alpha)100\%$ confidence interval for θ_i is given by

$$\hat{\theta}_i \pm t_{\frac{\alpha}{2}; ddf_{H_single}} \sqrt{\frac{1}{r} MS_{denOR_single}} \quad (24)$$

and a test for $H_0: \theta_i = \theta_0$ can be made by comparing $t = (\hat{\theta}_i - \theta_0) / \sqrt{\frac{1}{r} MS_{denOR_single}}$ to a t distribution with degrees of freedom ddf_{H_single} . Equivalent expressions can be derived for the DBM procedure, but I omit those since I recommend instead the more robust approach described in the next section.

5.3 Single test inference using only corresponding data

A confidence interval and hypotheses test for θ_i can alternatively be based only on the data for the i th test. When we consider the data for test i , the OR model (1) reduces to a one-factor random effects ANOVA with the covariance for each pair of errors equal to Cov_2 . Since this single-test model makes no assumptions about the variances or covariances of the error terms corresponding to other tests, the resulting confidence interval should be more robust than confidence interval (24). Let $\widehat{Cov}_2^{(i)}$ denote the average of the fixed reader covariance estimates for the i th test and define

$$MSR^{(i)} = \frac{\sum_{j=1}^r (\hat{\theta}_{ij} - \hat{\theta}_i)^2}{r-1}$$

$$MS_{denOR_single}^{(i)} = MSR^{(i)} + \max[r\widehat{Cov}_2^{(i)}, 0]$$

and

$$ddf_{H_single}^{(i)} = \frac{\{MSR^{(i)} + \max[r\widehat{Cov}_2^{(i)}, 0]\}^2}{\frac{(MSR^{(i)})^2}{r-1}}$$

An approximate $(1-\alpha)100\%$ confidence interval for θ_i is given by

$$\hat{\theta}_i \pm t_{\frac{\alpha}{2}; ddf_{H_single}^{(i)}} \sqrt{\frac{1}{r} MS_{denOR_single}^{(i)}} \quad (25)$$

A similar result is also given by Obuchowski and Rockette [9], but instead of $ddf_{H_single}^{(i)}$ they use $r-1$; thus we see that $ddf_{H_single}^{(i)}$ yields a less conservative result, since $ddf_{H_single}^{(i)} \geq r-1$.

Similarly, the DBM procedure reduces to a conventional reader×case ANOVA of pseudovalues when we consider only data for a single test. Let $MS(R)_{pseudo}^{(i)}$, $MS(C)_{pseudo}^{(i)}$, and

$MS(R^*C)^{(i)}_{pseudo}$ denote mean squares using only pseudovalues corresponding to the i th test. An equivalent expression for confidence interval (25) is given by

$$\hat{\theta}_i \pm t_{\frac{\alpha}{2}; ddf_{H_single}^{(i)}} \sqrt{\frac{1}{rc} MSden_{DBM_single}^{(i)}}$$

where

$$MSden_{DBM_single}^{(i)} = MS(R)^{(i)}_{pseudo} + \max \left[MS(C)^{(i)}_{pseudo} - MS(R^*C)^{(i)}_{pseudo}, 0 \right]$$

and

$$ddf_{H_single}^{(i)} = \frac{\left\{ MS(R)^{(i)}_{pseudo} + \max \left[MS(C)^{(i)}_{pseudo} - MS(R^*C)^{(i)}_{pseudo}, 0 \right] \right\}^2}{\frac{\left(MS(R)^{(i)}_{pseudo} \right)^2}{r-1}}$$

A test for $H_0: \theta_i = \theta_0$ can be made by comparing either $t = (\hat{\theta}_i - \theta_0) / \sqrt{\frac{1}{r} MSden_{OR_single}^{(i)}}$ or

$t = (\hat{\theta}_i - \theta_0) / \sqrt{\frac{1}{rc} MSden_{DBM_single}^{(i)}}$ to a t distribution with degrees of freedom

$ddf_{H_single}^{(i)}$.

5.4 Fixed readers

If readers are treated as fixed for the OR model (1) then the only random effects are the error terms. Treating the variance and covariances of the error terms as known, the method of generalized least squares can be used, as discussed by Obuchowski and Rockette [9]. The test statistic, given by

$$\frac{(t-1)MS(T)\hat{\theta}_{ij}}{\left[\hat{\sigma}_\epsilon^2 - \widehat{Cov}_1 + (r-1)(\widehat{Cov}_2 - \widehat{Cov}_3) \right]}$$

has an approximate chi-squared distribution with $t-1$ degrees of freedom. Hillis et al [11] show that the DBM analysis, treating readers as fixed, will give the same test statistic when suitably normalized and approximately the same p -value for typical ROC studies. I see no problems with these methods in terms of the degrees of freedom.

6 SIMULATION STUDY

In a simulation study I compare ddf_H , ddf_O , and ddf_D with respect to the empirical significance level for testing the null hypothesis of no test effect; and with respect to the width of a 95% confidence interval for the difference of the test AUCs. The simulation model of Roe and Metz [22] provides continuous decision-variable outcomes generated from a binormal model that treats both cases and readers as random. I use this simulation model for simulating rating data, taking integer values from one to five, by transforming the continuous outcomes to discrete ratings using the same cutpoints as Dorfman et al [8]; the combinations of reader and case sample sizes, AUC values, and variance components are the same as used in Roe and Metz [22] and Dorfman et al [8]. Briefly, rating data are simulated for 144 combinations of three reader-sample sizes (readers = 3, 5, and 10), four case sample sizes (10+/90-, 25+/25-, 50+/50-, and 100+/100-, where “+” indicates a diseased case and “-” indicates a normal case), three AUC values (AUC = .702, .855, and .961), and four combinations of reader and case variance components. Two thousand samples are generated for each of the 144 combinations; within each simulation, all Monte Carlo readers read the same cases for each of two tests. Since these

are null simulations the test effect in the model is set to zero. The simulation design is summarized in Table 3. [Table 3]

In terms of the OR procedure, each sample is analyzed using two different methods for estimating the AUC and σ_{ϵ}^2 , Cov_1 , Cov_2 and Cov_3 . For one method I estimate the AUC using maximum likelihood estimation assuming a binormal model [23,24] and obtain $\hat{\sigma}_{\epsilon}^2$, $\widehat{\text{Cov}}_1$, $\widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$ using the jackknife method; for the other method I estimate the AUC using the trapezoidal-rule (trapezoid) method and obtain $\hat{\sigma}_{\epsilon}^2$, $\widehat{\text{Cov}}_1$, $\widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$ using the DeLong method. These estimation methods are referred to as the MLE-jack and trapezoid-DeLong methods, respectively. For each method I compute F_{OR} , ddf_H , ddf_O , and ddf_D using equations (5,7,4,18). The null hypothesis of no test effect is rejected if $F_{\text{OR}} > F_{.05;1,\text{ddf}}$. For each of the 144 combinations the empirical significance level is the proportion of samples for which the null hypothesis is rejected. For each simulation I also compute the width of the 95% confidence interval for the difference of the two test AUCs using equation (22) and report the mean width for the 144·2000 simulations.

Identical results can also be obtained using the DBM procedure, as discussed in Hillis et al [11]. MLE-jack results can be obtained by computing the normalized pseudovalues corresponding to the MLE AUC estimates, computing $\text{MS}(\text{T})_{\text{pseudo}}$, $\text{MS}(\text{T}^*\text{R})_{\text{pseudo}}$, $\text{MS}(\text{T}^*\text{C})_{\text{pseudo}}$, and $\text{MS}(\text{T}^*\text{R}^*\text{C})_{\text{pseudo}}$, computing F_{DBM} (14) which will be the same as F_{OR} , and then computing ddf_O , ddf_H , and ddf_D using equations (4,20,17). A 95% confidence for the difference of the two test AUCs is given by equation (23). Trapezoid-DeLong results can similarly be obtained using the DBM procedure with quasi pseudovalues as described in Hillis et al [11]. I exploit these OR-DBM equivalent relationships by using the DBM procedure for the MLE-jack simulations and the OR procedure for the Trapezoid-DeLong simulations. Simulations are performed using the IML procedure in SAS 9.1 [25] running under Windows XP. The MLE AUC pseudovalues are computed using a dynamic linked library (DLL), written in Fortran 90 by Don Dorfman and Kevin Schartz, that is accessed from within the IML procedure; this DLL is available on request.

The empirical significance levels from the simulation study are described in Table 4 and displayed in dot plots in Figure 1. The mean significance levels, with ranges indicated in parentheses, for ddf_O , ddf_H , and ddf_D , respectively, are 0.018 (0.041), 0.051 (0.052), and 0.043 (0.050) for MLE-jackknife estimation, and 0.019 (0.047), 0.054 (0.051), and 0.047 (0.058) for trapezoid-DeLong estimation. Thus we see that the average significance levels for ddf_H and ddf_O are much closer to the nominal .05 level than for ddf_D , with ddf_D slightly more conservative than ddf_H in accord with relationship (19). The dot plots show the ultraconservative performance of ddf_O : several significance levels are 0.00 and all of the significance levels are less than .05. For ddf_H and ddf_D the dot plots show bell-shaped distributions without outliers. For ddf_H , 93% (134/144) of the significance levels are within the interval [0.03, 0.07] for both estimation methods and the standard deviations of the significance levels are 0.011 (MLE-jackknife) and 0.010 (Trapezoid-DeLong) computed across the 144 combination of design factors; for ddf_D , 86% (124/144) and 92% (133/144) of the significance levels are within the interval [0.03, 0.07] for MLE-jackknife and trapezoid-DeLong estimation, respectively, and the standard deviations of the significance levels are 0.011 for both estimation methods. I conclude from the closeness of the mean significance level to the nominal level, the relatively small standard deviation, the high proportion of significance levels within .02 of the nominal level, and the absence of outliers that both ddf_H and ddf_{DBM} perform satisfactorily with respect to significance levels. [Table 4] [Figure 1]

The extremely large mean confidence interval widths corresponding to ddf_D in Table 4 (mean CI width: $\text{ddf}_D = 2.36\text{E}+121$ vs. $\text{ddf}_H = 0.173$ for MLE-jackknife estimation; $\text{ddf}_D = 2.74\text{E}$

+121 vs. $ddf_H = 0.184$ for Trapezoid-DeLong estimation) can be attributed to a small proportion of samples for which ddf_D approaches zero. For example, $ddf_D \leq 1$ in 0.002 of the samples; when I set $ddf_D = 1$ for these samples, then the resulting mean confidence interval widths are comparable (although somewhat larger) than those for ddf_H . Each ddf_D value less than one corresponds to an inadequate Satterthwaite ddf approximation, using the adequacy guidelines provided by Gaylor and Hopper [21]. I conclude that although the ddf_H and ddf_D methods are comparable with respect to significance levels, ddf_H performs better when confidence interval width is also considered.

Table 5 shows the mean empirical significance levels for ddf_H by factor level for each of the four simulation study factors. The results suggest that high AUC (.961) is associated with mild conservatism, while low AUC (.701) and low number of readers (3) are associated with mild liberalism. Accordingly, 17 of the 18 lowest empirical ddf_H significance levels in Figure 1(a) and 18 of the lowest 19 empirical ddf_H significance levels in Figure 1(b) correspond to AUC = .961; in contrast, 22 of the highest 23 values in Figure 1(a) and 21 of the 22 highest values in Figure 1(b) corresponding either to AUC = .701 or readers = 3. [Table 5]

7 EXAMPLE

My example comes courtesy of Carolyn Van Dyke, MD. The study [26] compared the relative performance of single spin-echo magnetic resonance imaging (SE MRI) to cinematic presentation of MRI (CINE MRI) for the detection of thoracic aortic dissection. There were 45 patients with an aortic dissection and 69 patients without a dissection imaged with both SE MRI and CINE MRI. Five radiologists independently interpreted all of the images using a five-point ordinal scale: 1 = definitely no aortic dissection, 2 = probably no aortic dissection, 3 = unsure about aortic dissection, 4 = probably aortic dissection, and 5 = definitely aortic dissection.

The analysis of this study using trapezoid AUC estimates and DeLong covariance estimates is displayed in Table 6 and the empirical ROC curves are displayed in Figure 2. Although Table 6 only displays the analysis using the OR procedure, identical results can be obtained with the DBM procedure using quasi pseudovalues. For testing $H_0: \theta_1 = \theta_2$ we have $F_{OR} = 4.485$, $ddf_O = 4$ ($p = .102$), $ddf_H = 15.07$ ($p = .051$), and $ddf_D = 13.81$ ($p = .053$). Thus ddf_O and ddf_H give somewhat different results, while ddf_H and ddf_D give similar results, with $ddf_H > ddf_D$ in accord with relationship (19). [Table 6] [Figure 2]

Single test 95% confidence intervals using only the corresponding data, as given by equation (25), are given in part (f) of Table 5. We see that the confidence interval for CINE (.897±0.0717) is roughly 50% wider than the interval for SE MRI (.941±0.0467). When based upon the combined data using equation (24), the single test confidence intervals are given by $\hat{\theta}_i \pm 0.0591$, $i = 1, 2$. I prefer the first approach since it does not assume that the AUCs have the same variances and covariances for each test, as does the combined data approach.

8 DISCUSSION

The motivation for this paper was the recent finding by Hillis et al [11] that the DBM and OR procedures yield identical test statistics when based on the same accuracy measure and covariance estimation methods, but their different ddf methods, ddf_D and ddf_O , can result in considerably different inferences. I proposed a new ddf estimator, ddf_H , that overcomes problems with the ddf_D and ddf_O methods. I derived ddf_H by showing that the null distribution of the OR test statistic can be approximated by an F_{t-1, ddf_H} distribution. I showed how ddf_H can be used with both the OR and DBM procedures and how it can also be derived from the

DBM model. The p -value corresponding to ddf_H will always be less than or equal to that corresponding to ddf_D or ddf_O .

Although ddf_H can be derived from the DBM model assumptions, the derivation in terms of the OR model is more important because the OR model provides an acceptable conceptual model. In contrast, the DBM model does not provide an acceptable conceptual model because pseudovalues have no intrinsic meaning; hence Hillis et al [11] characterize the DBM model as a “working model” that allows one to fit the OR model using conventional ANOVA software.

In simulations ddf_H performed better than ddf_O with respect to significance level, with the ddf_H mean significance level much closer to the nominal level. The shape and spread of the distributions of the significance levels were similar for the ddf_H and ddf_D methods, with ddf_H closer to the nominal level for binormal AUC estimation than ddf_D (mean significance levels: $ddf_H = .051$, $ddf_D = .043$) but deviating approximately the same amount from the nominal level for trapezoid AUC estimation (mean significance levels: $ddf_H = .054$, $ddf_D = .047$). However, a drawback of ddf_D was that confidence intervals were sometimes extremely wide due to ddf_D being close to zero. For these reasons I concluded that ddf_H performs better than either ddf_D or ddf_O . A SAS macro for implementing the OR-DBM procedure using ddf_H is available on request.

A limitation of this study is that I used only typical ROC sample sizes where the number of readers is low (≤ 10) and the number of cases is moderate (≥ 50) in the simulation study. Although the derivation of ddf_H involved several approximations, I discussed in Section 3.1 why ddf_H should perform adequately for the typical ROC diagnostic study where the number of cases is moderate to large, regardless of the number of readers. However, the performance of the new method may not be acceptable when the number of cases is less; certainly, further research is required before using it in this situation. Although the simulations only looked at small reader sample sizes, I see no reason why the new method should not work satisfactorily when the number of readers is moderate or large, but again I caution that our simulations have only been for typical ROC study sample sizes.

A topic to consider for future research is the robustness of the OR and DBM procedures to violations of the model assumptions. Since the procedures can be viewed as equivalent, only the less restrictive OR model assumptions need to be considered. The OR model assumes that the variances and covariances for the accuracy estimates do not vary by test. I conjecture that, similar to conventional ANOVA models, the test of the null hypothesis of no test effect will be fairly robust to departures from this assumption, but further investigation is required to support this conjecture. Although I would not expect single test confidence intervals based on this model to be robust to violations of this assumption, this problem can be circumvented by basing the confidence interval only on data for the corresponding test, as discussed in Section 5.3.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health, grant R01EB000863. I thank two anonymous referees for their excellent suggestions which greatly improved the paper, and also Nancy Obuchowski and Kevin Berbaum for their helpful suggestions in the final stage of preparing the manuscript. The views expressed in this article are those of the author and do not necessarily represent the views of the Department of Veterans Affairs.

A Appendix

In this section I show that the OR mean squares, MS(T), MS(R), and MS(T^{*}R), have expectations as given in Table 1 and are independently distributed as follows:

$$\begin{aligned} \text{MS(T)} &\sim \frac{c_1}{t-1} \chi_{t-1, \lambda}^2, \text{ with} \\ c_1 &= \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 + (r-1)(\text{Cov}_2 - \text{Cov}_3) \end{aligned} \quad (\text{A1})$$

and noncentrality parameter $\lambda = \frac{r}{c_1} \sum_{i=1}^t (\tau_i - \tau)^2$; $\text{MS(R)} \sim \frac{E[\text{MS(R)}]}{(r-1)} \chi_{(r-1)}^2$; and

$$\begin{aligned} \text{MS(T}^*\text{R)} &\sim \frac{E[\text{MS(T}^*\text{R)}]}{(t-1)(r-1)} \chi_{(t-1)(r-1)}^2. \text{ Furthermore, from Table 2 we see that } E[\text{MS(T)} | H_0] = \\ c_1; \text{ hence if } H_0: \tau_1 = \dots = \tau_t \text{ is true, then } \text{MS(T)} &\sim \frac{E[\text{MS(T)}]}{t-1} \chi_{t-1}^2. \end{aligned}$$

Let $\hat{\theta} = (\hat{\theta}_{11}, \dots, \hat{\theta}_{1r}, \dots, \hat{\theta}_{t1}, \dots, \hat{\theta}_{tr})'$ denote the vector of outcomes for the OR model (1) and let $\theta = (\theta_{11}, \dots, \theta_{1r}, \dots, \theta_{t1}, \dots, \theta_{tr})'$ denote the corresponding mean vector, where $\theta_{ij} = \mu + \tau_i$. Then $\theta \sim N(\theta, \Sigma)$ where $\Sigma = \text{Cov}(\theta)$. I assume that Σ is positive definite.

I first show that Σ has the following form:

$$\Sigma = \mathbf{I}_t \otimes (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r) + \mathbf{M}_t \otimes (x_3 \mathbf{I}_r + \text{Cov}_3 \mathbf{M}_r) \quad (\text{A2})$$

where $\otimes$ denotes the Kronecker product operator, $\mathbf{I}_p$ denotes the $p \times p$ identity matrix, $\mathbf{M}_p$ denotes a $p \times p$ matrix of ones,

$$x_1 = \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3 \quad (\text{A3})$$

$$x_2 = \text{Cov}_2 - \text{Cov}_3 \quad (\text{A4})$$

and

$$x_3 = \sigma_R^2 + \text{Cov}_1 - \text{Cov}_3 \quad (\text{A5})$$

It follows from the OR model (1) that $\text{Var}(\hat{\theta}_{ij}) = \sigma_R^2 + \sigma_{\tau R}^2 + \sigma_e^2$ and

$$\text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}) = \begin{cases} \text{Cov}_1 + \sigma_R^2 & i \neq i', j = j' \text{ (different test, same reader)} \\ \text{Cov}_2 & i = i', j \neq j' \text{ (same test, different reader)} \\ \text{Cov}_3 & i \neq i', j \neq j' \text{ (different test, different reader)} \end{cases}$$

Define $\sigma_{\theta}^2 = \sigma_R^2 + \sigma_{\tau R}^2 + \sigma_e^2$. Then, for example, for $t = 2$ and $r = 3$ the covariance matrix of $\hat{\theta} = (\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13}, \hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23})'$ is given by

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{bmatrix}$$

where

$$\Sigma_1 = \begin{bmatrix} \sigma_{\theta}^2 & \text{Cov}_2 & \text{Cov}_2 \\ \text{Cov}_2 & \sigma_{\theta}^2 & \text{Cov}_2 \\ \text{Cov}_2 & \text{Cov}_2 & \sigma_{\theta}^2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \text{Cov}_1 + \sigma_R^2 & \text{Cov}_3 & \text{Cov}_3 \\ \text{Cov}_3 & \text{Cov}_1 + \sigma_R^2 & \text{Cov}_3 \\ \text{Cov}_3 & \text{Cov}_3 & \text{Cov}_1 + \sigma_R^2 \end{bmatrix}$$

We can write

$$\Sigma = \begin{bmatrix} \Sigma_1 - \Sigma_2 & 0 \\ 0 & \Sigma_1 - \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_2 & \Sigma_2 \\ \Sigma_2 & \Sigma_2 \end{bmatrix} = \mathbf{I}_2 \otimes (\Sigma_1 - \Sigma_2) + \mathbf{M}_2 \otimes \Sigma_2 \quad (\text{A6})$$

where 0 denotes a matrix of zeros. Since

$$\Sigma_1 - \Sigma_2 = \begin{bmatrix} \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 & \text{Cov}_2 - \text{Cov}_3 & \text{Cov}_2 - \text{Cov}_3 \\ \text{Cov}_2 - \text{Cov}_3 & \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 & \text{Cov}_2 - \text{Cov}_3 \\ \text{Cov}_2 - \text{Cov}_3 & \text{Cov}_2 - \text{Cov}_3 & \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 \end{bmatrix} = x_1 \mathbf{I}_3 + x_2 \mathbf{M}_3$$

and $\Sigma_2 = x_3 \mathbf{I}_3 + \text{Cov}_3 \mathbf{M}_3$, where x_1 , x_2 , and x_3 are defined by equations (A3-A5), then substituting in (A6) gives

$$\Sigma = \mathbf{I}_2 \otimes (x_1 \mathbf{I}_3 + x_2 \mathbf{M}_3) + \mathbf{M}_2 \otimes (x_3 \mathbf{I}_3 + \text{Cov}_3 \mathbf{M}_3)$$

which is equation (A2) with $t = 2$; $r = 3$. It can similarly be shown for arbitrary t and r that the covariance matrix for $\theta = (\hat{\theta}_{11}, \dots, \hat{\theta}_{1r}, \dots, \hat{\theta}_{t1}, \dots, \hat{\theta}_{tr})'$ is given by equation (A2).

Define $\mathbf{C}_p = \mathbf{I}_p - \frac{1}{p} \mathbf{M}_p$. Smith and Lewis [27] give matrix expressions for balanced ANOVA sums of squares; from their results we have $\text{MS(T)} = \hat{\theta}' \mathbf{A}_1 \hat{\theta} / (t - 1)$, $\text{MS(T}^*\text{R)} = \hat{\theta} \mathbf{A}_2 \hat{\theta}' / [(t - 1)(r - 1)]$, and $\text{MS(R)} = \hat{\theta}' \mathbf{A}_3 \hat{\theta} / (r - 1)$, where $\mathbf{A}_1 = \frac{1}{r} \mathbf{C}_t \otimes \mathbf{M}_r$, $\mathbf{A}_2 = \mathbf{C}_t \otimes \mathbf{C}_r$, and $\mathbf{A}_3 = \frac{1}{t} \mathbf{M}_t \otimes \mathbf{C}_r$.

Let A and B be any two matrices and let $\text{tr}(\cdot)$ denote the trace function. Results that I utilize for the mean square distribution derivations are given below. For properties (a-c) I also assume that A and B are $tr \times tr$ symmetric matrices.

- a. $E(\hat{\theta}' \mathbf{A} \hat{\theta}) = \text{tr}(\mathbf{A} \Sigma) + \theta' \mathbf{A} \theta$
- b. $\hat{\theta}' \mathbf{A} \hat{\theta} \sim \chi_{m, \lambda}^2$ where $m = \text{rank}(\mathbf{A} \Sigma)$ and $\lambda = \theta' \mathbf{A} \theta$ if and only if $\mathbf{A} \Sigma$ is idempotent
- c. $\hat{\theta}' \mathbf{A} \hat{\theta}$ and $\hat{\theta}' \mathbf{B} \hat{\theta}$ are distributed independently if and only if $\mathbf{A} \Sigma \mathbf{B} = \mathbf{0}$
- d. $\mathbf{0} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{0} = \mathbf{0}$
- e. $a \mathbf{A} \otimes b \mathbf{B} = ab \mathbf{A} \otimes \mathbf{B}$
- f. $(\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D}) = (\mathbf{AB}) \otimes (\mathbf{CD})$
- g. $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$; $\mathbf{C} \otimes (\mathbf{A} + \mathbf{B}) = \mathbf{C} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{B}$
- h. $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A}) \text{rank}(\mathbf{B})$
- i. if $\mathbf{B}$ is nonsingular then $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$
- j. if $\mathbf{A}$ is idempotent then $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$
- k. $\mathbf{C}_p \mathbf{C}_p = \mathbf{C}_p$
- l. $\mathbf{M}_p \mathbf{M}_p = p \mathbf{M}_p$
- m. $\mathbf{M}_p \mathbf{C}_p = \mathbf{C}_p \mathbf{M}_p = \mathbf{0}$
- n. $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ are idempotent
- o. $\text{tr}(\mathbf{A}_1) = t-1$; $\text{tr}(\mathbf{A}_2) = (t-1)(r-1)$; $\text{tr}(\mathbf{A}_3) = (r-1)$

$$\mathbf{p}. \quad \mathbf{A}_1 \mathbf{A}_2 = \mathbf{A}_1 \mathbf{A}_3 = \mathbf{A}_2 \mathbf{A}_3 = \mathbf{0}$$

Properties (a-c) are well known [28], (d-j) are standard matrix results [29], and properties (k-p) are easily derived.

To derive the distribution of $\text{MS}(\text{T}) = \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}} / (t-1)$, I show below that $\mathbf{A}_1 \Sigma = c_1 \mathbf{A}_1$, where c_1 is defined by equation (A1). Using property (a) we then have

$$E[\text{MS}(\text{T})] = \frac{1}{t-1} E[\hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}}] = \frac{1}{t-1} [\text{tr}(\mathbf{A}_1 \Sigma) + \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}}] = \frac{1}{t-1} [\text{tr}(c_1 \mathbf{A}_1) + \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}}]$$

Since $\theta_{ij} = \mu$, then $\hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}} = r \sum_{i=1}^t (\theta_{i.} - \theta_{..})^2 = r \sum_{i=1}^t (\tau_i - \tau_{..})^2$; also, from property (o) we have $\text{tr}(c_1 \mathbf{A}_1) = c_1(t-1)$. Thus

$$E[\text{MS}(\text{T})] = \frac{1}{t-1} [\text{tr}(c_1 \mathbf{A}_1) + \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}}] = \frac{1}{t-1} \left[c_1(t-1) + r \sum_{i=1}^t (\tau_i - \tau_{..})^2 \right] = \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 + (r-1)(\text{Cov}_2 - \text{Cov}_3) + \frac{r}{t-1} \sum_{i=1}^t (\tau_i - \tau_{..})^2$$

as given in Table 1. Since $\mathbf{A}_1 \Sigma = c_1 \mathbf{A}_1$, then from property (n) it follows that $\frac{1}{c_1} \mathbf{A}_1 \Sigma$ is

idempotent, and hence properties (b,i,j,o) imply $\frac{1}{c_1} \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}} \sim \chi_{t-1, \lambda}^2$ where the noncentrality

parameter is $\lambda = \frac{1}{c_1} \hat{\boldsymbol{\theta}}' \mathbf{A}_1 \hat{\boldsymbol{\theta}} = \frac{r}{c_1} \sum_{i=1}^t (\tau_i - \tau_{..})^2$. It follows that $\text{MS}(\text{T}) \sim \frac{c_1}{t-1} \chi_{t-1, \lambda}^2$.

To show $\mathbf{A}_1 = c_1 \mathbf{A}_1$, using (A2) we have

$$\mathbf{A}_1 \Sigma = \frac{1}{r} \mathbf{C}_t \otimes \mathbf{M}_r [\mathbf{I}_t \otimes (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r) + \mathbf{M}_t \otimes (x_3 \mathbf{I}_r - \text{Cov}_3 \mathbf{M}_r)] = \frac{1}{r} \mathbf{C}_t \otimes \mathbf{M}_r [\mathbf{I}_t \otimes (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r)] + \frac{1}{r} \mathbf{C}_t \otimes \mathbf{M}_r [\mathbf{M}_t \otimes (x_3 \mathbf{I}_r - \text{Cov}_3 \mathbf{M}_r)]$$

Using property (f) it follows that

$$\mathbf{A}_1 \Sigma = \frac{1}{r} (\mathbf{C}_t \mathbf{I}_t) \otimes [\mathbf{M}_r (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r)] + \frac{1}{r} (\mathbf{C}_t \mathbf{M}_t) \otimes [\mathbf{M}_r (x_3 \mathbf{I}_r - \text{Cov}_3 \mathbf{M}_r)]$$

Using properties (d,e,g,l,m) we have

$$\mathbf{A}_1 \Sigma = \frac{x_1}{r} \mathbf{C}_t \otimes \mathbf{M}_r + \frac{x_2}{r} \mathbf{C}_t \otimes \mathbf{M}_r \mathbf{M}_r + 0 = \frac{(x_1 + r x_2)}{r} \mathbf{C}_t \otimes \mathbf{M}_r = (x_1 + r x_2) \mathbf{A}_1 = c_1 \mathbf{A}_1$$

I similarly derive the distribution of $\text{MS}(\text{T}^* \text{R}) = \hat{\boldsymbol{\theta}}' \mathbf{A}_2 \hat{\boldsymbol{\theta}} / [(t-1)(r-1)]$ by showing below that $\mathbf{A}_2 \Sigma = c_2 \mathbf{A}_2$, where

$$c_2 = \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3$$

Since $\hat{\boldsymbol{\theta}}' \mathbf{A}_2 \hat{\boldsymbol{\theta}} = \sum_{i=1}^t \sum_{j=1}^r (\theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..})^2 = 0$, then similar to the derivation for $\text{MS}(\text{T})$ distribution it follows that

$$E[\text{MS}(\text{T}^* \text{R})] = \frac{1}{(t-1)(r-1)} [\text{tr}(c_2 \mathbf{A}_2) + \hat{\boldsymbol{\theta}}' \mathbf{A}_2 \hat{\boldsymbol{\theta}}] = c_2 = \sigma_{\tau R}^2 + \sigma_e^2 - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3$$

as given in Table 1, and $\frac{1}{c_2} \hat{\boldsymbol{\theta}}' \mathbf{A}_2 \hat{\boldsymbol{\theta}} \sim \chi_{(t-1)(r-1)}^2$. Since $E[\text{MS}(\text{T}^* \text{R})] = c_2$, then it follows that

$$\text{MS}(\text{T}^* \text{R}) \sim \frac{E[\text{MS}(\text{T}^* \text{R})]}{(t-1)(r-1)} \chi_{(t-1)(r-1)}^2$$

To show $\mathbf{A}_2 \Sigma = c_2 \mathbf{A}_2$, using (A2) we have

$$\mathbf{A}_2 \Sigma = \mathbf{C}_t \otimes \mathbf{C}_r [\mathbf{I}_t \otimes (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r) + \mathbf{M}_t \otimes (x_3 \mathbf{I}_r + \text{Cov}_3 \mathbf{M}_r)] = \mathbf{C}_t \otimes \mathbf{C}_r [\mathbf{I}_t \otimes (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r)] + \mathbf{C}_t \otimes \mathbf{C}_r [\mathbf{M}_t \otimes (x_3 \mathbf{I}_r + \text{Cov}_3 \mathbf{M}_r)]$$

Using property (f) it follows that

$$\mathbf{A}_2 \mathbf{Z} = (\mathbf{C}_t \mathbf{I}_t) \otimes [\mathbf{C}_r (x_1 \mathbf{I}_r + x_2 \mathbf{M}_r)] + (\mathbf{C}_t \mathbf{M}_t) \otimes [\mathbf{C}_r (x_3 \mathbf{I}_r + \text{Cov}_3 \mathbf{M}_r)]$$

Using properties (d,e,g,m) we have

$$\mathbf{A}_2 \mathbf{Z} = \mathbf{C}_t \otimes (x_1 \mathbf{C}_r + \mathbf{0}) + \mathbf{0} = x_1 \mathbf{C}_t \otimes \mathbf{C}_r = x_1 \mathbf{A}_2 = c_2 \mathbf{A}$$

From property (c), $\hat{\theta}' \mathbf{A}_1 \hat{\theta}$ and $\hat{\theta}' \mathbf{A}_2 \hat{\theta}$ are independent if $\mathbf{A}_1 \Sigma \mathbf{A}_2 = 0$. To show $\mathbf{A}_1 \Sigma \mathbf{A}_2 = 0$, note that since $\mathbf{A}_1 \Sigma = c_1 \mathbf{A}_1$, then $\mathbf{A}_1 \Sigma \mathbf{A}_2 = c_1 \mathbf{A}_1 \mathbf{A}_2 = 0$ using property (p). It follows that MS(T) and MS(T*R) are independent. Similarly, the distribution of MS(R) can be derived and it can be shown that MS(R) is independent of MS(T) and MS(T*R).

References

1. Swets, JA.; Pickett, RM. Evaluation of diagnostic systems: methods from signal detection theory. Academic Press; New York: 1982.
2. Metz CE. ROC methodology in radiologic imaging. Investigative Radiology 1986;21:720–733. [PubMed: 3095258]
3. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–1293. [PubMed: 3287615]
4. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Critical Reviews in Diagnostic Imaging 1989;29:307–335. [PubMed: 2667567]
5. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36. [PubMed: 7063747]
6. Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song HH, Wagner RF. Multi-case ROC analyses: an empirical comparison of five methods. Academic Radiology 2004;11:980–995. [PubMed: 15350579] DOI:10.1016/j.acra.2004.04.014
7. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Investigative Radiology 1992;27:723–731. [PubMed: 1399456]
8. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Academic Radiology 1998;5:591–602. [PubMed: 9750888]
9. Obuchowski NA, Rockette HE. Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: an ANOVA approach with dependent observations. Communications in Statistics: Simulation and Computation 1995;24:285–308.
10. Obuchowski NA. Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations. With rejoinder. Academic Radiology 1995;2(Suppl 1):S22–S29. [PubMed: 9419702]
11. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for receiver operating characteristic (ROC) data. Statistics in Medicine 2005;24:1579–1607. [PubMed: 15685718] DOI:10.1002/sim.2024
12. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–844. [PubMed: 3203132]
13. Satterthwaite FE. Synthesis of variance. Psychometrika 1941;6:309–316.
14. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometric Bulletin 1946;2:110–114.
15. Pavur R, Nath R. Exact F tests in an ANOVA procedure for dependent observations. Multivariate Behavioral Research 1984;19:408–420.
16. Quenoille MH. Approximate tests of correlation in time series. Journal of the Royal Statistical Society, Series B 1949;11:68–84.
17. Quenoille MH. Notes on bias in estimation. Biometrika 1956;43:353–360.

18. Tukey JW. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* 1958;29:614.
19. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. *Academic Radiology* 2004;11:1260–1273. [PubMed: 15561573]DOI:10.1016/j.acra.2004.08.009
20. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Academic Radiology* 2005;12:1534–1541. [PubMed: 16321742]DOI:10.1016/j.acra.2005.07.012
21. Gaylor DW, Hopper FN. Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics* 1969;11:691–706.
22. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. *Academic Radiology* 1997;4:298–303. [PubMed: 9110028]
23. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology* 1969;6:487–496.
24. Dorfman, DD. RSCORE II. In: Swets, JA.; Pickett, RM., editors. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press; San Diego, CA: 1982. p. 212-232.
25. SAS for Windows, Version 9.1. SAS Institute Inc.; Cary, NC, USA: copyright (c) 2002-2003 by
26. Van Dyke, CW.; White, RD.; Obuchowski, NA.; Geisinger, MA.; Lorig, RJ.; Meziene, MA. Cine MRI in the diagnosis of thoracic aortic dissection. 79th RSNA Meetings; Chicago, IL. November 28 - December 3, 1993;
27. Smith JH, Lewis TO. Determining the effects of intra-class correlation on factorial-experiments. *Communications in Statistics Part A-Theory and Methods* 1980;9:1353–1364.
28. Searle, SR. *Linear Models*. Wiley; New York: 1971. p. 55-59.
29. Harville, DA. *Matrix Algebra From a Statistician's Perspective*. Springer; New York: p. 335-338.

Figure 1.
Dot plots of the 144 empirical significance levels using MLE-jackknife and trapezoid-DeLong estimation with each ddf method. The nominal significance level is .05.

Figure 2. Trapezoid ROC curves of the five readers for SE MRI and CINE MRI in the detection of aortic dissection. The average areas under the empirical curves are .941 (SE) and .897 (CINE).

Table 1.
Expected mean squares for the Obuchowski-Rockette model.

Mean square	Expected mean square
MS(T)	$\frac{r}{t-1} \sum_{i=1}^t (\tau_i - \tau_{.})^2 + \sigma_{\tau R}^2 + \sigma_{\epsilon}^2 - \text{Cov}_1 + (r-1)(\text{Cov}_2 - \text{Cov}_3)$
MS(R)	$t\sigma_R^2 + \sigma_{\tau R}^2 + \sigma_{\epsilon}^2 - \text{Cov}_2 + (t-1)(\text{Cov}_1 - \text{Cov}_3)$
MS(T*R)	$\sigma_{\tau R}^2 + \sigma_{\epsilon}^2 - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3$

Table 2.

Denominator degrees of freedom summary.

a) In terms of the OR mean squares and covariances estimates:

$ddf_0 = (t-1)(r-1)$

$$ddf_D = \begin{cases} \frac{[MS(T^*R) + r(\widehat{Cov}_2 - \widehat{Cov}_3)]^2}{(t-1)(r-1) + \frac{[\hat{\sigma}_\epsilon^2 - \widehat{Cov}_1 + (r-1)(\widehat{Cov}_2 - \widehat{Cov}_3)]^2}{(t-1)(c-1)} + \frac{[\hat{\sigma}_\epsilon^2 - \widehat{Cov}_1 - \widehat{Cov}_2 + \widehat{Cov}_3]^2}{(t-1)(r-1)(c-1)}}, & \widehat{Cov}_2 - \widehat{Cov}_3 > 0 \\ (t-1)(r-1), & \widehat{Cov}_2 - \widehat{Cov}_3 \leq 0 \end{cases}$$

$$ddf_H = \frac{[MS(T^*R) + \max[r(\widehat{Cov}_2 - \widehat{Cov}_3), 0]]^2}{\frac{[MS(T^*R)]^2}{(t-1)(r-1)}}$$

b) Relationships:

$ddf_0 \leq ddf_H$

$$ddf_0 = (t-1)(r-1) \begin{cases} \leq ddf_H, & \widehat{Cov}_2 - \widehat{Cov}_3 > 0 \\ = ddf_H, & \widehat{Cov}_2 - \widehat{Cov}_3 = 0 \end{cases}$$

ddf_D can be larger or smaller than ddf_0

c) In terms of the DBM mean squares:

$ddf_0 = (t-1)(r-1)$

$$ddf_D = \begin{cases} \frac{[MS(T^*R)_{pseudo} + MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo}]^2}{\frac{MS(T^*R)_{pseudo}^2}{(t-1)(r-1)} + \frac{MS(T^*C)_{pseudo}^2}{(t-1)(c-1)} + \frac{MS(T^*R^*C)_{pseudo}^2}{(t-1)(r-1)(c-1)}}, & MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo} > 0 \\ (t-1)(r-1), & MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo} \leq 0 \end{cases}$$

$$ddf_H = \frac{\{MS(T^*R)_{pseudo} + \max[MS(T^*C)_{pseudo} - MS(T^*R^*C)_{pseudo}, 0]\}^2}{MS(T^*R)_{pseudo}^2 / [(t-1)(r-1)]}$$

Table 3.

Simulation study design.

Factor	Number of levels	Description of levels
reader-sample size	3	readers = 3, 5, 10
case-sample size	4	cases = 10+/90-, 25+/25-, 50+/50-, 100+/100-
AUC	3	AUC = .702, .855, .961
variance components	4	HH, HL, LH, LL

Notes: For each factor combination $n = 2000$ samples are simulated for $t = 2$ tests, with each reader reading the same cases using each test. "+" = diseased case, "-" = normal case. See Roe and Metz [22] for the definitions of HH, HL, LH, and LL.

Table 4.

Results of the simulation study for the 144 combinations of reader-sample size, case-sample size, AUC, and variance components.

Estimation method	ddf method	Significance levels						CI width mean
		N	Mean	Min	Max	Range	SD	
MLE-jackknife	O	144	0.018	0.000	0.041	0.041	0.0144	0.231
	H	144	0.051	0.025	0.077	0.052	0.0105	0.173
	D	144	0.043	0.017	0.067	0.050	0.0109	2.36E+121
Trapezoid-Delong	O	144	0.019	0.000	0.047	0.047	0.0124	0.245
	H	144	0.054	0.032	0.082	0.051	0.0100	0.184
	D	144	0.047	0.023	0.080	0.058	0.0105	2.74E+121

Notes: Min: minimum; Max: maximum; SD: standard deviation; CI width: width of a 95% confidence interval for the difference of the AUC estimates; MLE-jackknife: binormal maximum likelihood AUC estimation and jackknife covariance estimation; Trapezoid-DeLong: trapezoid AUC estimation and DeLong covariance estimation.

Table 5.
Simulation study mean significance levels for ddf_H by factor level.

factor	factor level	N	Estimation method	
			MLE-jackknife (mean = 0.051)	Trapezoid-Delong (mean = 0.054)
case-sample size	10+/90-	36	0.047	0.053
	25+/25-	36	0.049	0.053
	50+/50-	36	0.052	0.055
	100+/100-	36	0.054	0.056
reader-sample size	3	48	0.056	0.060
	5	48	0.048	0.052
	10	48	0.048	0.051
	0.702	48	0.057	0.060
AUC	0.855	48	0.052	0.056
	0.961	48	0.043	0.047
variance components	HH	36	0.051	0.054
	HL	36	0.055	0.057
	LH	36	0.046	0.051
	LL	36	0.050	0.054

Table 6.

Analysis of Van Dyke et al [26] data using trapezoid AUC estimation and DeLong covariance estimation for $t = 2$ tests and $r = 5$ readers.

a) Trapezoid AUCs:			
	1 (CINE)	test	2 (Spin Echo)
reader (j)	$\hat{\theta}_{1j}$		$\hat{\theta}_{2j}$
1	0.9196		0.9478
2	0.8588		0.9053
3	0.9039		0.9217
4	0.9731		0.9994
5	0.8298		0.9300
	$\hat{\theta}_{1\cdot} = .8970$		$\hat{\theta}_{2\cdot} = .9408$
b) ANOVA table:			
Source	df	Sum of squares	Mean square
T	1	0.00479617	0.00479617
R	4	0.01534480	0.00383620
T*R	4	0.00220412	0.00055103
c) Covariance estimates computed from DeLong covariance matrix:			
$\hat{\sigma}_{\epsilon}^2 = .000792133$, $\widehat{\text{Cov}}_1 = .000342009$, $\widehat{\text{Cov}}_2 = .000339526$, $\widehat{\text{Cov}}_3 = .000235850$			
d) $\text{FOR} = \text{MS(T)} / [\text{MS(T*R)} + \max\{r(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3), 0\}] = 4.4849$			