HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2015 November 10.

Published in final edited form as:

Stat Med. 2014 January 30; 33(2): 330–360. doi:10.1002/sim.5926.

A Marginal-Mean ANOVA Approach for Analyzing Multireader Multicase Radiological Imaging Data

Stephen L. Hillis, Ph.D.

Departments of Radiology and Biostatistics, The University of Iowa, 3710 Medical Laboratories, 200 Hawkins Drive, Iowa City, IA 52242-1077, U.S.A.; Comprehensive Access and Delivery Research and Evaluation (CADRE) Center, Iowa City VA Health Care System

Stephen L. Hillis: steve-hillis@uiowa.edu

Abstract

The correlated-error ANOVA method proposed by Obuchowski and Rockette (OR) has been a useful procedure for analyzing reader-performance outcomes, such as the area under the receiver-operating-characteristic curve, resulting from multireader multicase radiological imaging data. This approach, however, has only been formally derived for the test-by-reader-by-case factorial study design. In this paper I show that the OR model can be viewed as a marginal-mean ANOVA model. Viewing the OR model within this marginal-mean ANOVA framework is the basis for the marginal-mean ANOVA approach, the topic of this paper. This approach (1) provides an intuitive motivation for the OR model, including its covariance-parameter constraints; (2) provides easy derivations of OR test statistics and parameter estimates, as well as their distributions and confidence intervals; and (3) allows for easy generalization of the OR procedure to other study designs. In particular, I show how one can easily derive OR-type analysis formulas for any balanced study design by following an algorithm which only requires an understanding of conventional ANOVA methods.

Keywords

Receiver operating characteristic (ROC) curve; correlated ANOVA; diagnostic radiology

1. INTRODUCTION

Receiver operating characteristic (ROC) curve analysis is a well established method for evaluating and comparing the performance of diagnostic tests. In radiological imaging studies such tests typically involve a human reader (usually a radiologist) evaluating an image or images resulting from an imaging modality (such as mammography for breast cancer) for a case (i.e., subject) with respect to confidence of disease. In such situations it is important that conclusions generalize to both the case and reader populations. A typical design for comparing diagnostic tests is the balanced test×reader×case factorial study design where each image is assigned a disease-confidence rating by each reader using each diagnostic test. Throughout I use *test* to refer to a diagnostic test, modality, or treatment.

Correspondence to: Stephen L. Hillis, steve-hillis@uiowa.edu.

The methods proposed by Obuchowski and Rockette (OR) [1, 2] and Dorfman, Berbaum, andMetz (DBM) [3, 4] are the most commonly used methods for analyzing such multireader multicase studies (often referred to as MRMC studies) and have performed well in simulations. The OR procedure fits a correlated-error test×reader ANOVA to reader-performance outcomes such as the area under the ROC curve (AUC), while the DBM procedure fits a test×reader×case conventional ANOVA to case-specific pseudovalues. Although the two methods have been shown to be equivalent [5, 6] when based on the same procedural parameters, I find the OR procedure more intuitive and its parameters more interpretable because it models observed reader-performance outcomes rather than pseudovalues. For this reason the OR procedure will be the focus of this paper.

Previously published derivations of OR model statistical properties [6] are tedious to derive, do not provide motivation for the model, and have been derived only for the balanced text×reader×case factorial study design. In this paper I show that the OR model is the same as the model for the marginal mean of a conventional ANOVA model with independent errors, where the mean is computed across cases. Viewing the OR model within this marginal-mean ANOVA framework is the basis for the *marginal-mean ANOVA approach* (mm-ANOVA approach), the topic of this paper. This approach (1) provides an intuitive motivation for the OR model, including its covariance-parameter constraints; (2) provides easy derivations of OR test statistics and parameter estimates, as well as their distributions and confidence intervals; and (3) allows for easy generalization of the OR procedure to other study designs.

In particular, I show how one can easily derive OR-type analysis formulas for any balanced study design by following an algorithm which only requires an understanding of conventional ANOVA methods. This development is important because for many situations other designs are more suitable than the text×reader×case factorial study design. For example, diagnostic tests may be mutually exclusive for various reasons, such as high radiation dose or invasiveness of the test, and thus can not be given to each patient; readers may be trained to read under only one of the tests; or power considerations may show that it is advantageous to have replicated readings or to have groups of readers read different cases.

The outline of this paper is as follows. I review the OR method in Section 2. In Sections 3–4 and Appendices A–C I describe and justify steps of an algorithm for motivating the OR model and deriving its properties using the marginal-mean ANOVA approach. Steps are stated in a general form so that analogous OR-type procedures can be formulated for other study designs. In Section 5 I summarize the algorithm and illustrate how the algorithm can be used to develop OR-type procedures for six other study designs. A discussion and concluding remarks are given in Section 6.

2. THE OBUCHOWSKI-ROCKETTE (OR) METHOD

2.1. Design and notation

Throughout this section I assume the data have been collected using a balanced test×reader×case study factorial design. This commonly used diagnostic-radiology study design specifies that each case be subjected to each test, with the resulting images evaluated

once by each reader. In addition, each case is classified as diseased or nondiseased according to an available reference standard. Typically the number of cases is 25–200 while the number of readers is 3–15. Let Z_{ijk} denote a confidence-of-disease rating assigned to the *k*th case by the *j*th reader using the *i*th test. For example, often an ordinal five-level ordinal integer scale or a quasi-continuous 0% to 100% confidence scale is used. The observed rating data consists of the Z_{ijk}, with $i = 1, \ldots, t, j = 1, \ldots, r, k = 1, \ldots, c$, where t is the number of tests, r the number of readers, and c the number of cases.

2.2. Model and test statistic

Let $\hat{\theta}_{ij}$ denote the AUC estimate (or other ROC-curve accuracy estimate) for the *i*th test and *j*th reader. Obuchowski and Rockette [1] use a test × reader factorial ANOVA model for the AUC estimates, but unlike a conventional ANOVA model they allow the errors to be correlated to account for correlation due to each reader evaluating the same cases. Their model, which I refer as the *OR model*, can be written as

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

$i = 1, \ldots, t, j = 1, \ldots, r$, 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. Without loss of generality I assume $\sum_{i=1}^{t}\tau_i=0$. The R_j and $(\tau R)_{ij}$ are assumed to be mutually independent and normally distributed with zero means and respective variances σ_R^2 and σ_{TR}^2. The ε_{ij} are assumed to be normally distributed with zero mean and variance σ_ε^2 and are assumed independent of the R_j and $(\tau R)_{ij}$. Equi-covariance of the errors between readers and tests is assumed, resulting in three possible covariances given by

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

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 × reader effects. Based on clinical considerations Obuchowski and Rockette [1] suggest the following ordering for the covariances:

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

In Section 3.4 I show that these constraints can replaced by the less restrictive constraints

$$\mathrm{Cov}_1 \geq \mathrm{Cov}_3, \mathrm{Cov}_2 \geq \mathrm{Cov}_3, \mathrm{Cov}_3 \geq 0 \quad (3)$$

Alternatively, the model can be described in terms of the error correlations, defined by $\rho_i=\mathrm{Cov}_i/\sigma_\varepsilon^2, i=1,2,3$.

When Cov_2 and Cov_3 are known, the OR statistic for testing the null hypothesis of no test effect (H_0: $\tau_i = 0$; $i = 1, \ldots t$) is given by

$$F^*_{\mathrm{OR}}=\frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)} \quad (4)$$

where MS(*T*) and MS(*T* * *R*) are the test and test × reader mean squares; i.e.,

$\mathrm{MS}(T)=\frac{r}{t-1}\sum_{i=1}^{t}\left(\hat{\theta}_{i\bullet}-\hat{\theta}_{\bullet\bullet}\right)^2$ and

$\mathrm{MS}(T*R)=\frac{1}{(t-1)(r-1)}\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\hat{\theta}_{ij}-\hat{\theta}_{i\bullet}-\hat{\theta}_{\bullet j}+\hat{\theta}_{\bullet\bullet}\right)^2$. A subscript replaced by a dot indicates that values are averaged across the missing subscript index; for example,

$\hat{\theta}_{\bullet\bullet}=\frac{1}{tr}\sum_{i=1}^{t}\sum_{j=1}^{r}\hat{\theta}_{ij}$.

In practice the statistic actually used is

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

where $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ denote estimates for Cov_2 and Cov_3, respectively. Note that (5) incorporates the constraints specified by (3) by setting $\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3$ 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 × reader effects, they can be estimated using methods that treat cases as random but readers as fixed, such as jackknifing, bootstrapping, parametric methods, or the method proposed by DeLong et al [7] for trapezoidal-rule (or empirical) AUC estimates [8]. The OR estimates obtained from averaging corresponding fixed-reader AUC variances and covariances are denoted by $\hat{\sigma}^2_{\varepsilon}$, $\widehat{\mathrm{Cov}}_1$, $\widehat{\mathrm{Cov}}_2$, and $\widehat{\mathrm{Cov}}_3$. Hillis [6] shows that F_{OR} has an approximate $F_{t-1;\mathrm{ddf_H}}$ null distribution, where

$$\mathrm{ddf_H}=\frac{\left\{\mathrm{MS}(T*R)+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]\right\}^2}{\frac{[\mathrm{MS}(T*R)]^2}{(t-1)(r-1)}} \quad (6)$$

More generally, F_{OR} has an $F_{t-1,\mathrm{df_2};\lambda}$ distribution where

$\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma^2_{TR}+\sigma^2_{\tilde{\varepsilon}}-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)}$ and

$\mathrm{df_2}=\frac{\left[\sigma^2_{TR}+\sigma^2_{\tilde{\varepsilon}}-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)\right]^2}{\left[\sigma^2_{TR}+\sigma^2_{\tilde{\varepsilon}}-\mathrm{Cov}_1-\mathrm{Cov}_2+\mathrm{Cov}_3\right]^2/[(t-1)(r-1)]}$.

Letting θ_i denote the expected reader performance measure for test *i* (i.e., $\theta_i = E(\hat{\theta}_{i\bullet})$), an approximate (1 − α) 100% confidence interval for contrast $\sum_{i=1}^{t}l_i\theta_i\left(\sum_{i=1}^{t}l_i=0\right)$ is given by $\sum_{i=1}^{t}l_i\hat{\theta}_{i\bullet}\pm t_{\alpha/2;\mathrm{ddf}_H}\sqrt{\hat{V}}$ where

$\hat{V}=\frac{1}{r}\left(\sum_{i=1}^{t}l_i^2\right)\left\{\mathrm{MS}(T*R)+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]\right\}$. An approximate (1 − α) 100% confidence interval for θ_i, using a standard error computed from all of the data, is given by

$\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2} \sqrt{\hat{V}}$, where $\hat{V}=\frac{1}{tr}\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+tr\max\left(\widehat{\mathrm{Cov}}_2,0\right)\right]$ and $\mathrm{df}_2=\frac{\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+tr\max(\widehat{\mathrm{Cov}}_2,0)\right]^2}{[\mathrm{MS}(R)]^2/(r-1)+[(t-1)\mathrm{MS}[T*R]]^2/[(t-1)(r-1)]}$. Alternatively, an approximate (1 − α) 100% confidence interval for test *i*, using a standard error computed only from data for test *i*, is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2^{(i)}} \sqrt{\hat{V}(i)}$, where $\hat{V}^{(i)}=\frac{1}{r}\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]$ and $\mathrm{df}_2^{(i)}=\frac{\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]^2}{[\mathrm{MS}(R)^{(i)}]^2/(r-1)}$; here MS $(R)^{(i)}$ and $\widehat{\mathrm{Cov}}_2^{(i)}$ are computed only from test *i* data. I recommend this latter formula for single AUC confidence intervals, since it does not depend on assuming equal error covariances and variances for each test. All of these results have been previously presented [6].

Expected mean squares are given in Table 1a; proofs for these results are given by Hillis [6]. Expressions for the variance components, in terms of the expected mean squares and covariances are presented in Table 1b; these relationships follow directly from Table 1a. Estimated variance components result by replacing expected mean squares by mean squares and covariance parameters by estimates; for example,

$$\hat{\sigma}^2_{TR}=\mathrm{MS}(T*R)-\hat{\sigma}^2_{\varepsilon}+\widehat{\mathrm{Cov}}_1+\max\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3,0\right)$$

Typically the variance component estimates are changed to zero if the computed values are negative.

2.3. Real-data example

To illustrate the OR method for the factorial design, I compare reader AUCs for hard- and soft-copy computed radiography chest images selected randomly from a medical intensive care unit. In the study [9] four radiologists blindly read both hard- and soft-copy images obtained with computed radiography from the same patients. Six months separated the end of the hard-copy readings and the start of the soft-copy readings. A five-point ordinal scale was used to rate the likelihood of presence of the condition (which I will refer to as “disease”) implied by the reason for requesting the corresponding examination. Ninety-five images, consisting of 29 diseased and 66 nondiseased images, were read under each test condition.

The analysis of this study using empirical AUC estimates and jackknife covariance estimates is displayed in Table 2. The AUCs for soft- and hard-copy images, averaged across the four readers, are 0.804 and 0.841, respectively. The test for the null hypothesis of no test effect (i.e., the population average AUC across readers is the same for soft- and hard-copy images) is not significant (F_{OR} = 6.01, ddf_H = 3, p = .092); the 95% confidence interval for the difference of the population AUCs (hard- minus soft-copy) is (−0.011, 0.086). Parts (i) and (j) give 95% confidence intervals for the single-test AUCs, based on all

of the data and only on data for the specific test, respectively. The confidence intervals from the two methods are similar; this is expected because the AUCs are similar.

Although this study showed a nonsignificant difference between soft- and hard-copy image reader performance, the confidence interval for the difference of the AUCs showed a difference as large as 0.086 to be commensurate with the data. In such a situation, the researcher may decide to design a future study that would produce a more precise estimate of the difference. Increased precision could result from an increase in the number of cases, the number of readers, or from replicated readings where each reader reads each image 2 or more times. If increasing the number of cases and readers is not feasible, then a replicated study is a natural choice for increasing power; however, OR analysis methodology has been developed only for the nonreplicated factorial design. I use the algorithm described in this paper to derive the OR-type procedure for the replicated factorial design, including the test-statistic nonnull distribution, which allows for power and sample size estimation. Using this result, I illustrate efficiency computations comparing the nonreplicated and replicated designs in Section 5.6.

In this study the same radiologists also similarly rated 95 hard-copy chest images obtained with screen-film; these images were from different patients than the computed radiographs. Because the original OR method assumes a factorial study design with readers reading the same cases under each test, it cannot be used to compare the screen-film AUC outcomes with the AUC outcomes from either the soft- or hard-copy computed radiograph images. In Section 5.3 I show how the OR approach can be adapted for this situation, which represents a split-plot study design with cases nested within test, and illustrate the analysis of these data.

2.4. Previous derivations of OR properties

Derivations of OR-procedure properties have previously been derived starting with the OR model (1, 2). For what is essentially the OR model, Pavur and Nath [10] show that, for testing the null hypothesis of equal tests, the F statistic that is appropriate when the errors are independent can be used if corrected by a multiplicative factor. The multiplicative factor is a function of the correlations, which are assumed known, and the distribution for this corrected F statistic is the same as for the uncorrected F statistic when the errors are independent. The approach taken by Obuchowski and Rockette [1] was to modify this result by replacing the assumed-known correlations by estimated correlations. This approach yielded valid ANOVA statistics but unsatisfactory degrees of freedom, resulting in overly conservative tests [6]. Alternatively, Hillis [6] directly derived properties, but the proofs are tedious and nonintuitive.

3. MM-ANOVA APPROACH – STEP 1: DERIVE THE MM-ANOVA MODEL

In Sections 3–4 and Appendices A–C I show how the properties of the OR model can easily be derived using an algorithm, based on the mm-ANOVA approach, that only requires knowing how to determine conventional ANOVA test statistics and expected mean squares. I describe and illustrate the steps in the algorithm for the typical balanced test×reader×case study design discussed in the previous section. The steps are stated in a general form so that

they can be applied to other balanced study designs. The mm-ANOVA approach and corresponding algorithm have not been previously described and are the main contribution of this paper.

3.1. Step 1a: Define the conventional ANOVA model that corresponds to the study design as if each reader-performance measure was the mean of case outcomes

Let Y_{ijk} denote a *hypothetical* outcome for test i, reader j, and case k. For our purposes Y_{ijk} is used only to illustrate the marginal ANOVA model approach; i.e., it does not represent an actual study outcome and should be distinguished from the observed rating Z_{ijk}. I assume that the Y_{ijk} follow a three-way conventional ANOVA model that corresponds to the study design.

Thus the distribution of Y_{ijk} is given by the following test × reader × case ANOVA model that treats test as a fixed factor and reader and case as random factors:

$$Y_{ijk}=\mu+\tau_i+R_j+C_k+(\tau R)_{ij}+(\tau C)_{ik}+(RC)_{jk}+(\tau RC)_{ijk}+\varepsilon_{ijk} \quad (7)$$

$i = 1, \ldots, t, j = 1, \ldots, r, k = 1, \ldots, c$, where τ_i denotes the fixed effect of test i with $\sum_{i=1}^{t}\tau_i=0$, R_j denotes the random effect of reader j, C_k denotes the random effect of case k, the multiple symbols in parentheses denote random interactions, and ε_{ijk} is the error term. The random effects are assumed to be mutually independent and normally distributed with zero means and respective variances $\sigma_R^2, \sigma_C^2, \sigma_{TR}^2, \sigma_{TC}^2, \sigma_{RC}^2, \sigma_{TRC}^2$, and σ_ε^2. Because there are no replications, for estimation purposes σ_{TRC}^2 and σ_ε^2 are inseparable; hence I define

$$\sigma^2=\sigma_{TRC}^2+\sigma_\varepsilon^2$$

Results for this model, such as mean square distributional properties and ANOVA test statistics, are well known (e.g., [11]) and will be stated without references.

3.2. Step 1b: From the conventional ANOVA model defined in step 1a, derive the mm-ANOVA model by averaging across cases and defining the mm-ANOVA model error term equal to the mean, across cases, of the sum of the conventional ANOVA model error term and random effects involving case

I say that a random effect "involves case" if it is subscripted according to case. Let $\tilde{Y}_{ij}$ denote the marginal mean resulting from averaging over cases; i.e.,

$$\tilde{Y}_{ij}=Y_{ij\bullet} \quad (8)$$

I use the term *marginal-mean ANOVA model* (mm-ANOVA model) to refer to the model implied by the conventional 3-way ANOVA model (7) for the marginal mean (8). It follows from (7) that

$$\tilde{Y}_{ij}=\mu+\tau_i+R_j+(\tau R)_{ij}+\tilde{\varepsilon}_{ij} \quad (9)$$

where

$$\tilde{\varepsilon}_{ij}=C_{\bullet}+(\tau C)_{i\bullet}+(RC)_{j\bullet}+(\tau RC)_{ij\bullet}+\varepsilon_{ij\bullet} \quad (10)$$

the R_j and $(\tau R)_{ij}$ are mutually independent and normally distributed with zero means and respective variances σ^2_R and σ^2_{TR}, and the $\tilde{\varepsilon}_{ij}$ are independent of the R_j and $(\tau R)_{ij}$.

3.3. Step 1c: Express the mm-ANOVA model error variance and covariances in terms of the conventional ANOVA model variance components

From (10) it follows that the $\tilde{\varepsilon}_{ij}$ are normally distributed with mean 0, variance

$$\sigma^2_{\tilde{\varepsilon}}=\frac{1}{c}(\sigma^2_C+\sigma^2_{TC}+\sigma^2_{RC}+\sigma^2_{TRC}+\sigma^2_{\varepsilon}) \quad (11)$$

and equi-correlated with

$$\mathrm{Cov}_1 \equiv \mathrm{cov}(\tilde{\varepsilon}_{ij},\tilde{\varepsilon}_{i'j})=\frac{1}{c}(\sigma^2_C+\sigma^2_{RC}) \quad (12)$$

$$\mathrm{Cov}_2 \equiv \mathrm{cov}(\tilde{\varepsilon}_{ij},\tilde{\varepsilon}_{ij'})=\frac{1}{c}(\sigma^2_C+\sigma^2_{TC}) \quad (13)$$

and

$$\mathrm{Cov}_3 \equiv \mathrm{cov}(\tilde{\varepsilon}_{ij},\tilde{\varepsilon}_{i'j'})=\frac{1}{c}\sigma^2_C \quad (14)$$

where $i \neq i'$ and $j \neq j'$.

3.4. Step 1d: Determine the mm-ANOVA model covariance constraints implied by step 1c

The covariance constraints given by (3) follow from (12–14). Thus the mm-ANOVA model for $\tilde{Y}_{ij}$ is defined by (9) and (3). It also follows from (11–14) that $\sigma^2_{\tilde{\varepsilon}} \geq (\mathrm{Cov}_1+\mathrm{Cov}_2+\mathrm{Cov}_3)$, but I do not include this constraint as part of the definition of the mm-ANOVA model because this constraint is implied from the relationship $\mathrm{Var}(\tilde{\varepsilon}_{11}-\tilde{\varepsilon}_{12}-\tilde{\varepsilon}_{21}+\tilde{\varepsilon}_{22}) \geq 0$.

3.5. Remarks

3.5.1. One-to-one relationship between parameters of the 3-way conventional ANOVA and corresponding mm-ANOVA models—

In terms of the mm-ANOVA model parameters (μ, τ_i, σ^2_R, σ^2_{TR}, $\sigma^2_{\tilde{\varepsilon}}$, Cov_1, Cov_2, and Cov_3), the parameters for the corresponding three-way ANOVA model (7) are given by μ, τ_i, $\sigma^2_R, \sigma^2_{TR}, \sigma^2_{\varepsilon}=c(\sigma^2_{\tilde{\varepsilon}}-\mathrm{Cov}_1-\mathrm{Cov}_2-\mathrm{Cov}_3), \sigma^2_C=c\mathrm{Cov}_3, \sigma^2_{TC}=c(\mathrm{Cov}_2-\mathrm{Cov}_3)$, and $\sigma^2_{RC}=c(\mathrm{Cov}_1-\mathrm{Cov}_3)$. Thus there is a one-to-one relationship between the parameters of the two models. *Hence for any mm-ANOVA model, defined by* (9) *and* (3), *there is a*

corresponding conventional 3-way ANOVA model (7) *that implies that model for the marginal means.* These relationships between the two models are presented in Table 3.

3.5.2. Equivalence of the OR and mm-ANOVA models—Note that the mm-ANOVA model (9, 3) has the same form as the OR model (1, 2), with the only difference being that the mm-ANOVA model covariance constraints (3) are less restrictive. Since the OR covariance constraints (2) were suggested by Obuchowski and Rockette [1] based only on clinical considerations, to simplify comparison of the models *I now modify the definition of the OR model to include the less restrictive mm-ANOVA model constraints* (3); *i.e., the OR model is now considered to be defined by* equations (1) *and* (3). *With this change the OR and the mm-ANOVA model become equivalent.*

3.5.3. Definition of the mm-ANOVA approach—Because the OR and mm-ANOVA model are identical, statistical properties for the ROC accuracy estimates, the $\hat{\theta}_{ij}$, are the same as for the marginal means, the $\tilde{Y}_{ij}$, for an mm-ANOVA model having the same parameter values as the OR model. *The mm-ANOVA approach consists of deriving statistical properties for the OR model* (1, 3) *by recognizing that it is equivalent to the mm-ANOVA model* (9, 3), *and then deriving properties of the mm-ANOVA model by utilizing its relationship with the conventional three-way ANOVA model.* The advantage of this approach is that properties of the conventional three-way ANOVA model are well known.

3.5.4. Motivation for the OR model—The mm-ANOVA approach provides an intuitive motivation for the OR model (1, 3) as follows. Suppose, hypothetically, that the reader performance outcome $\hat{\theta}_{ij}$ is the mean of case-specific outcomes; that is, suppose that $\hat{\theta}_{ij} = Y_{ij\bullet}$ for some outcome Y_{ijk}, with $k = 1, \ldots, c$. A typical way to account for variation in $\hat{\theta}_{ij}$ due to readers and cases would be to assume the three-way ANOVA model (7), which implies the mm-ANOVA model (9, 3) and hence also the equivalent OR model (1, 3) for $\hat{\theta}_{ij}$. Of course, in practice $\hat{\theta}_{ij}$ is not a marginal mean, but rather a nonlinear function of the case-specific confidence-of-disease ratings and truth-state (i.e., reference standard) indicator values. However, the mm-ANOVA approach shows that the OR model accounts for reader and case variation using the covariance structure implied by a conventional three-way ANOVA model, as if the accuracy estimate was a marginal mean.

4. MM-ANOVA APPROACH – STEP 2: DERIVE THE MM-ANOVA MODEL TEST STATISTIC AND ITS NULL DISTRIBUTION FOR A HYPOTHESIS EXPRESSED IN TERMS OF TEST ACCURACIES

In this section I show how to derive the mm-ANOVA model test statistic and its null distribution for testing the null hypothesis of equal test accuracies. I define *test accuracy* as the expected reader-performance measure for a particular test level. However, more generally these steps can be applied to any hypothesis that can be expressed in terms of linear functions of expected reader-performance outcomes.

4.1. Step 2a: State the hypothesis of interest in terms of the mm-ANOVA model

For the mm-ANOVA model (9, 3) let θ_i denote the test accuracy for test *i*; i.e., $\theta_i = E(\tilde{Y}_{i\bullet})$ is the expected reader-performance outcome for test *i* across the population of readers. The hypothesis of interest is the global null hypothesis of equal test accuracies, i.e., $H_0 : \theta_1 = \ldots = \theta_t$, or equivalently, $H_0 : \tau_1 = \ldots = \tau_t = 0$.

4.2. Step 2b: Express the hypothesis from step 2a in terms of the conventional ANOVA model

Noting that

$$\theta_i = E\left(\tilde{Y}_{i\bullet}\right) = E\left(Y_{i\bullet\bullet}\right) = \mu + \tau_i$$

it follows that $H_0 : \theta_1 = \ldots = \theta_t$ is equivalent to $H_0 : \tau_1 = \ldots = \tau_t = 0$ for the conventional ANOVA model (7).

4.3. Step 2c: Create the expected-mean-square table for the conventional ANOVA model

Let MS(*T*), MS(*R*), and MS(*C*) denote the conventional ANOVA mean squares due to test, reader, and case, respectively, with interaction mean squares notated in the usual manner. The expected mean squares for the conventional ANOVA model are presented in Table 4. These relationships will be utilized in other steps.

4.4. Step 2d: Determine the conventional ANOVA F statistic corresponding to the step 2b hypothesis

The conventional ANOVA test statistic for testing for $H_0 : \tau_1 = \ldots = \tau_t = 0$ is given by

$$F = \frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R) + \mathrm{MS}(T*C) - \mathrm{MS}(T*R*C)} \quad (15)$$

I refer to *F* as an *ANOVA statistic* because its numerator and denominator have the same expectation under H_0, but the numerator has a larger expectation than the denominator under $H_1 : \tau_i \neq \tau_j$ for some $i \neq j$.

4.5. Step 2e: Express mm-ANOVA mean squares in terms of conventional ANOVA mean squares

For the mm-ANOVA model let $\tilde{\mathrm{MS}}(T), \tilde{\mathrm{MS}}(R)$, and $\tilde{\mathrm{MS}}(T*R)$ denote the test, reader, and test×reader mean squares; i.e.,

$$\tilde{\mathrm{MS}}(T) = \frac{r}{t-1}\sum_{i=1}^{t}\left(\tilde{Y}_{i\bullet} - \tilde{Y}_{\bullet\bullet}\right)^2, \tilde{\mathrm{MS}}(R) = \frac{t}{r-1}\sum_{j=1}^{r}\left(\tilde{Y}_{\bullet j} - \tilde{Y}_{\bullet\bullet}\right)^2 \text{ and}$$

$\tilde{\mathrm{MS}}(T*R)=\frac{1}{(t-1)(r-1)}\sum_{i=1}^{t}\sum_{j=1}^{r}\left(\tilde{Y}_{ij}-\tilde{Y}_{i\bullet}-\tilde{Y}_{\bullet j}+\tilde{Y}_{\bullet\bullet}\right)^2$. Noting that $\mathrm{MS}(T)=\frac{rc}{t-1}\sum_{i=1}^{t}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2, \mathrm{MS}(R)=\frac{tc}{r-1}\sum_{i=1}^{t}(Y_{\bullet j\bullet}-Y_{\bullet\bullet\bullet})^2, \mathrm{MS}(T*R)=\frac{c}{(t-1)(r-1)}\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij\bullet}-Y_{i\bullet\bullet}-Y_{\bullet j\bullet}+Y_{\bullet\bullet\bullet})^2$, it follows that

$$\tilde{\mathrm{MS}}(T)=\frac{1}{c}\mathrm{MS}(T) \quad (16)$$

$$\tilde{\mathrm{MS}}(R)=\frac{1}{c}\mathrm{MS}(R)$$

$$\tilde{\mathrm{MS}}(T*R)=\frac{1}{c}\mathrm{MS}(T*R) \quad (17)$$

4.6. Step 2f: Express F from step 2d in terms of mm-ANOVA model mean squares and U, where U is a linear function of conventional ANOVA model mean squares that involve case

It follows from (16–17) that (15) can be written in the form

$$F=\frac{\tilde{\mathrm{MS}}(T)}{\tilde{\mathrm{MS}}(T*R)+U} \quad (18)$$

where

$$U=\frac{1}{c}[\mathrm{MS}(T*C)-\mathrm{MS}(T*R*C)]$$

Note that *U* is a linear function of conventional ANOVA model mean squares involving case and (18) is an ANOVA statistic.

4.7. Step 2g: Express E (U) in terms of conventional ANOVA model variance components, and then in terms of mm-ANOVA model error covariance parameters using the relationships from step 1c

From Table 4 we have $E[\mathrm{MS}(T*C)]=r\sigma^2_{TC}+\sigma^2$ and *E* [MS (*T* * *R* * *C*)] = σ^2. It follows that

$$E(U)=\frac{1}{c}E[\mathrm{MS}(T*C)-\mathrm{MS}(T*R*C)]=\frac{r}{c}\sigma^2_{TC} \quad (19)$$

Using (13) and (14) we can write the right side of (19) in terms of the mm-ANOVA covariances: $\frac{r}{c}\sigma^2_{TC}=r(\mathrm{Cov}_2-\mathrm{Cov}_3)$. Hence

$$E(U)=r(\mathrm{Cov}_2-\mathrm{Cov}_3) \quad (20)$$

4.8. Step 2h: Modify F (18) from step 2f to produce the mm-ANOVA statistic F^*_{OR} by replacing U by E (U), expressed as a linear function of mm-ANOVA covariance parameters

Replacing *U* in equation (18) by its expectation (20) results in

$$F^*_{\mathrm{OR}}=\frac{\tilde{\mathrm{MS}}(T)}{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)} \quad (21)$$

which is the OR test statistic F^*_{OR} (4) when we treat the $\tilde{Y}_{ij}$ as the OR model outcomes $\hat{\theta}_{ij}$. Because (18) is an ANOVA statistic, it follows that F^*_{OR} (21) is also an ANOVA statistic.

4.9. Step 2i: Derive F_{OR} by replacing covariance parameters in FOR* by estimates that take into account the constraints from step 1d

An obvious estimate of $\mathrm{Cov}_2-\mathrm{Cov}_3$ that takes into account covariance constraints (3) is given by $\max\left[\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]$, where $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are estimates as discussed in Section 2.2. Replacing $\mathrm{Cov}_2-\mathrm{Cov}_3$ in (21) by this estimate results in

$$F_{\mathrm{OR}}=\frac{\tilde{\mathrm{MS}}(T)}{\tilde{\mathrm{MS}}(T*R)+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]} \quad (22)$$

which is the OR statistic F_{OR} (5) when we replace the $\tilde{Y}_{ij}$ by the OR model outcomes $\hat{\theta}_{ij}$.

4.10. Step 2j: Determine the approximate null distribution of F_{OR}

Null-distribution result—Write the denominator of F_{OR} in the form

$$b\left(\sum_{i=1}^{I}a_i\tilde{\mathrm{MS}}_i+\hat{d}\right) \quad (23)$$

where the $\tilde{\mathrm{MS}}_i$, $i = 1, \ldots, I$ are mm-ANOVA mean squares, $\hat{d}$ is a function of the covariance parameter estimates and the a_i and b are constants. Then F_{OR} will have an approximate $F_{\mathrm{df}_1,\mathrm{df}_2}$ null distribution, where df_1 is the numerator degrees of freedom for the conventional ANOVA model test statistic in step 2d and df_2 is given by

$$\mathrm{df}_2=\frac{\left[\sum_{i=1}^{I}a_i\tilde{\mathrm{MS}}_i+\hat{d}\right]^2}{\sum_{i=1}^{I}\frac{\left[a_i\tilde{\mathrm{MS}}_i\right]^2}{\mathrm{df}\left(\tilde{\mathrm{MS}}_i\right)}} \quad (24)$$

where $\mathrm{df}\left(\tilde{\mathrm{MS}}_i\right)$ is the degrees of freedom for $\tilde{\mathrm{MS}}_i$, and hence also for MS_i. I have stated this result generally so that it can be easily applied to other designs. See Appendix A for a derivation of this result.

To apply this result to the balanced test×reader×case factorial study design, note that the denominator of F_{OR} (22) is given by (23) with $I = 1$, $a_1 = 1$, $b = 1$, $\tilde{\mathrm{MS}}_1 = \tilde{\mathrm{MS}}(T*R)$, and $\hat{d} = \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right), 0\right]$. Using (24), the null-distribution result states that F_{OR} (22) has an approximate F_{t-1,df_2} null distribution, where

$$\mathrm{df}_2 = \frac{\left\{\tilde{\mathrm{MS}}(T*R) + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right), 0\right]\right\}^2}{\frac{\left[\tilde{\mathrm{MS}}(T*R)\right]^2}{(t-1)(r-1)}} \quad (25)$$

Note that the equation for df_2 (25), with $\tilde{Y}_{ij}$ replaced by $\hat{\theta}_{ij}$, is the same as the equation for ddf_H (6) for the OR model.

4.11. Remark: Derivation of mm-ANOVA expected mean square and variance component expressions

For the mm-ANOVA model an expected mean square table, such as Table 1a, can be created as follows. Write the mm-ANOVA expected mean squares in terms of the conventional ANOVA variance components and fixed effects using the relationships given in steps 2c and 2e. For example, for the factorial model we have

$$E\left[\tilde{\mathrm{MS}}(T)\right] = \frac{1}{c}E[\mathrm{MS}(T)] = \frac{1}{c}\left[\frac{rc}{(t-1)}\sum_{i=1}^{t}\tau_i^2 + c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2\right] \quad (26)$$

From step 1c it follows that the conventional ANOVA variance components in (26) involving case (i.e., the corresponding random effects are subscripted according to case) can be written in terms of the mm-ANOVA covariances: $\sigma_{TC}^2 = c(\mathrm{Cov}_2 - \mathrm{Cov}_3)$ and $\sigma^2 = c(\sigma_{\tilde{\varepsilon}}^2 - \mathrm{Cov}_1 - \mathrm{Cov}_2 + \mathrm{Cov}_3)$. Replacing these variance components in (26) by their corresponding mm-ANOVA covariance expressions yields $\tilde{\mathrm{MS}}(T) = \frac{r}{(t-1)}\sum_{i=1}^{t}\tau_i^2 + \sigma_{TR}^2 + \sigma_{\tilde{\varepsilon}}^2 - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)$, the first line in Table 1a. Similarly, the other expressions in Table 1a can be derived. A table of mm-ANOVA variance component formulas, such as Table 1b, can then be created from the mm-ANOVA expected mean square table by solving for the variance components.

5. Mm-ANOVA algorithm summary and examples

In Sections 3–4 steps 1 and 2 of the mm-ANOVA algorithm were presented. These two steps illustrated the essence of the mm-ANOVA approach. Steps 3 and 4, which are presented later in Appendices B and C, extend this approach by showing how to derive confidence intervals and the non-null distribution of the test statistic.

Table 5 presents a succinct summary of the mm-ANOVA algorithm. This summary is intended to make it easy to use the algorithm to determine the properties of OR-type models corresponding to other study designs. Note that Table 5 shows the steps for deriving the confidence interval formula, not only for a linear combination of test accuracy parameters,

but also for a single accuracy parameter. Table 6 illustrates the application of Table 5 to the typical test×reader×case study design previously discussed in Sections 3 and 4.

Using the algorithm in Table 5, I derive results for several other study designs and summarize these results in the remainder of this section. For each study design the corresponding algorithm results, in a format similar to Table 6, are presented in the referenced supplementary tables that are available in the online version of this article. Note that in the summaries below the reader performance measure is denoted by $\hat{\theta}_{ij}$ instead of $\tilde{Y}_{ij}$ to make it clear that, although these are mm-ANOVA models, the outcome is not restricted to a marginal mean but can be any reader-performance measure. In addition, I omit the tilde symbol over the mean squares and error term since it is clear that they are for the mm-ANOVA model rather than the corresponding conventional ANOVA model. Standard nesting notation is used; e.g., subscript (*i*) *j* denotes that the factor indexed by *j* is nested within the factor indexed by *i*, and MS[*R* (*T*)] is the mean square for reader nested within test.

5.1. Example 1: Reader×case study design (one test)

In this study design there is only one test and each reader reads each case. Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S1. The derivation begins with a conventional reader×case study-design ANOVA model that treats reader and case as random factors and includes their interaction. Averaging across cases produces the corresponding mm-ANOVA model: a one-way ANOVA model with reader as its only factor.

This mm-ANOVA model is given by $\hat{\theta}_j = \mu + R_j + \varepsilon_{ij}$, $j = 1, \ldots, r$, where r is the number of readers. The R_j are mutually independent and normally distributed with zero mean and variance σ_R^2; the ε_{ij} are normally distributed with zero mean and variance σ_ε^2 and are independent of the R_j; and $\text{Cov}_2 \equiv \text{Cov}(\varepsilon_j, \varepsilon_{j'}) \geq 0$, $j \neq j'$. Thus reader is a random factor and the covariance between error terms is assumed constant. Because there is only one test, only the formula for computing a confidence interval for the single test accuracy is presented.

An approximate (1 − a) 100% confidence interval for a single test accuracy, $\theta = E(\hat{\theta}_j)$, is given by $\hat{\theta}_\bullet \pm t_{\alpha/2;\text{df}_2}\sqrt{\hat{V}}$, where

$$\hat{V} = \frac{1}{r}\left[\text{MS}(R) + r\max\left(\widehat{\text{Cov}}_2, 0\right)\right], \quad \text{df}_2 = \frac{\left[\text{MS}(R) + r\max\left(\widehat{\text{Cov}}_2, 0\right)\right]^2}{[\text{MS}(R)]^2/(r-1)}, \text{ and}$$

$$\text{MS}(R) = \frac{1}{r}\sum_{j=1}^{r}\left(\hat{\theta}_j - \hat{\theta}_\bullet\right)^2.$$

A hypothesis test for the single test accuracy can be based on this confidence interval. Although Hillis [6] discusses this single-test confidence interval formula, he does not provide a derivation of the result.

This confidence interval result can also be used with the test×reader×case study design to yield single test confidence intervals, each based only on data for the corresponding test, as was illustrated in the analysis of the example data in Section 2.3. Because properties of this confidence interval do not depend on assumptions about the variance components and

covariances corresponding to the other tests, we expect these single-test confidence intervals to be more robust than those where the standard error is based on all of the data.

5.2. Example 2: Reader-nested-within-test study design

In this study design readers read images from only one test; i.e., readers are nested within test. This study design is natural when readers are trained to read under only one of the tests. The study design is balanced with an equal number of readers reading all cases using each test. Thus reader is nested within test and is crossed with case. Obuchowski [12] discusses this design and refers to this as a *paired-case, unpaired-reader* design. This can be viewed as a split-plot design with readers being the "whole plots," case the split-plot (or within-plot) factor, and test the whole-plot (or between-plot) factor. This design is schematically illustrated in Table 7a.

Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S2. The derivation begins with a conventional split-plot ANOVA model corresponding to the study design (i.e., with reader nested within test and crossed with case) that treats reader and case as random factors and includes all possible interactions. Averaging across cases produces the corresponding mm-ANOVA model: a reader-nested-within-test ANOVA model with reader as a random factor.

The mm-ANOVA model is given by $\hat{\theta_{ij}} = \mu + \tau_i + R_{(i)j} + \varepsilon_{ij}$, $i = 1, \ldots, t$, $j = 1, \ldots, r$ where t is the number of tests, r is the number of readers, τ_i denotes the fixed effect of test, and $\sum_{i=1}^{t} \tau_i = 0$. The reader effects, the $R_{(i)j}$, are mutually independent and normally distributed with zero mean and variance $\sigma^2_{R(T)}$, where "$R(T)$" is read "reader nested within test". The ε_{ij} are normally distributed with zero mean and variance σ^2_{ε}. The ε_{ij} are independent of the $R_{(i)j}$; $Cov_2 = Cov(\varepsilon_{ij}, \varepsilon_{i\hat{j}})$ with $j \neq j'$ and $Cov_3 = Cov(\varepsilon_{ij}, \varepsilon_{ij})$ with $i \neq i'$, with $Cov_2 \geq Cov_3 \geq 0$.

Thus there are two error covariances, Cov_2 and Cov_3, $Cov_2 \geq Cov_3 \geq 0$, defined as the covariances between errors for the same test and different readers, and for different tests and different readers, respectively. Note that the definition $Cov_3 \equiv Cov(\varepsilon_{ij}, \varepsilon_{i'j'})$, $i \neq i'$ does not require $j \neq j'$ because $i \neq i'$ implies different readers. There is no Cov_1 parameter because the design does not allow for one reader reading under two tests.

Let $\theta_i \equiv E(\hat{\theta_{i\bullet}})$ denote the expected reader performance measure for test i. The test statistic for the null hypothesis of equal test accuracies ($H_0 : \theta_1 = \ldots = \theta_t$) is

$$F_{\mathrm{OR}} = \frac{\mathrm{MS}(T)}{\mathrm{MS}[R(T)] + r\max\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3, 0\right)}$$

where MS(T) is defined as for the factorial model and

$\mathrm{MS}[R(T)] = \frac{1}{t(r-1)} \sum_{i=1}^{t} \sum_{j=1}^{r} \left(\hat{\theta}_{ij} - \hat{\theta}_{i\bullet}\right)^2$. Under H_0, $F_{\mathrm{OR}} \sim^{\cdot} F_{t-1,\mathrm{df}_2}$ where

$$\mathrm{df}_2=\frac{\left[\mathrm{MS}[R(T)]+r\max\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3,0\right)\right]^2}{\{\mathrm{MS}[R(T)]\}^2/[t(r-1)]} \quad (27)$$

More generally, $F_{\mathrm{OR}} \sim^{\cdot} F_{t-1,\mathrm{df}_2;\lambda}$, where $\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma^2_{R(T)}+\sigma^2_{\varepsilon}+(r-1)\mathrm{Cov}_2-r\mathrm{Cov}_3}$ and $\mathrm{df}_2=\frac{\left[\sigma^2_{R(T)}+\sigma^2_{\varepsilon}+(r-1)\mathrm{Cov}_2-r\mathrm{Cov}_3\right]^2}{\left(\sigma^2_{R(T)}+\sigma^2_{\varepsilon}-\mathrm{Cov}_2\right)^2/[t(r-1)]}$.

An approximate (1 − α) 100% confidence interval for contrast $\sum_{i=1}^{t} l_i\theta_i$ is given by $\sum_{i=1}^{t} l_i\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$ where $\hat{V}=\frac{1}{r}\left(\sum_{i=1}^{t} l_i^2\right)\left\{\mathrm{MS}[R(T)]+r\max\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3,0\right)\right\}$ and df_2 is given by (27). An approximate (1 − α) 100% confidence interval for θ_i is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where $\hat{V}=\frac{1}{r}\left\{\mathrm{MS}[R(T)]+\max\left(r\widehat{\mathrm{Cov}}_2,0\right)\right\}$ and $\mathrm{df}_2=\frac{\left\{\mathrm{MS}[R(T)]+\max\left(r\widehat{\mathrm{Cov}}_2,0\right)\right\}^2}{\{\mathrm{MS}[R(T)]\}^2/[t(r-1)]}$. Alternatively, an approximate (1 − α) 100% confidence interval for θ_i, using a standard error computed only from data for test *i*, is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2^{(i)}}\sqrt{\hat{V}^{(i)}}$, where $\hat{V}=\frac{1}{r}\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]$ and $\mathrm{df}_2^{(i)}=\frac{\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]^2}{\left[\mathrm{MS}(R)^{(i)}\right]^2/(r-1)}$, where MS $(R)^{(i)}$ and $\widehat{\mathrm{Cov}}_2^{(i)}$ are computed only from test *i* data; note that this is the result from Section 5.1.

5.3. Example 3: Case-nested-within-test split-plot study design

In this study design each case is imaged under only one test, with the same number of cases imaged for each test. Each reader interprets all of the images from each test. This is often called a *paired-reader, unpaired-case* design. Obuchowski [12] notes that this design is needed when the diagnostic tests are mutually exclusive, e.g., if they are invasive, administer a high radiation dose, or carry a risk of contrast reactions. This can be viewed as a split-plot design with cases being the whole plots, reader the split-plot factor, and test the whole-plot factor. This design is schematically illustrated in Table 7b.

Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S3. The derivation begins with a conventional split-plot ANOVA model corresponding to the study design that treats reader and case as random factors and includes all possible interactions. Averaging across cases produces the corresponding mm-ANOVA model, which is the same as the factorial mm-ANOVA model but with Cov_1 and Cov_3 constrained to zero; i.e., the model is defined by equation (1) and constraints $\mathrm{Cov}_2 \geq 0$, $\mathrm{Cov}_1 = \mathrm{Cov}_3 = 0$. It follows that hypotheses-test, confidence-interval and sample-size formulas can be derived from those for the factorial model by setting $\mathrm{Cov}_1 = \mathrm{Cov}_3 = 0$.

Thus the test statistic for the null hypothesis of equal test accuracies is

$$F_{\mathrm{OR}}=\frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R)+\max\left(r\widehat{\mathrm{Cov}}_2,0\right)}$$

Under H_0, $F_{\mathrm{OR}} \sim^{\cdot} F_{t-1,\mathrm{df}_2}$ where

$$\mathrm{df}_2=\frac{\left\{\mathrm{MS}(T*R)+\max\left(r\widehat{\mathrm{Cov}}_2,0\right)\right\}^2}{[\mathrm{MS}(T*R)]^2/[(t-1)(r-1)]} \quad (28)$$

More generally, $F_{\mathrm{OR}} \sim^{\cdot} F_{t-1,\mathrm{df}_2;\lambda}$, where $\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma_{TR}^2+\sigma_{\varepsilon}^2+(r-1)(\mathrm{Cov}_2)}$ and

$\mathrm{df}_2=\frac{\left[\sigma_{TR}^2+\sigma_{\varepsilon}^2+(r-1)(\mathrm{Cov}_2)\right]^2}{\left[\sigma_{TR}^2+\sigma_{\varepsilon}^2-\mathrm{Cov}_2\right]^2/[(t-1)(r-1)]}$.

Letting θ_i denote $E(\hat{\theta_{i\bullet}})$, an approximate $(1-\alpha)$ 100% confidence interval for contrast $\sum_{i=1}^{t} l_i\theta_i$ is given by $\sum_{i=1}^{t} l_i\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where df_2 is given by (28) and $\hat{V}=\frac{1}{r}\left(\sum_{i=1}^{t} l_i^2\right)\left\{\mathrm{MS}(T*R)+\max\left[r\widehat{\mathrm{Cov}}_2,0\right]\right\}$. An approximate $(1-\alpha)$ 100% confidence interval for θ_i is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where

$\hat{V}=\frac{1}{tr}\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+tr\max\left(\widehat{\mathrm{Cov}}_2,0\right)\right]$ and

$\mathrm{df}_2=\frac{\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+tr\max\left(\widehat{\mathrm{Cov}}_2,0\right)\right]^2}{[\mathrm{MS}(R)]^2/(r-1)+\{(t-1)\mathrm{MS}[T*R]\}^2/[(t-1)(r-1)]}$. Alternatively, an approximate $(1-\alpha)$ 100% confidence interval for θ_i, using a standard error computed only from data for test *i*, is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2^{(i)}}\sqrt{\hat{V}^{(i)}}$, where

$\hat{V}=\frac{1}{r}\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]$ and $\mathrm{df}_2^{(i)}=\frac{\left[\mathrm{MS}(R)^{(i)}+r\max\left(\widehat{\mathrm{Cov}}_2^{(i)},0\right)\right]^2}{[\mathrm{MS}(R)^{(i)}]^2/(r-1)}$, where MS $(R)^{(i)}$ and $\widehat{\mathrm{Cov}}_2^{(i)}$ are computed only from test *i* data. Note that these single-test confidence-interval formulas are the same as those for the factorial design.

5.3.1. Real-data example—Using the Kundel et al [9] data that were discussed in Section 2.3, I now compare soft-copy computed radiographs with screen-film radiographs. The images are from different patients for each type of radiograph, with 95 images in each group (soft-copy computed radiograph: 66 nondiseased, 29 diseased; screen-film radiograph: 68 nondiseased, 27 diseased). Because the images for each method are from different patients, this is an example of a case-nested-within-test study design. The analysis of this study using empirical AUC estimates and jackknife covariance estimates is displayed in Table 8. The AUCs for soft-copy and screen-film images, averaged across the four readers, are 0.804 and

0.829, respectively. The test for the null hypothesis of no AUC difference between soft-copy and screen-film is not significant (F_{OR} = 0.31, df_2 = 164.4, p = 0.58); the 95% confidence interval for the difference of the population AUCs (screen-film minus soft-copy) is (−0.064, 0.114). Part (h) gives 95% confidence intervals for the single-test AUCs based only on data for the specific test.

5.4. Example 4: Case-nested-within-reader split-plot study design

In this study design each reader interprets a different set of cases using all of the diagnostic tests. The study design is balanced with each reader reading the same number of cases under each test. This can be viewed as a split-plot design with cases being the whole plots, reader the whole-plot factor, and test the split-plot factor. Obuchowski [12] refers to this as a *hybrid* design. The advantage of this design is that for equivalent power each reader must interpret fewer cases than for the factorial design, but the disadvantage is that the total number of cases is higher [13]. Thus this design is appropriate when a large number of verified cases are available and reading time per reader is limited or relatively expensive. This design is schematically illustrated in Table 7c.

Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S4. The derivation begins with a conventional split-plot ANOVA model corresponding to the study design that treats reader and case as random factors and includes all possible interactions. Averaging across cases produces the corresponding mm-ANOVA model, which is the same as the factorial model except with Cov_2 and Cov_3 constrained to zero; i.e., the model is defined by (1) and constraints: $Cov_1 \geq 0$, $Cov_2 = Cov_3 = 0$. Because this model is the same as the factorial model with Cov_2 and Cov_3 constrained to zero, hypotheses-test, confidence-interval, and sample-size formulas can be derived from those for the factorial model by setting $Cov_2 = Cov_3 = 0$.

Thus the test statistic for the null hypothesis of equal test accuracies is

$$F_{\mathrm{OR}}=\frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R)}$$

Under H_0, $F_{OR} \sim^{\cdot} F_{t-1,df_2}$ where

$$\mathrm{df}_2=(t-1)(r-1) \quad (29)$$

More generally, $F_{OR} \sim^{\cdot} F_{t-1,df_2;\lambda}$, where $\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma_{TR}^2+\sigma_{\varepsilon}^2-\mathrm{Cov}_1}$ and df_2 is given by (29).

Letting θ_i denote $E(\hat{\theta_{i\bullet}})$, an approximate $(1-\alpha)$ 100% confidence interval for contrast $\sum_{i=1}^{t} l_i\theta_i$ is given by $\sum_{i=1}^{t} l_i\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where df_2 is given by (29) and $\hat{V}=\frac{1}{r}\left(\sum_{i=1}^{t} l_i^2\right)\mathrm{MS}(T*R)$. An approximate $(1-\alpha)$ 100% confidence interval for θ_i is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where $\hat{V}=\frac{1}{tr}\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)\right]$ and

$\mathrm{df}_2 = \frac{[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)]^2}{[\mathrm{MS}(R)]^2/(r-1)+[(t-1)\mathrm{MS}[T*R]]^2/[(t-1)(r-1)]}$. Alternatively, an approximate (1 − α) 100% confidence interval for θ_i, using a standard error computed only from data for test *i*, is given by $\hat{\theta}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where $\hat{V} = \frac{1}{r}\mathrm{MS}(R)^{(i)}$ and $\mathrm{df}_2 = r - 1$.

5.5. Example 5: Reader-and-case-crossed-and-nested-within-group split-plot study design

In this study design there are several groups (or blocks) of readers and cases such that (1) each reader and each case belongs to only one group and (2) within each group all readers read all cases under each test. I assume a balanced design where each group has the same number of readers and cases. Obuchowski [13] discusses this design and refers to it as a *mixed* design; I will refer to it as a *mixed split-plot* design. The motivation for this study design is to reduce the number of reader interpretations for each reader, compared to the factorial study, without requiring as many cases to be verified as the hybrid design. This design is schematically illustrated in Table 7d. Although not explicitly stated, Obuchowski [13] assumes that there is no group effect for this design; e.g., cases and readers are randomly assigned to the groups (personal communication, Nancy Obuchowski, 2012). In contrast, I allow for a group effect; e.g., readers are assigned to groups according to experience level. Obuchowski et al [14] provide a real-data example that shows how this design can be particularly useful for studying multiple imaging tests.

Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S5. The derivation begins with a conventional split-plot ANOVA model corresponding to the study design (reader and case crossed and nested within group) that treats reader and case as random factors and group and test as fixed factors. All possible interactions are included. Averaging across cases produces the corresponding mm-ANOVA model: a three-way ANOVA model with group, test, and reader as factors.

Let $\hat{\theta}_{hij}$ denote the reader-performance estimate for reader *j* under test *i*, with both belonging to group *h*. The mm-ANOVA model is given by $\hat{\theta}_{hij} = \mu + \gamma_h + \tau_i + (\gamma\tau)_{hi} + R_{(h)i} + (\tau R)_{(h)ij} + \varepsilon_{hij}$, $h = 1, \ldots, g$, $i = 1, \ldots, t$, $j = 1, \ldots, r$, where *g* is the number of groups, *t* is the number of tests, *r* is the number of readers, τ_i denotes the fixed effect of test *i*, γ_h denotes the fixed effect of group *h*, and $(\gamma\tau)_{hi}$ denotes the fixed group-by-test interaction with $\sum_{i=1}^{t}\tau_i = \sum_{h=1}^{g}\gamma_h = \sum_{h=1}^{g}(\gamma\tau)_{hi} = \sum_{i=1}^{t}(\gamma\tau)_{hi} = 0$. The $R_{(h)j}$ and $(\tau R)_{(h)ij}$ are random reader and test-by-reader effects, nested within group; they are mutually independent and normally distributed with zero means and respective variances $\sigma^2_{R(G)}$ and $\sigma^2_{\tau R(G)}$. The ε_{hij} are normally distributed with zero mean and variance σ^2_{ε}. The ε_{hij} are independent of the $R_{(h)j}$ and $(\tau R)_{(h)ij}$. In summary, the mm-ANOVA model contains fixed effects for group, test, and their interaction, and random effects for reader nested within group and the test-by-reader interaction nested within group.

Cov_1, Cov_2, and Cov_3 are defined and constrained similar to corresponding covariances for the typical test×reader×case factorial design, but with this difference: here they are not defined between errors corresponding to different groups because the covariance of those

errors is zero. Specifically, $\mathrm{Cov}_1 \equiv \mathrm{Cov}\ (\varepsilon_{hij}, \varepsilon_{hi'j})$, $\mathrm{Cov}_2 \equiv \mathrm{Cov}\ (\varepsilon_{hij}, \varepsilon_{hij'})$, and $\mathrm{Cov}_3 \equiv \mathrm{Cov}\ (\varepsilon_{hij}, \varepsilon_{hij})$ where $i \neq i'$, $j \neq j'$ and $\mathrm{Cov}_1 \geq \mathrm{Cov}_3$, $\mathrm{Cov}_2 \geq \mathrm{Cov}_3$, and $\mathrm{Cov}_3 \geq 0$.

The null hypothesis of equal test accuracies is $H_0 : \theta_1 = \ldots = \theta_t$, where $\theta_i = E\ (\hat{\theta}_{\bullet i \bullet})$. The corresponding test statistic is

$$F_{\mathrm{OR}} = \frac{\mathrm{MS}(T)}{\mathrm{MS}[T*R(G)] + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right), 0\right]}$$

Under H_0, $F_{\mathrm{OR}} \sim^{\cdot} F_{t-1,\mathrm{df}_2}$ where

$$\mathrm{df}_2 = \frac{\left\{\mathrm{MS}[T*R(G)] + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right), 0\right]\right\}^2}{\{\mathrm{MS}[T*R(G)]\}^2/[g(t-1)(r-1)]} \quad (30)$$

and MS[T * R(G)] denotes the mean square for test-by-reader interaction nested within group. More generally, F_{OR} has an approximate $F_{t-1,\mathrm{df}_2;\lambda}$ distribution, where

$$\lambda = \frac{gr\sum_{i=1}^{t}\tau_i^2}{\sigma^2_{TR(G)} + \sigma^2_{\varepsilon} - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)} \text{ and}$$

$$\mathrm{df}_2 = \frac{\left[\sigma^2_{TR(G)} + \sigma^2_{\varepsilon} - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)\right]^2}{\left[\sigma^2_{TR(G)} + \sigma^2_{\varepsilon} - \mathrm{Cov}_1 - \mathrm{Cov}_2 + \mathrm{Cov}_3\right]^2/[g(t-1)(r-1)]}.$$

An approximate $(1 - \alpha)$ 100% confidence interval for contrast $\sum_{i=1}^{t} l_i\theta_i$ is given by $\sum_{i=1}^{t} l_i\hat{\theta}_{\bullet i \bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where df_2 is given by (30) and $\hat{V} = \frac{1}{r}\sum_{i=1}^{t} l_i^2\left\{\mathrm{MS}[R(T)] + r\max\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3, 0\right)\right\}$. An approximate $(1 - \alpha)$ 100% confidence interval for θ_i is given by $\hat{\theta}_{\bullet i \bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where

$$\hat{V} = \frac{1}{gtr}\left[\mathrm{MS}[R(G)] + (t-1)\mathrm{MS}[T*R(G)] + tr\max\left(\widehat{\mathrm{Cov}}_2, 0\right)\right] \text{ and}$$

$$\mathrm{df}_2 = \frac{\{\mathrm{MS}[R(G)] + (t-1)\mathrm{MS}[T*R(G)] + tr\max(\widehat{\mathrm{Cov}}_2, 0)\}^2}{\{\mathrm{MS}[R(G)]\}^2/[g(r-1)] + \{(t-1)\mathrm{MS}[T*R(G)]\}^2/[g(t-1)(r-1)]}.$$

5.6. Example 6: Replicated factorial study design

This study design is the same as the factorial study design except that each reader reads each case n times. Typically sessions corresponding to different readings are separated by a suitable period of time to reduce the probability that the reader will recognize cases from the earlier session. This study design has two advantages over the factorial design with one replication: it allows for estimation of within-reader reliability between two readings of the same cases, and it provides more power for the same number of cases and readers. This last aspect can be important if the number of available cases and readers is limited. In the example later in this section, I show how to estimate the gain in power based on pilot data.

Derivation of results using the mm-ANOVA algorithm is presented in Supplementary Table S6. The derivation begins with a conventional three-way replicated factorial ANOVA model with reader and case as random factors and test as a fixed factor. There are *n* replications. All possible interactions are included between reader, case and test. Averaging across cases for each replication produces the corresponding mm-ANOVA model: a two-way replicated factorial ANOVA model with test and reader as factors.

Let $\hat{\theta}_{ijm}$ denote the reader-performance estimate for reader *j* under test *i* based on the *m*th reading of the data. The mm-ANOVA model is given by $\hat{\theta}_{ijm} = \mu + \tau_i + R_j + (\tau R)_{ij} + \varepsilon_{ijm}$ $i = 1, \ldots, t, j = 1, \ldots, r, m = 1, \ldots, n$ where *t* is the number of tests, *r* is the number of readers, *n* is the number of replications, 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, ε_{ijm} is the error term, and $\sum_{i=1}^{t} \tau_i = 0$. The R_j and $(\tau R)_{ij}$ are assumed to be mutually independent and normally distributed with zero means and respective variances σ_R^2 and σ_{TR}^2. The ε_{ij} are assumed to be normally distributed with zero mean and variance σ_ε^2 and are assumed independent of the R_j and $(\tau R)_{ij}$. The errors are equi-covariant with four possible covariances given by

$$\mathrm{Cov}(\varepsilon_{ijm}, \varepsilon_{i'j'm'}) = \begin{cases} \mathrm{Cov}_0 & i=i, j=j, m \neq m' (\text{same test and reader, different replication}) \\ \mathrm{Cov}_1 & i \neq i', j=j' (\text{different test, same reader}) \\ \mathrm{Cov}_2 & i=i', j \neq j' (\text{same test, different reader}) \\ \mathrm{Cov}_3 & i \neq i', j \neq j' (\text{different test, different reader}) \end{cases}$$

and subject to the following constraints:

$$\mathrm{Cov}_0 \geq \mathrm{Cov}_1 \geq \mathrm{Cov}_3; \mathrm{Cov}_0 \geq \mathrm{Cov}_2 \geq \mathrm{Cov}_3; \mathrm{Cov}_3 \geq 0$$

Let $\theta_i \equiv E(\hat{\theta}_{i\bullet\bullet})$ denote the expected reader performance measure for test *i*. The test statistic for the null hypothesis of equal test accuracies ($H_0 : \theta_1 = \ldots = \theta_t$) is

$$F_{\mathrm{OR}} = \frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R) + nr \max\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3, 0\right)}$$

where $\mathrm{MS}(T) = \frac{nr}{t-1} \sum_{i=1}^{t} \left(\hat{\theta}_{i\bullet\bullet} - \hat{\theta}_{\bullet\bullet\bullet}\right)^2$ and $\mathrm{MS}(T*R) = \frac{n}{(t-1)(r-1)} \sum_{i=1}^{t} \sum_{j=1}^{r} \left(\hat{\theta}_{ij\bullet} - \hat{\theta}_{i\bullet\bullet} - \hat{\theta}_{\bullet j\bullet} + \hat{\theta}_{\bullet\bullet\bullet}\right)^2$. Under H_0, $F_{\mathrm{OR}} \sim \cdot F_{t-1,\mathrm{df}_2}$ where

$$\mathrm{df}_2 = \frac{\left[\mathrm{MS}(T*R) + nr \max\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3, 0\right)\right]^2}{\{\mathrm{MS}(T*R)\}^2 / [(t-1)(r-1)]} \quad (31)$$

More generally, $F_{\mathrm{OR}} \sim \cdot F_{t-1,\mathrm{df}_2;\lambda}$, where

$$\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma_{TR}^2+\sigma_{\varepsilon}^2/n-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)+[(n-1)/(n)]\mathrm{Cov}_0} \quad (32)$$

and

$$\mathrm{df}_2=\frac{\left[\sigma_{TR}^2+\sigma_{\varepsilon}^2/n-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)+[(n-1)/n]\mathrm{Cov}_0\right]^2}{[\sigma_{TR}^2+\sigma_{\varepsilon}^2/n-\mathrm{Cov}_1-(\mathrm{Cov}_2-\mathrm{Cov}_3)+[(n-1)/n]\mathrm{Cov}_0]^2/[(t-1)(r-1)]} \quad (33)$$

An approximate (1 − α) 100% confidence interval for contrast $\sum_{i=1}^{t} l_i\theta_i$ is given by $\sum_{i=1}^{t} l_i\hat{\theta}_{i\bullet\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$ where

$\hat{V}=\frac{1}{nr}\left(\sum_{i=1}^{t} l_i^2\right)\left\{\mathrm{MS}(T*R)+nr\max\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3,0\right)\right\}$ and df_2 is given by (31). An approximate (1 − α) 100% confidence interval for θ_i is given by $\hat{\theta}_{i\bullet\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where

$\hat{V}=\frac{1}{ntr}\left\{\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+\max\left(ntr\widehat{\mathrm{Cov}}_2,0\right)\right\}$ and

$\mathrm{df}_2=\frac{\left[\mathrm{MS}(R)+(t-1)\mathrm{MS}(T*R)+ntr\max\left(\widehat{\mathrm{Cov}}_2,0\right)\right]^2}{[\mathrm{MS}(R)]^2/(r-1)+[(t-1)\mathrm{MS}(T*R)]^2/[(t-1)(r-1)]}$.

Consider $\mathrm{Cov}_2 \equiv \mathrm{cov}(\hat{\theta}_{ijm}, \theta_{ij'm'})$ where $j \neq j'$ and either $m = m'$ or $m \neq m'$. It follows that Cov_2 can be computed from one set of replications ($m = m'$) or from different sets of replications ($m \neq m'$). For example, for test i and readers j and j', with $n = 2$ we have $\mathrm{Cov}_2 = \mathrm{cov}(\hat{\theta}_{ij1}, \theta_{ij'1}) = \mathrm{cov}(\hat{\theta}_{ij1}, \theta_{ij'2}) = \mathrm{cov}(\hat{\theta}_{ij2}, \theta_{ij'1}) = \mathrm{cov}(\theta_{ij2}, \theta_{ij'2})$. Thus an obvious estimate for Cov_2 that utilizes all of the data is given by

$$\widehat{\mathrm{Cov}}_2=\frac{2}{n^2tr(r-1)}\sum_{i=1}^{t}\sum_{j<j'1\le m\le n,1\le m'\le n}\sum\widehat{\mathrm{cov}}\left(\hat{\theta}_{ijm},\theta_{ij'm'}\right)$$

where $\widehat{\mathrm{cov}}\left(\hat{\theta}_{ijm},\theta_{ij'm'}\right)$ is a fixed-reader covariance estimate, as discussed in Section 2.2. Similarly, estimates for Cov_1 and Cov_3 can be estimated by averaging fixed-reader covariance estimates, computed for each of the n^2 possible (m, m') pairs of replications, across corresponding test-reader combinations. Obvious estimates for Cov_0 and σ_{ε}^2 are

$\widehat{\mathrm{Cov}}_0=\frac{2}{n(n-1)tr}\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{m<m'}\widehat{\mathrm{cov}}\left(\hat{\theta}_{ijm},\theta_{ijm'}\right)$ and

$\hat{\sigma}_{\varepsilon}^2=\frac{1}{ntr}\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{m=1}^{n}\widehat{\mathrm{var}}\left(\hat{\theta}_{ijm}\right)$, where $\widehat{\mathrm{var}}\left(\hat{\theta}_{ijm}\right)=\widehat{\mathrm{cov}}\left(\hat{\theta}_{ijm},\hat{\theta}_{ijm}\right)$.

5.6.1. Real-data example—In Section 2.3 I compared AUCs for hard- and soft-copy computed radiography chest images. Both types of images were obtained for each patient and were read by each of the readers. Thus this was a factorial study design, which could be analyzed by the standard OR procedure. Although there was not a significant difference

between the two types of images, the resulting confidence interval showed that an AUC difference as large as 0.086 was commensurate with the data. In such a situation the researcher might want to plan a similar experiment that is sized to have more power.

Increased power can be obtained by increasing the number of readers, the number of cases, or the number of replications. I now compute the number of cases needed to obtain .80 power to detect an AUC difference of .04 with alpha = .05. Because $F_{OR} \sim^{\cdot} F_{t-1,df_2;\lambda}$, power is approximated by Pr ($F_{1,df_2,\lambda} > F_{.95;1,df_2}$) where λ and df_2 are defined by (32) and (33) and $F_{.95;1,df_2}$ is the 95th percentile of a central F distribution with degrees of freedom 1 and df_2.

For the power computations I use the following estimates, obtained from Section 2.3: $\hat{\sigma}^2_{\varepsilon}$=.0022034331, $\widehat{\mathrm{Cov}}_1$=.0011163046, $\widehat{\mathrm{Cov}}_2$=.0.0008438255, $\widehat{\mathrm{Cov}}_3$=.0008871752, and $\hat{\sigma}^2_{TR}$=0. An estimate of Cov_0 is not available from the data because there are no replicated readings; however, the similarity of the two tests (hard- and soft-copy) suggests that the within-reader correlation between replications for the same test and reader, $\rho_0 = \mathrm{Cov}_0/\sigma^2_{\varepsilon}$, should be only slightly higher than the within-reader correlation based on one replication between two tests, given by $\hat{p}_1 = \widehat{\mathrm{Cov}}_1/\hat{\sigma}^2_{\varepsilon} = 0.507$ from Table 2. Thus I set ρ_0 = 0.60 for the power computations; it follows that $\widehat{\mathrm{Cov}}_0 = .6\hat{\sigma}^2_{\varepsilon} = 0.00132206$. Following Hillis et al [15] I assume that the covariances are inversely proportion to the number of cases c, and hence multiply $\hat{\sigma}^2_{\varepsilon}$, $\widehat{\mathrm{Cov}}_1$, $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ by the factor $\frac{95}{c}$ (recall that 95 is the number of cases for the example); the resulting values are used in place of σ^2_{ε}, Cov_1, Cov_2, and Cov_3 in (32) and (33) when computing power for c cases.

The numbers of cases need to achieve 0.80 power for combinations of 4–8 readers and 1–2 replications are presented in Table 9. For example, achieving 0.80 power with 8 readers and one replication requires 173 cases versus 103 cases with two replications. Thus if cases are expensive to obtain or validate and it is difficult to obtain more than 8 readers, then using two replications appears to be an attractive option.

6. Discussion

The mm-ANOVA approach allows for analysis of ROC and other reader-performance outcomes that result from any balanced study design that has reader and case as random factors and any number of fixed factors. In addition, by providing the non-null distribution of the test statistic it allows for sample size estimation for such studies and efficiency comparisons between different types of studies. Although steps were fully justified only for the test×reader×case factorial study design, justification can be similarly established for other designs. Until now researchers have been limited to using the test×reader×case study design with the OR method because analysis methods were not developed for other designs. This work allows researchers to choose designs that are most appropriate for their study. A SAS macro for fitting some of these designs using the mm-ANOVA approach is available on request from the author.

As noted in Section 2.4, Obuchowski and Rockette [1] derived their *F* statistic by modifying the *F* statistic described by Pavur and Nath [10]. Although Pavur and Nath [10] give results only for two-factor models, their approach, which is based on results given by Pavur and Lewis [19], could conceivably be applied to other correlated-error ANOVA models; as such it would provide an alternative to the approach described in this paper. However, the results of Pavur and Lewis do not extend beyond specifying the correct form for the *F* test when correlations are known; in particular, they do not indicate how to implement their approach when the correlations must be estimated, do not discuss derivation of confidence interval formulas for contrasts, give little motivation for the correlated error models, and do not discuss power computations.

Explicit formulas can be derived [20, 21, 22] for the variances of reader-performance outcomes that are U-statistics [23], such as reader empirical-AUC averages and their differences. Replacing parameters in these formulas by sample estimates yields variance estimates with excellent statistical properties. However, this approach is limited to U-statistic estimators, such as the empirical AUC and presently incorporates an adaptation of the OR degrees of freedom formula. Advantages include explicit variance formulas and applicability to a wide variety of multireader study designs, including unbalanced designs.

Another alternative approach for analyzing multireader data is the *marginal model* approach proposed by Song and Zhou [24] for empirical AUC estimates. An advantage of their approach is that case-specific covariates can be included; disadvantages include being limited to empirical AUC outcomes, based on large-sample inferences, and thus far developed only for the factorial model.

Limitations of the mm-ANOVA approach include the following: (1) It is presently limited to balanced study designs; i.e., the number of levels for each factor does not depend on any other factor. However, because case is treated as one factor it is possible to have different numbers of normal and abnormal cases. I am currently investigating models that are not balanced with regard to case. (2) It assumes that the number of cases is large enough so that covariance estimates can be treated like known values for computing the denominator degrees of freedom. (3) It assumes that the fixed-reader measurement errors, the ε_{ij}, are normally distributed. This is a reasonable assumption when the number of cases is moderate because most typical reader-performance outcomes, such as AUC, have asymptotic normal distributions for a fixed reader. (4) It assumes that the latent reader-performance outcomes (i.e., $R_j + (\tau R)_{ij}$) have a normal distribution. If these normal distribution assumptions do not appear to be reasonable, one possible remedy is to transform the outcome, e.g., using a logarithmic or logit transformation for AUC. (5) It assumes the errors have an equi-covariance structure. I am currently investigating the robustness of the mm-ANOVA approach to this assumption.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This research was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), grants R01EB000863 and R01EB013667. I thank Dr. Harold Kundel for sharing his data set.

Appendix

A. DERIVATION OF THE NULL-DISTRIBUTION RESULT USED IN STEP 2J

To derive the null-distribution result given in step 2j, I approximate the distribution of F_{OR} (22) by deriving an approximate distribution for F^{*}_{OR} (21), where Cov_2 and Cov_3 are known. Each $\tilde{\mathrm{MS}}_i$ is equal to its corresponding conventional three-way ANOVA model mean square, denoted by MS_i, multiplied by $\frac{1}{c}$, with $\mathrm{MS}_i \sim E(\mathrm{MS}_i)\chi^2_{\mathrm{df}(\mathrm{MS}_i)}/\mathrm{df}(\mathrm{MS}_i)$ under $\mathrm{H}_0 : \theta_1 = \ldots = \theta_t$. It follows that the $\tilde{\mathrm{MS}}_i$ are mutually independent, each $\tilde{\mathrm{MS}}_i$ has the same degrees of freedom as its corresponding MS_i and $\tilde{\mathrm{MS}}_i \sim E\left(\tilde{\mathrm{MS}}_i\right)\chi^2_{\mathrm{df}(\tilde{\mathrm{MS}}_i)}/\mathrm{df}\left(\tilde{\mathrm{MS}}_i\right)$.

In general, a chi-squared-distribution approximation [25, 26] for a random variable X is given by

$$E(X)\chi^2_{\mathrm{df}}/\mathrm{df}$$

where

$$\mathrm{df} = \frac{2[E(X)]^2}{\mathrm{var}(X)}$$

It follows that a chi-square approximation for

$$X = b\left(\sum_{i=1}^{I} a_i \tilde{\mathrm{MS}}_i + d\right)$$

where the a_i, b and d are constants, is given by

$$b\left(\sum_{i=1}^{I} a_i E(\tilde{\mathrm{MS}}_i) + d\right)\chi^2_{\mathrm{df}}/\mathrm{df} \quad \text{(A1)}$$

where

$$\mathrm{df} = \frac{\left[\sum_{i=1}^{I} a_i E\left(\tilde{\mathrm{MS}}_i\right) + d\right]^2}{\sum_{i=1}^{I} \frac{[a_i E(\tilde{\mathrm{MS}}_i)]^2}{\mathrm{df}(\tilde{\mathrm{MS}}_i)}} \quad \text{(A2)}$$

Replacing $E\left(\tilde{\mathrm{MS}}_i\right)$ by $\tilde{\mathrm{MS}}_i$ and d by an estimate $\hat{d}$ in (A2) results in the approximation for df given by df_2 (24).

It follows using (A1) with $i = 1$, $a_1 = 1$, $\tilde{\mathrm{MS}}_1 = \tilde{\mathrm{MS}}(T*R)$, $d = r\,(\mathrm{Cov}_2 - \mathrm{Cov}_3)$ and (A2) estimated by (24) that a chi-squared approximation for $\tilde{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)$, the denominator of F^* (21), is given by

$$\tilde{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3) \dot{\sim} \{E\left[\tilde{\mathrm{MS}}(T*R)\right] + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)\}\chi^2_{\mathrm{df}_2}/\mathrm{df}_2 \quad \text{(A3)}$$

where df_2 is given by (25) and "$\dot{\sim}$" stands for "is approximately distributed as." See Reference [6] for a more detailed derivation and justification of df_2 (referred to as ddf_H in the reference.)

Because F^*_{OR} (21) is an ANOVA statistic, $E\left[\tilde{\mathrm{MS}}(T)\right] = E\left[\tilde{\mathrm{MS}}(T*R)\right] + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)$ under H_0. Combining this result with the chi-squared approximation (A3) for $\tilde{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)$ and the independence of $\tilde{\mathrm{MS}}(T)$ and $\tilde{\mathrm{MS}}(T*R)$, it follows under H_0 that

$$F^*_{\mathrm{OR}} = \frac{\tilde{\mathrm{MS}}(T)}{\tilde{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)} = \frac{\frac{\tilde{\mathrm{MS}}(T)}{E[\tilde{\mathrm{MS}}(T)]}}{\frac{\tilde{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)}{E[\tilde{\mathrm{MS}}(T*R)] + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)}} = \frac{U/(t-1)}{W/\mathrm{df}_2}$$

where $U \sim \chi^2_{t-1}$, W is approximately $\chi^2_{\mathrm{df}_2}$, and U and W are independent. Thus F^*_{OR} has an approximate $F_{(t-1),\mathrm{df}_2}$ null distribution, with df_2 given by (25). Because F_{OR} (22) approximates F^*_{OR} (21), it is reasonable to approximate the null distribution of F_{OR} by $F_{(t-1),\mathrm{df}_2}$, which is the null distribution derived by Hillis [6] for F_{OR}, discussed in Section 2.2.

B. MM-ANOVA APPROACH STEP 3: DERIVE CONFIDENCE INTERVALS FOR A LINEAR FUNCTION *g*(θ) OF TEST ACCURACIES

In this section I show how to compute a confidence interval for a linear function of test accuracy parameters. Specifically, for the balanced test×reader×case factorial study design with $\theta_i \equiv E(\hat{\theta}_{i\bullet})$ denoting the expected reader-performance outcome for test i across readers, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_t)'$, and $\boldsymbol{l} = (l_1, \ldots, l_t)'$ denoting a t-dimensional contrast vector (i.e., $\sum_{i=1}^{t} l_i = 0$), I illustrate how to derive a confidence interval for $g(\boldsymbol{\theta}) \equiv \boldsymbol{l}'\boldsymbol{\theta}$. More generally this step can be used to determine a confidence interval for $g(\boldsymbol{\theta})$, where $g(\cdot)$ is any linear function and $\boldsymbol{\theta}$ any vector of test accuracy parameters; this general result is given in step 3k.

B.1. Step 3a: Write the test accuracy parameter vector θ in terms of the mm-ANOVA model

In terms of the mm-ANOVA model parameterization, treating $\tilde{Y}_{ij}$ as $\hat{\theta}_{ij}$, we have $\theta_i = E(\tilde{Y}_{i\bullet}) = \mu + \tau_i$.

B.2. Step 3b: Write θ in terms of the conventional ANOVA model

Since $\theta_i = E(\tilde{Y}_{i\bullet}) = E(Y_{i\bullet\bullet}) = \mu + \tau_i$, then in terms of the conventional ANOVA model we also have $\theta_i = \mu + \tau_i$.

B.3. Step 3c: Determine the conventional ANOVA estimate for θ, denoted by $\hat{\theta}$

The conventional unbiased ANOVA estimate for θ is given by $\hat{\theta} = (\hat{\theta_1}, \ldots, \hat{\theta_t})'$ with $\hat{\theta_i} = Y_{i\bullet\bullet}$.

B.4. Step 3d: Determine the variance V of g ($\hat{\theta}$) in terms of conventional ANOVA parameters

From (7) it follows that

$$g\left(\hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{t} l_i \hat{\boldsymbol{\theta}}_i = \sum_{i=1}^{t} l_i Y_{i\bullet\bullet} = \sum_{i=1}^{t} l_i \tau_i + \sum_{i=1}^{t} l_i [(\tau R)_{i\bullet} + (\tau C)_{i\bullet} + (\tau RC)_{i\bullet\bullet} + \varepsilon_{i\bullet\bullet}]$$

Thus

$$V \equiv \mathrm{Var}\left(g\left(\hat{\boldsymbol{\theta}}\right)\right) = \sum_{i=1}^{t} l_i^2 \left[\frac{\sigma_{TR}^2}{r} + \frac{\sigma_{TC}^2}{c} + \frac{\sigma^2}{rc}\right] = \frac{1}{rc} \sum_{i=1}^{t} l_i^2 [c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2]$$

Because $\hat{\theta}$ has a multivariate normal distribution, it follows that

$$g\left(\hat{\boldsymbol{\theta}}\right) \sim N(\boldsymbol{l}'\theta, V)$$

where

$$V = \frac{1}{rc} \sum_{i=1}^{t} l_i^2 (c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2)$$

B.5. Step 3e: Write V from step 3d in the form V = bE (∑$a_i$$MS_i$) for constants b and a_i

Expected values of the conventional ANOVA mean squares are given in Table 4. It follows that

$$V = \frac{1}{rc} \sum_{i=1}^{t} l_i^2 E[\mathrm{MS}(T*R) + \mathrm{MS}(T*C) - \mathrm{MS}(T*R*C)]$$

B.6. Step 3f: Write V from step 3e in the form $V = \tilde{b}E(\sum \tilde{a}_i \tilde{\mathrm{MS}}_i + U)$ where $\tilde{b}$ and $\tilde{a}_i$ are constants and U is a linear function of conventional ANOVA mean squares that involve case

We have

$$V=\frac{1}{r}\sum_{i=1}^{t} l_i^2 E\left[\tilde{\mathrm{MS}}(T*R)+U\right]$$

where

$$U=\frac{1}{c}[\mathrm{MS}(T*C)-\mathrm{MS}(T*R*C)]$$

B.7. Step 3g: Express E (U) in terms of conventional ANOVA model variance components and then in terms of mm-ANOVA model error covariance parameters, using the relationships from step 1c; then rewrite V using this expression for E (U)

We did the first part of this step in step 2g where we showed

$$E(U)=r(\mathrm{Cov}_2-\mathrm{Cov}_3)$$

Using this expression we have

$$V=\frac{1}{r}\sum_{i=1}^{t} l_i^2\{E\left[\tilde{\mathrm{MS}}(T*R)\right]+r(\mathrm{Cov}_2-\mathrm{Cov}_3)\}$$

B.8. Step 3h: Derive the variance estimate $\hat{V}$ from V by replacing expected mean squares by mean squares and replacing covariances by estimates that take into account the constraints from step 1d

We have

$$\hat{V}=\frac{1}{r}\sum_{i=1}^{t} l_i^2\{\tilde{\mathrm{MS}}(T*R)+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]\}$$

B.9. Step 3i: Derive the degrees of freedom df_2 for $\hat{V}$ using the general formula for df_2 (24) given in step 2j

It follows that the degrees of freedom is given by (25), which is the same as ddf_H (6).

B.10. Step 3j: Write $\hat{\theta}$ from step 3c in terms of the mm-ANOVA model

Since $\hat{\theta_i} = Y_{i\bullet\bullet} = \tilde{Y}_{i\bullet\bullet}$, then in terms of the mm-ANOVA model $\hat{\theta_i} = \tilde{Y}_{i\bullet\bullet}$.

B.11. Step 3k: General confidence-interval result: In terms of the mm-ANOVA model, an approximate (1 − α) 100% confidence interval for g (θ) is given by $g\left(\hat{\boldsymbol{\theta}}\right) \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$ where $\hat{V}$ is determined in step 3h, df_2 in step 3i and $\hat{\theta}$ in step 3j

This result yields the following (1 − α) 100% confidence interval for $l'\theta$:

$$\sum_{i=1}^{t} l_i \tilde{Y}_{i\bullet} \pm t_{\alpha/2;\mathrm{ddf_H}} \sqrt{\frac{1}{r}\left(\sum_{i=1}^{t} l_i^2\right)\left[\tilde{\mathrm{MS}}(T*R)+r\max(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3,0)\right]} \quad \text{(B1)}$$

where $\mathrm{ddf_H}$ is given by (25). Letting "F_{OR}-test denominator" denote the denominator of the F_{OR} statistic (22) for testing $\mathrm{H_0}: \theta_1 = \ldots = \theta_t$, we can write (B1) as

$$\sum_{i=1}^{t} l_i \tilde{Y}_{i\bullet} \pm t_{\alpha/2;\mathrm{ddf_H}} \sqrt{\frac{1}{r}\left(\sum_{i=1}^{t} l_i^2\right)\{F_{\mathrm{OR}}-\text{test denominator}\}}$$

B.12. Derivation of the general confidence-interval result given in step 3k

I now derive the step 3k result for the test×reader×case factorial study design with $g(\boldsymbol{\theta}) \equiv \boldsymbol{l}'\boldsymbol{\theta}$ and $\boldsymbol{l} = (l_1, \ldots, l_t)'$ denoting a t-dimensional contrast vector (i.e., $\sum_{i=1}^{t} l_i = 0$). We have shown in the previous steps that $g(\hat{\boldsymbol{\theta}}) \sim N[g(\boldsymbol{\theta}), V]$, where

$$V = \frac{1}{r}\sum_{i=1}^{t} l_i^2\{E\left[\tilde{\mathrm{MS}}(T*R)\right]+r(\mathrm{Cov}_2-\mathrm{Cov}_3)\}$$

Define V^* by replacing $E\left[\tilde{\mathrm{MS}}(T*R)\right]$ by $\tilde{\mathrm{MS}}(T*R)$:

$$V^* = \frac{1}{r}\sum_{i=1}^{t} l_i^2\{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)\}$$

Using the same argument as given in Appendix A and noting that $V = E(V^*)$, we can show that a chi-squared-distribution approximation for V^* is given by $V\chi^2_{\mathrm{df}_2}/\mathrm{df}_2$ with df_2 given by (25). Furthermore, independence of $g(\hat{\boldsymbol{\theta}})$ and $\tilde{\mathrm{MS}}(T*R)$ for the mm-ANOVA model, and hence independence of $g(\hat{\boldsymbol{\theta}})$ and V^*, follows from the independence of $g(\hat{\boldsymbol{\theta}})$ and $\mathrm{MS}(T*R)$ for the conventional ANOVA model (7). Thus for the mm-ANOVA model

$$t = \frac{g\left(\hat{\theta}\right)-g(\theta)}{\sqrt{\frac{1}{r}\sum_{i=1}^{t} l_i^2\{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)}} = \frac{g\left(\hat{\theta}\right)-g(\theta)}{\sqrt{V^*}} = \frac{\frac{g(\hat{\theta})-g(\theta)}{\sqrt{V}}}{\sqrt{\frac{(V^*)\mathrm{df}_2}{V}/\mathrm{df}_2}} = \frac{Z}{\sqrt{W/\mathrm{df}_2}}$$

where $Z \sim N(0, 1)$, W is approximately $\chi^2_{\mathrm{df}_2}$, and Z and W are independent. Thus

$$t = \frac{g\left(\hat{\theta}\right)-g(\theta)}{\sqrt{\frac{1}{r}\sum_{i=1}^{t} l_i^2\{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)}}$$

has an approximate t_{df_2} distribution with df_2 given by (25). In practice we replace r (Cov_2 − Cov_3) by $\max[r(\widehat{Cov}_2 - \widehat{Cov}_3), 0]$ and base tests and confidence intervals on

$$t=\frac{g\left(\hat{\theta}\right)-g(\theta)}{\sqrt{\frac{1}{r}\sum_{i=1}^{t}l_i^2\{\tilde{\mathrm{MS}}(T*R)+\max[r(\widehat{Cov}_2-\widehat{Cov}_3),0]\}}}=\frac{g\left(\hat{\theta}\right)-g(\theta)}{\sqrt{\hat{V}}} \quad \text{(B2)}$$

which we treat as having an approximate t_{df_2} distribution; the confidence interval result in step 3k follows.

The general result for with $g(\cdot)$ being any linear function can be similarly proved, with the main difference being the formula for V.

C. MM-ANOVA APPROACH – STEP 4: DERIVE THE NON-NULL DISTRIBUTION OF F_{OR}

Power and sample size estimation for the step 2a hypothesis requires specification of the distribution of the F_{OR} statistic, derived in step 2i, when the null hypothesis is not true. A noncentral F distribution approximation for the non-null distribution is specified by steps 4a–d below. These steps are justified in Section C.5.

C.1. Step 4a: Compute the noncentrality parameter in terms of the conventional ANOVA model

Express the noncentrality parameter in terms of the conventional ANOVA model using

$$\lambda=\frac{\mathrm{df}(\mathrm{MS_{num}})\mathrm{MS_{num}}|_{\boldsymbol{Y}=E(\boldsymbol{Y})}}{E(\mathrm{MS_{num}}|\mathrm{H_0})} \quad \text{(C1)}$$

where MS_{num} is the numerator mean square from the conventional ANOVA F statistic given in step 2d, $df(MS_{num})$ is its degrees of freedom, $E(MS_{num}|H_0)$ is its expected value under H_0, and $MS_{num}|_{\mathbf{Y}=E(\mathbf{Y})}$ is the mean square evaluated with outcomes replaced by their expected values.

For the balanced test×reader×case factorial design we have MS_{num} = MS (T) from step 2d. From Table 4 we have $E[\mathrm{MS}(T)]=\frac{rc}{(t-1)}\sum_{i=1}^{t}\tau_i^2+c\sigma_{TR}^2+r\sigma_{TC}^2+\sigma^2$. Thus $E[\mathrm{MS}(T)|\mathrm{H_0}]=c\sigma_{TR}^2+r\sigma_{TC}^2+\sigma^2$ under H_0:$\tau_1 = \ldots = \tau_t = 0$. Noting that $E(Y_{ijk}) = \mu + \tau_i$, we have $\mathrm{MS}(T)|_{\boldsymbol{Y}=E(\boldsymbol{Y})}=\frac{rc}{t-1}\sum_{i=1}^{t}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2|_{Y_{ijk}=\mu+\tau_i}=\frac{rc}{t-1}\sum_{i=1}^{t}\tau_i^2$. Noting that $df[MS(T)] = t - 1$, then from (C1) it follows that

$$\lambda=\frac{rc\sum_{i=1}^{t}\tau_i^2}{c\sigma_{TR}^2+r\sigma_{TC}^2+\sigma^2} \quad \text{(C2)}$$

C.2. Step 4b: Express λ in terms of mm-ANOVA parameters

Replace variance components in (C2) corresponding to random effects involving case by mm-ANOVA covariances. From the relationships determined in step 1c and presented in Table 3 we have

$$r\sigma_{TC}^2+\sigma^2=c[\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)]$$

(Recall that $\sigma_{\tilde{\varepsilon}}^2$ is the error variance for the mm-ANOVA model.) Thus in terms of mm-ANOVA parameters

$$\lambda=\frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma_{TR}^2+\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)} \quad \text{(C3)}$$

C.3. Step 4c: Determine the denominator degrees of freedom in terms of mm-ANOVA parameters

Write the denominator of F_{OR}^* from step 2h in the form $b\left(\sum_{i=1}^{I}a_i\tilde{\mathrm{MS}}_i+d\right)$. The denominator degrees of freedom is given by

$$\mathrm{df}_2=\frac{\left[\sum_{i=1}^{I}\alpha_i E\left(\tilde{\mathrm{MS}}_i\right)+d\right]^2}{\sum_{i=1}^{I}\left[\alpha_i E\left(\tilde{\mathrm{MS}}_i\right)\right]^2/\mathrm{df}\left(\tilde{\mathrm{MS}}_i\right)} \quad \text{(C4)}$$

which is the same as (A2). Note that (C4) contains the expected mean square values and the true value of *d*, in contrast to approximation (24) that replaces these values by sample estimates. The reason for this difference is that approximation (24) will be used for hypotheses testing and confidence intervals for a study data set; in contrast, (C4) will be used for sample-size and power estimation for a future study and will be based on parameter values that are either conjectured or estimated from pilot data.

Express the expected mean squares in (C4) in terms of mm-ANOVA model parameters by determining their expected values in terms of the conventional ANOVA parameters and then replacing variance components that involve case by mm-ANOVA covariances. For example, for the balanced test×reader×case factorial study design, the denominator of F_{OR}^* from step 2h is given by $\sum_{i=1}^{I}a_i\tilde{\mathrm{MS}}_i+d=\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)$. From (17) and Tables 3–4 it follows that

$$E\left(\tilde{\mathrm{MS}}(T*R)\right)=\frac{1}{c}E(\mathrm{MS}(T*R))=\frac{1}{c}(c\sigma_{TR}^2+\sigma^2)$$

with $\sigma^2=c(\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1-\mathrm{Cov}_2+\mathrm{Cov}_3)$. Thus

$$E\left(\tilde{\mathrm{MS}}(T*R)\right)=\sigma_{TR}^2+\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1-\mathrm{Cov}_2+\mathrm{Cov}_3$$

and hence, using (C4),

$$\mathrm{df}_2=\frac{\left[\sigma_{TR}^2+\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1+(r-1)(\mathrm{Cov}_2-\mathrm{Cov}_3)\right]^2}{\frac{\left[\sigma_{TR}^2+\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1-\mathrm{Cov}_2+\mathrm{Cov}_3\right]^2}{(t-1)(r-1)}} \quad \text{(C5)}$$

Hillis et al [15] illustrate how these formulas can be used in practice to estimate power and sample size using pilot-data or conjectured parameter estimates.

C.4. Step 4d: General non-null distribution result

An approximation for the non-null distribution of F_{OR} is given by

$$F_{\mathrm{df}_1,\mathrm{df}_2;\lambda}$$

where λ is given in step 4b, df_1 is the degrees of freedom for the numerator mean square from the conventional ANOVA *F* statistic given in step 2d and df_2 is given by (C4), expressed in terms of the mm-ANOVA parameters. Thus for the balanced test×reader×case factorial study design, λ is given by (C3),$\mathrm{df}_1 = t - 1$, and df_2 is given by (C5).

C.5. Justification of steps 4a–d

The non-null distribution result given in step 4d can derived for the test×reader×case study design along the same lines as the derivation of the null distribution result given in Section A. One difference is that $\tilde{\mathrm{MS}}(T)$, the numerator numerator mean square in F_{OR} (22) has a noncentral chi-square distribution when appropriately normalized under H_1. The distribution for MS(*T*) is given by

$$(t-1)\frac{\mathrm{MS}(T)}{E(\mathrm{MS}(T|\mathrm{H}_0))}\sim\chi_{t-1;\lambda}^2$$

where λ is given by (C3). Because $\tilde{\mathrm{MS}}(T)=\frac{1}{c}\mathrm{MS}(T)$, it follows that

$$(t-1)\frac{\tilde{\mathrm{MS}}(T)}{E\left(\tilde{\mathrm{MS}}(T|\mathrm{H}_0)\right)}\sim\chi_{t-1;\lambda}^2$$

Using the Section A approach but with this one difference, we can show that

$$F_{OR}^*=\frac{\tilde{\mathrm{MS}}(T)}{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)}=\frac{\frac{\tilde{\mathrm{MS}}(T)}{E[\tilde{\mathrm{MS}}(T)|\mathrm{H}_0]}}{\frac{\tilde{\mathrm{MS}}(T*R)+r(\mathrm{Cov}_2-\mathrm{Cov}_3)}{E[\tilde{\mathrm{MS}}(T*R)]+r(\mathrm{Cov}_2-\mathrm{Cov}_3)}}=\frac{U/(t-1)}{W/\mathrm{df}_2}$$

where $U \sim \chi^2_{t-1;\lambda}$ W is approximately $\chi^2_{df_2}$ with df_2 given by (C5), and U and W are independent. Thus F^*_{OR} has an approximate $F_{(t-1),df_2;\lambda}$ distribution. Because F_{OR} (22) approximates F^*_{OR} (21), it is reasonable to approximate the null distribution of F_{OR} by $F_{(t-1),df_2;\lambda}$.

References

1. 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.
2. 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]
3. 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]
4. 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]
5. 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]
6. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine. 2007; 26:596–619. [PubMed: 16538699]
7. 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]
8. 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]
9. Kundel HL, Gefter W, Aronchick J, Miller W, Hatabu H, Whitfill CH. Accuracy of bedside chest hard-copy screen-film versus hard-and soft-copy computed radiographs in a medical intensive care unit: receiver operating characteristic analysis. Radiology. 1997; 205:859–863. [PubMed: 9393548]
10. Pavur R, Nath R. Exact F tests in an ANOVA procedure for dependent observations. Multivariate Behavioral Research. 1984; 19:408–420.
11. Searle, SR. Linear Models. New York: Wiley; 1971. p. 55-59.
12. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. Academic Radiology. 1995; 2(Suppl 1):S22–S29. [PubMed: 9419702]
13. Obuchowski NA. Reducing the number of reader interpretations in MRMC studies. Academic Radiology. 2009; 16:209–217. [PubMed: 19124107]
14. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. Academic Radiology. 2012; 19:1508–1517. [PubMed: 23122570]
15. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods: An updated and unified approach. Academic Radiology. 2011; 18:129–142. doi: [PubMed: 21232681]
16. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. Academic Radiology. 2000; 7:516–525. [PubMed: 10902960]
17. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis, and validation. Medical Physics. 2004; 31:2313–2330. [PubMed: 15377098]

18. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. Free-response approach to the measurement and characterization of radiographic-observer performance. Journal of Applied Photographic Engineering. 1978; 4:166–171.
19. Pavur RJ, Lewis TO. Unbiased F-tests for factorial-experiments for correlated data. Communications in Statistics-Theory and Methods. 1983; 12:829–840.
20. Gallas BD. One-shot estimate of MRMC variance: AUC. Academic Radiology. 2006; 13:353–362. [PubMed: 16488848]
21. Gallas BD, Pennelo GA, Myers KJ. Multireader multicase variance analysis for binary data. JOSA A. 2007; 24:B70–B80. [PubMed: 18059916]
22. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators. Communications in Statistics-Theory and Methods. 2009; 38:2586–2603.
23. Hoeffding W. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics. 1948; 19:293–325.
24. Song X, Zhou XH. A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. Biostatistics. 2005; 6:303–312. [PubMed: 15772108]
25. Satterthwaite FE. Synthesis of variance. Psychometrika. 1941; 6:309–316.
26. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometric Bulletin. 1946; 2:110–114.

Table 1

Expected mean square and variance component formulas for the Obuchowski-Rockette model.

a. Expected mean squares

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

b. Variance components

Variance component	Equivalent function of expected mean squares and covariances
σ_R^2	$\frac{1}{t}E\{\mathrm{MS}(R)-\mathrm{MS}(T*R)\}-\mathrm{Cov}_1+\mathrm{Cov}_3$
σ_{TR}^2	$E[\mathrm{MS}(T*R)]-\sigma_\varepsilon^2+\mathrm{Cov}_1+(\mathrm{Cov}_2-\mathrm{Cov}_3)$

Table 2

Obuchowski-Rockette analysis of Kundel et al [9] data for soft- and hard-copy computed radiographs using trapezoid AUC estimation and jackknife covariance estimation for $t = 2$ tests, $r = 4$ readers, $c = 95$ cases (66 nondiseased, 29 diseased).

a. Trapezoid AUCs:

	Test	
	1 (Soft-copy)	**2 (Hard-copy)**
Reader (j)	$\hat{\theta}_{1j}$	$\hat{\theta}_{2j}$
1	0.815	0.854
2	0.767	0.812
3	0.831	0.900
4	0.803	0.798
	$\hat{\theta}_{1\cdot} = .804$	$\hat{\theta}_{2\cdot} = .841$

b. ANOVA table:

Source	df	Sum of squares	Mean square
T	1	0.00281054	0.00281054
R	4	0.00715054	0.00238351
T*R	4	0.00140392	0.00046797

c. Fixed-reader covariance and corresponding correlation estimates computed from jackknife covariance matrix:

$\hat{\sigma}^2_{\varepsilon} = .0022034331$, $\hat{\mathrm{Cov}}_1 = .0011163046$, $\hat{\mathrm{Cov}}_2 = .0008438255$, $\hat{\mathrm{Cov}}_3 = .0008871752$, $\hat{\rho}_1 = 0.507$, $\hat{\rho}_2 = 0.383$, $\hat{\rho}_3 = 0.403$

d. Variance component estimates using Table 1b formulas:

$$\hat{\sigma}^2_R = \frac{1}{t}\{\mathrm{MS}(R) - \mathrm{MS}(T*R)\} - \hat{\mathrm{Cov}}_1 + \hat{\mathrm{Cov}}_3 = 0.0007286397$$

$$\hat{\sigma}^2_{TR} = \mathrm{MS}(T*R) - \hat{\sigma}^2_{\varepsilon} + \hat{\mathrm{Cov}}_1 + \max\left(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3, 0\right) = -0.000662504 \text{(typically this would be changed to zero)}$$

e.

$$F_{\mathrm{OR}} = \frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R) + r\max(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3, 0)} = 6.00576$$

f. Denominator degrees of freedom:

$$\mathrm{ddf_H} = \frac{[\mathrm{MS}(T*R) + \max[r(\hat{\mathrm{C}}\mathrm{ov}_2 - \hat{\mathrm{C}}\mathrm{ov}_3), 0]]^2}{\frac{[\mathrm{MS}(T*R)]^2}{(t-1)(r-1)}} = 3$$

g. P -value for H_0: $\theta_1 = \theta_2$: $p = \Pr(F_{(t-1),\,\mathrm{ddf_H}} \geq F_{\mathrm{OR}}) = .092$

h. 95% CI for $\theta_2 - \theta_1 : \hat{\theta}_{2\cdot} - \hat{\theta}_{1\cdot} \pm t_{\mathrm{ddf_H}} \sqrt{\frac{2}{r}\left\{\mathrm{MS}(T*R) + r\max\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3, 0\right)\right\}} = (-0.0111940, .086168)$

i. Single-test 95% confidence intervals based on all of the data. Note:
$\mathrm{StdErr} = \frac{1}{tr}\left[\mathrm{MS}(R) + (t-1)\mathrm{MS}(T*R) + tr\max\left(\widehat{\mathrm{Cov}}_2, 0\right)\right]$.

i	$\hat{\theta}_i$	StdErr	df_2	95% CI
1 (Soft-copy)	0.804	.0346	46.9	0.734, 0.874
2 (Hard-copy)	0.841	.0346	46.9	0.772, 0.911

j. Single test 95% confidence intervals using only corresponding test data. Note:
$\mathrm{StdErr}^{(i)} = \sqrt{\frac{1}{r}\left[\mathrm{MS}(R)^{(i)} + r*\max\left(\widehat{\mathrm{Cov}}_2^{(i)}, 0\right)\right]}$.

i	$\hat{\theta}_i$	$\widehat{\mathrm{Cov}}_2^{(i)}$	$\mathrm{MS}(R)^{(i)}$	$\mathrm{StdErr}^{(i)}$	$\mathrm{df}_2^{(i)}$	95% CI
1 (Soft-copy)	0.804	0.000880	0.000735	0.0326	100.4	0.739, 0.867
2 (Hard-copy)	0.841	0.000808	0.002116	0.0366	19.2	0.765, 0.918

Table 3

Relationships between the 3-way ANOVA (7) and corresponding mm-ANOVA (9, 3) model parameters

3-way ANOVA parameter	Equivalent function of mm-ANOVA parameters
μ	= μ
τ_i	= τ_i
σ_R^2	$=\sigma_R^2$
σ_{TR}^2	$=\sigma_{TR}^2$
σ_C^2	= cCov$_3$
σ_{TC}^2	= c (Cov$_2$ − Cov$_3$)
σ_{RC}^2	= c (Cov$_1$ − Cov$_3$)
$\sigma^2 \equiv \sigma_{TRC}^2+\sigma_\varepsilon^2$	$=c\left(\sigma_{\tilde{\varepsilon}}^2-\mathrm{Cov}_1-\mathrm{Cov}_2+\mathrm{Cov}_3\right)$
mm-ANOVA parameter	**Equivalent function of 3-way ANOVA parameters**
μ	μ
τ_i	τ_i
σ_R^2	$=\sigma_R^2$
σ_{TR}^2	$=\sigma_{TR}^2$
$\sigma_{\tilde{\varepsilon}}^2$	$=\frac{1}{c}\left(\sigma_C^2+\sigma_{TC}^2+\sigma_{RC}^2+\sigma_\varepsilon^2\right)$
Cov$_1$	$=\frac{1}{c}\left(\sigma_C^2+\sigma_{RC}^2\right)$
Cov$_2$	$=\frac{1}{c}\left(\sigma_C^2+\sigma_{TC}^2\right)$
Cov$_3$	$=\frac{1}{c}\left(\sigma_C^2\right)$

These relationships assume covariance constraints (3) for the mm-ANOVA model and the same linear constraints for the τ_i (i.e., $\sum \tau_i = 0$) for both models.

Table 4

Expected mean squares for the conventional test-by-reader-by-case factorial ANOVA model (7).

Mean square	Expected mean square
MS (*T*)	$\frac{rc}{(t-1)}\sum_{i=1}^{t}\tau_i^2+c\sigma_{TR}^2+r\sigma_{TC}^2+\sigma^2$
MS (*R*)	$tc\sigma_R^2+c\sigma_{TR}^2+t\sigma_{RC}^2+\sigma^2$
MS (*C*)	$tr\sigma_C^2+r\sigma_{TC}^2+t\sigma_{RC}^2+\sigma^2$
MS (*T* * *R*)	$c\sigma_{TR}^2+\sigma^2$
MS (*T* * *C*)	$r\sigma_{TC}^2+\sigma^2$
MS (*R* * *C*)	$t\sigma_{RC}^2+\sigma^2$
MS (*T* * *R* * *C*)	$\sigma^2\equiv\sigma_{TRC}^2+\sigma_\varepsilon^2$

Table 5

Algorithm for deriving mm-ANOVA formulas

1 *Derive the mm-ANOVA model*

- **a.** Define the conventional ANOVA model that corresponds to the study design as if each reader-performance measure was the mean of case-level outcomes. (Note: Since reader-performance measures are measures of discrimination between diseased and nondiseased cases, disease status should not be included as a factor.)
- **b.** From the conventional ANOVA model defined in step 1a, derive the mm-ANOVA model by averaging across cases. Define the mm-ANOVA model error term equal to the mean, across cases, of the sum of the conventional ANOVA model error term and random effects involving case.
- **c.** Express the mm-ANOVA model error variance and covariances in terms of the conventional ANOVA model variance components.
- **d.** Determine the mm-ANOVA model covariance constraints implied by step 1c.

2 *Derive the mm-ANOVA model test statistic and its null distribution for a hypothesis express in terms of test accuracies (i.e., expected reader-performance measures)*

- **a.** State the hypothesis of interest in terms of the mm-ANOVA model.
- **b.** Express the hypotheses from step 2a in terms of the conventional ANOVA model.
- **c.** Create the expected-mean-square table for the conventional ANOVA model
- **d.** Determine the conventional ANOVA *F* statistic corresponding to the step 2b hypotheses.
- **e.** Express mm-ANOVA mean squares in terms of conventional ANOVA mean squares.
- **f.** Express *F* from step 2d in terms of the mm-ANOVA model mean squares and *U*, where *U* is a linear function of conventional ANOVA model mean squares that involve case.
- **g.** Express *E* (*U*) in terms of conventional ANOVA model variance components, and then in terms of mm-ANOVA model error covariance parameters using the relationships from step 1c.
- **h.** Modify *F* from step 2f to produce the mm-ANOVA statistic F^*_{OR} by replacing *U* by *E* (*U*), expressed as a linear function of mm-ANOVA covariance parameters.
- **i.** Derive F_{OR} by replacing covariance parameters in F^*_{OR} by estimates that take into account the constraints from step 1d.
- **j.** Determine the approximate null distribution of F_{OR} in the following way: Write the denominator of F_{OR} in the form $b\left(\sum_i a_i \tilde{\mathrm{MS}}_i + \hat{d}\right)$ where the $\tilde{\mathrm{MS}}_i$ are mm-ANOVA model mean squares, $\hat{d}$ is a function of the covariance parameter estimates, and the a_i and b are constants. Then F_{OR} will have an approximate $F_{\mathrm{df}_1,\mathrm{df}_2}$ null distribution, where df_1 is the numerator degrees of freedom for the conventional ANOVA model test statistic in step 2d and df_2 is given by

$$\mathrm{df}_2 = \frac{\left[\sum_i a_i \overline{\mathrm{MS}}_i + \hat{d}\right]^2}{\sum_i \frac{\left[a_i \overline{\mathrm{MS}}_i\right]^2}{\mathrm{df}\left(\overline{\mathrm{MS}}_i\right)}}$$

where $\mathrm{df}\left(\tilde{\mathrm{MS}}_i\right)$ is the degrees of freedom for $\tilde{\mathrm{MS}}_i$, and hence also for MS_i.

3 *Derive confidence intervals for a linear function g* (θ) *of test accuracy parameters*.

- **a.** Write the test accuracy parameter vector θ in terms of the mm-ANOVA model.
- **b.** Write θ in terms of the conventional ANOVA model.
- **c.** Determine the conventional ANOVA estimate for θ, denoted by $\hat{\theta}$.
- **d.** Determine the variance *V* of *g* ($\hat{\theta}$) in terms of conventional ANOVA parameters.
- **e.** Write *V* from step 3d in the form $V = bE\left(\sum a_i \mathrm{MS}_i\right)$ for constants *b* and a_i.
- **f.** Write *V* from step 3e in the form $V = \tilde{b}E\left(\sum \tilde{a}_i \tilde{\mathrm{MS}}_i + U\right)$ where $\tilde{b}$ and $\tilde{a}_i$ are constants and *U* is a linear function of conventional ANOVA mean squares that involve case.

g. Express $E(U)$ in terms of conventional ANOVA model variance components and then in terms of mm-ANOVA model error covariance parameters, using the relationships from step 1c; then rewrite V using this expression for $E(U)$.

h. Derive the variance estimate $\hat{V}$ from V by replacing expected mean squares by mean squares and replacing covariances by estimates that take into account the constraints from step 1d.

i. Derive the degrees of freedom df_2 for $\hat{V}$ using the general formula for df_2 given in step 2j.

j. Write $\hat{\boldsymbol{\theta}}$ from step 3c in terms of the mm-ANOVA model.

k. An approximate $(1-\alpha)$ 100% confidence interval for $g(\boldsymbol{\theta})$ is given by $g\left(\hat{\boldsymbol{\theta}}\right) \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\hat{V}}$, where $\hat{V}$ is determined in step 3h, df_2 in step 3i and $\hat{\boldsymbol{\theta}}$ in step 3j.

4 *Derive the non-null distribution of* F_{OR} from step 2i

a. Compute the noncentrality parameter in terms of the conventional ANOVA model: $\lambda = \dfrac{\mathrm{df}(\mathrm{MS}_{\mathrm{num}})\mathrm{MS}_{\mathrm{num}}|_{\boldsymbol{Y}=E(\boldsymbol{Y})}}{E(\mathrm{MS}_{\mathrm{num}}|\mathrm{H}_0)}$ where MS_{num} is the numerator mean square from the conventional ANOVA F statistic given in step 2d.

b. Express λ in terms of mm-ANOVA parameters by replacing variance components involving case by mm-ANOVA covariances.

c. Determine the denominator degrees of freedom in terms of mm-ANOVA parameters using $\mathrm{df}_2 = \dfrac{\left[\sum_i a_i E(\tilde{\mathrm{MS}}_{\mathrm{i}}) + d\right]^2}{\sum_i \left[a_i E(\tilde{\mathrm{MS}}_{\mathrm{i}})\right]^2 / \mathrm{df}(\tilde{\mathrm{MS}}_{\mathrm{i}})}$ where $b\left(\sum_i a_i \tilde{\mathrm{MS}}_{\mathrm{i}} + d\right)$ is the denominator of F^*_{OR} from step 2h

d. The non-null distribution is given by $F_{df_1,df_2;\lambda}$, where $df_1 = df(MS_{num})$, df_2 is determined in step 4c and λ in step 4b.

Table 6

Mm-ANOVA approach for typical test×reader×case factorial study design

1 Derive the mm-ANOVA model

a. Conventional ANOVA model: $Y_{ijk} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \varepsilon_{ijk}$, $i = 1, \ldots, t$; $j = 1, \ldots, r$; $k = 1, \ldots, c$, with variance components σ_R^2, σ_C^2, σ_{TR}^2, σ_{TC}^2, σ_{RC}^2, $\sigma_{\tau RC}^2$, and σ_ε^2 and constraint $\sum_{i=1}^{t} \tau_i = 0$. Define $\sigma^2 = \sigma_{TRC}^2 + \sigma_\varepsilon^2$.

b. Mm-ANOVA model (note: $\tilde{Y}_{ij} = Y_{ij\bullet}$):

$$\tilde{Y}_{ij} = \mu + \tau_i + R_j + (\tau R)_{ij} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} = C_\bullet + (\tau C)_{i\bullet} + (RC)_{j\bullet} + (\tau RC)_{ij\bullet} + \varepsilon_{ij\bullet} \text{ and } \sum_{i=1}^{t} \tau_i = 0$$

c. Mm-ANOVA error variance and covariances expressed in terms of conventional ANOVA variance components:

$$\sigma_\varepsilon^2 = \frac{1}{c}(\sigma_C^2 + \sigma_{TC}^2 + \sigma_{RC}^2 + \sigma^2), \text{Cov}_1 \equiv \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = \frac{1}{c}(\sigma_C^2 + \sigma_{RC}^2),$$

$$\text{Cov}_2 \equiv \text{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \frac{1}{c}(\sigma_C^2 + \sigma_{\tau C}^2), \text{Cov}_3 \equiv \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = \frac{1}{c}\sigma_C^2, \text{ where } i \neq i', j \neq j'$$

d. Covariance constraints: $\text{Cov}_1 \geq \text{Cov}_3$; $\text{Cov}_2 \geq \text{Cov}_3$; $\text{Cov}_3 \geq 0$

2 Derive the mm-ANOVA test statistic and its null distribution

a. Mm-ANOVA model hypothesis of equal test accuracies: $H_0 : \theta_1 = \cdots = \theta_t$ where $\theta_i = E(\tilde{Y}_{i\bullet})$

b. Conventional ANOVA model hypothesis: $\theta_i = E(Y_{i\bullet\bullet}) = \mu + \tau_i \Rightarrow H_0 : \tau_1 = \cdots = \tau_t$

c. Conventional ANOVA expected mean squares

Mean square	Expected mean square
MS(*T*)	$\frac{rc}{(t-1)}\sum_{i=1}^{t}\tau_i^2 + c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2$
MS(*R*)	$tc\sigma_R^2 + c\sigma_{TR}^2 + t\sigma_{RC}^2 + \sigma^2$
MS(*C*)	$tr\sigma_C^2 + r\sigma_{TC}^2 + t\sigma_{RC}^2 + \sigma^2$
MS(*T* * *R*)	$c\sigma_{TR}^2 + \sigma^2$
MS(*T* * *C*)	$r\sigma_{TC}^2 + \sigma^2$
MS(*R* * *C*)	$t\sigma_{RC}^2 + \sigma^2$
MS(*T* * *R* * *C*)	$\sigma^2 \equiv \sigma_{TRC}^2 + \sigma_\varepsilon^2$

d. Conventional ANOVA test statistic: $F = \frac{\text{MS}(T)}{\text{MS}(T*R) + \text{MS}(T*C) - \text{MS}(T*R*C)}$

e.

$$\overline{\text{MS}}(T) = \frac{1}{c}\text{MS}(T),\ \overline{\text{MS}}(T*R) = \frac{1}{c}\text{MS}(T*R),\ \overline{\text{MS}}(R) = \frac{1}{c}\text{MS}(R)$$

f. $F = \frac{\overline{\text{MS}}(T)}{\overline{\text{MS}}(T*R) + U}$ where $U = \frac{1}{c}\{\text{MS}(T*C) - \text{MS}(T*R*C)\}$

g.

$$E\{\mathrm{MS}(T*C)\} = r\sigma_{TC}^2 + \sigma^2,\ E\{\mathrm{MS}[T*R*C]\} = \sigma^2 \Rightarrow E(U) = \frac{1}{c}(r\sigma_{TC}^2) = r(\mathrm{Cov}_2 - \mathrm{Cov}_3).$$

h.

$$F_{\mathrm{OR}}^{*} = \frac{\overline{\mathrm{MS}}(T)}{\overline{\mathrm{MS}}(T*R) + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)}$$

i.

$$F_{\mathrm{OR}} = \frac{\overline{\mathrm{MS}}(T)}{\overline{\mathrm{MS}}(T*R) + r\max(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3, 0)}$$

j.

Under H_0, $F_{\mathrm{OR}} \approx F_{t-1,\mathrm{df}_2}$ where $\mathrm{df}_2 = \dfrac{\left[\overline{\mathrm{MS}}[T*R] + r\max(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3, 0)\right]^2}{\left[\overline{\mathrm{MS}}[T*R]\right]^2 / \left[(t-1)(r-1)\right]}$

3 Derive confidence intervals

(a) Mm-ANOVA test accuracy parameters: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_t)'$, with $\theta_i = E(\tilde{Y}_{i\bullet})$, $i = 1, \ldots, t$

(b) Corresponding conventional ANOVA parameters: $\theta_i = E(Y_{i\bullet\bullet}) = \mu + \tau_i$

(c) Conventional ANOVA estimate: $\hat{\theta}_i = Y_{i\bullet\bullet}$

CI for $\boldsymbol{l}'(\boldsymbol{\theta})$ with $\boldsymbol{l} = (l_1, \ldots, l_t)'$, $\sum_{i=1}^{t} l_i = 0$:

(d)

$$l'(\hat{\theta}) = \sum_{i=1}^{t} l_i\hat{\theta}_i = \sum_{i=1}^{t} l_i Y_{i\bullet\bullet} = \sum_{i=1}^{t} l_i\tau_i + \sum_{i=1}^{t} l_i[(\tau R)_{i\bullet} + (\tau C)_{i\bullet} + (\tau RC)_{i\bullet\bullet} + \varepsilon_{i\bullet\bullet}] \Rightarrow V = \sum_{i=1}^{t} l_i^2\left[\frac{\sigma_{TR}^2}{r} + \frac{\sigma_{TC}^2}{c} + \frac{\sigma^2}{rc}\right] = \frac{1}{rc}\sum_{i=1}^{t} l_i^2[c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2]$$

(e)

$$V = \frac{1}{rc}\sum_{i=1}^{t} l_i^2 E[\mathrm{MS}(T*R) + \mathrm{MS}(T*C) - \mathrm{MS}(T*R*C)]$$

(f)

$$V = \frac{1}{r}\sum_{i=1}^{t} l_i^2 E[\overline{\mathrm{MS}}(T*R) + U] \text{ where } U = \frac{1}{c}\{\mathrm{MS}(T*C) - \mathrm{MS}(T*R*C)\}$$

(g)

$$E(U) = \frac{r\sigma_{TC}^2}{c} = r(\mathrm{Cov}_2 - \mathrm{Cov}_3) \Rightarrow V = \frac{1}{r}\sum_{i=1}^{t} l_i^2\left\{E\left[\overline{\mathrm{MS}}(T*R)\right] + r(\mathrm{Cov}_2 - \mathrm{Cov}_3)\right\}$$

(h)

$$\hat{V} = \frac{1}{r}\sum_{i=1}^{t} l_i^2\left\{\overline{\mathrm{MS}}(T*R) + \max\left[r\left(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3\right), 0\right]\right\}$$

(i)

$$\mathrm{df}_2 = \frac{\left[\overline{\mathrm{MS}}(T*R) + r\max(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3, 0)\right]^2}{\left[\overline{\mathrm{MS}}(T*R)\right]^2 / \left[(t-1)(r-1)\right]}$$ (same as df_2 in step 2j)

(j) $\hat{\theta}_i = \tilde{Y}_{i\bullet}$

(k)

$$\mathrm{CI}: \sum_{i=1}^{t} l_i\tilde{Y}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\frac{1}{r}\sum_{i=1}^{t} l_i^2\left\{\overline{\mathrm{MS}}(T*R) + \max\left[r\left(\hat{\mathrm{Cov}}_2 - \hat{\mathrm{Cov}}_3\right), 0\right]\right\}}$$

CI for θ_i

(d)

$$\hat{\theta}_i = Y_{i\bullet\bullet} = \tau_i + R_\bullet + C_\bullet + (\tau R)_{i\bullet} + (\tau C)_{i\bullet} + (RC)_{\bullet\bullet} + (\tau RC)_{i\bullet\bullet} + \varepsilon_{i\bullet\bullet} \Rightarrow V = \frac{\sigma_R^2}{r} + \frac{\sigma_C^2}{c} + \frac{\sigma_{TR}^2}{r} + \frac{\sigma_{TC}^2}{c} + \frac{\sigma_{RC}^2}{rc} + \frac{\sigma^2}{rc} = \frac{1}{rc}(c\sigma_R^2 + r\sigma_C^2 + c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma_{RC}^2 + \sigma^2)$$

(e)

$$V = \frac{1}{trc}E[\mathrm{MS}(R) + (t-1)\mathrm{MS}(T*R) + \mathrm{MS}(C) - \mathrm{MS}(R*C) + (t-1)\mathrm{MS}(T*C) - (t-1)\mathrm{MS}(T*R*C)]$$

(f)

$$V = \frac{1}{tr}E\left[\overline{\mathrm{MS}}(R) + (t-1)\overline{\mathrm{MS}}(T*R) + U\right]$$

where

$$U = \frac{1}{c}\{\mathrm{MS}(C) - \mathrm{MS}(R*C) + (t-1)\mathrm{MS}(T*C) - (t-1)\mathrm{MS}(T*R*C)\}$$

(g)

$$E(U) = \frac{tr}{c}(\sigma_C^2 + \sigma_{TC}^2) = tr\mathrm{Cov}_2 \Rightarrow V = \frac{1}{tr}\left\{E\left[\overline{\mathrm{MS}}(R) + (t-1)\overline{\mathrm{MS}}(T*R)\right] + tr\mathrm{Cov}_2\right\}$$

(h)

$$\hat{V} = \frac{1}{tr}\left[\overline{\mathrm{MS}}(R) + (t-1)\overline{\mathrm{MS}}(T*R) + tr\max\left(\hat{\mathrm{Cov}}_2, 0\right)\right]$$

(i)

$$\mathrm{df}_2 = \frac{\left[\overline{\mathrm{MS}}(R) + (t-1)\overline{\mathrm{MS}}(T*R) + tr\max\left(\hat{\mathrm{Cov}}_2, 0\right)\right]^2}{\frac{[\overline{\mathrm{MS}}(R)]^2}{r-1} + \frac{[(t-1)\overline{\mathrm{MS}}(T*R)]^2}{(t-1)(r-1)}}$$

(j) $\hat{\theta}_i = \tilde{Y}_{i\bullet}$

(k)

$$\mathrm{CI}: \tilde{Y}_{i\bullet} \pm t_{\alpha/2;\mathrm{df}_2}\sqrt{\frac{1}{tr}\left[\overline{\mathrm{MS}}(R) + (t-1)\overline{\mathrm{MS}}(T*R) + tr\max\left(\hat{\mathrm{Cov}}_2, 0\right)\right]}$$

4 Derive the non-null distribution $F_{\mathrm{df}_1,\mathrm{df}_2;\lambda}$ of the step-2 F statistic

a. Step 2d F numerator: $\mathrm{MS}_{\mathrm{num}} = \mathrm{MS}(T)$, $E[\mathrm{MS}(T)] = \frac{rc}{(t-1)}\sum_{i=1}^{t}\tau_i^2 + c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2$, df (MS ($T$)) = $t - 1$,

$$E(Y_{ijk}) = \mu + \tau_i \Rightarrow \lambda = \frac{\mathrm{df}(\mathrm{MS}_{\mathrm{num}})\mathrm{MS}_{\mathrm{num}} \mid \mathbf{Y}=E(\mathbf{Y})}{E(\mathrm{MS}_{\mathrm{num}} \mid \mathrm{H}_0)} = \frac{rc\sum_{i=1}^{t}\tau_i^2}{c\sigma_{TR}^2 + r\sigma_{TC}^2 + \sigma^2}$$

b.

$$r\sigma_{TC}^2 + \sigma^2 = c\left[\sigma_\varepsilon^2 - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)\right] \Rightarrow \lambda = \frac{r\sum_{i=1}^{t}\tau_i^2}{\sigma_{TR}^2 + \sigma_\varepsilon^2 - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)}$$

c. Step 2h F^{*}_{OR} denominator $= \overline{MS}(T * R) + r(\text{Cov}_2 - \text{Cov}_3)$,

$$E\ \overline{MS}(T * R) = \frac{1}{c}E(\overline{MS}(T * R)) = \frac{1}{c}(c\sigma^2_{TR} + \sigma^2) = (\sigma^2_{TR} + \sigma^2_{\varepsilon} - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3) \Rightarrow \text{df}_2$$

$$= \left(\frac{[\sigma^2_{TR} + \sigma^2_{\varepsilon} - \text{Cov}_1 + (r-1)(\text{Cov}_2 - \text{Cov}_3)]^2}{\frac{[\sigma^2_{TR} + \sigma^2_{\varepsilon} - \text{Cov}_1 - \text{Cov}_2 + \text{Cov}_3]^2}{(t-1)(r-1)}}\right)$$

d. $F_{OR} \sim^{\cdot} F_{t-1,\text{df}_2;\lambda}$

Table 7

Split-lot design layouts. For nested factors, the level of the nesting factor is given in parentheses; e.g., reader (*t*) 1 in (a) denotes reader 1 nested within test *t*.

a) Reader nested within test. Y_{ijk} = rating for test *i* from reader *j* reading cases 1, …, *c*, with readers nested in test *i*; *i* = 1, …, *t*, *j* = 1, …, *r*, *k* = 1, …, *c*.

		case		
test	reader	1	…	*c*
1	(1)1	Y_{111}	⋯	Y_{11c}
⋮	⋮	⋮	⋱	⋮
1	(1)*r*	Y_{1r1}	⋯	Y_{1rc}
⋮	⋮	⋮	⋮	⋮
t	(*t*)1	Y_{t11}	⋯	Y_{t1c}
⋮	⋮	⋮	⋱	⋮
t	(*t*)*r*	Y_{tr1}	⋯	Y_{trc}

b) Case nested within test. Y_{ijk} = rating for test *i* from reader *j* reading cases 1, …, *c*, with readers nested in test *i*; *i* = 1, …, *t*, *j* = 1, …, *r*, *k* = 1, …, *c*.

		reader		
test	case	1	⋯	*r*
1	(1)1	Y_{111}	⋯	Y_{1r1}
⋮	⋮	⋮	⋱	⋮
1	(1)*c*	Y_{11c}	⋯	Y_{1rc}
⋮	⋮	⋮	⋮	⋮
t	(*t*)1	Y_{t11}	⋯	Y_{tr1}
⋮	⋮	⋮	⋱	⋮
t	(*t*)*c*	Y_{t1c}	⋯	Y_{trc}

c) Case nested within reader. Y_{ijk} = rating for test *i* from reader *j* reading cases 1, …, *c*, with cases nested in reader *j*;*i* = 1, … *t*, *j* = 1, …, *r*, *k* = 1, …, *c*.

		test		
reader	case	1	⋯	*t*
1	(1)1	Y_{111}	⋯	Y_{t11}
⋮	⋮	⋮	⋱	⋮
1	(1)*c*	Y_{11c}	⋯	Y_{t1c}
⋮	⋮	⋮	⋮	⋮
r	(*r*)1	Y_{1r1}	⋯	Y_{tr1}
⋮	⋮	⋮	⋱	⋮
r	(*r*)*c*	Y_{1rc}	⋯	Y_{trc}

d)Reader and case crossed and nested within group. Y_{hijk} = rating assigned by the *j*th reader in group *h* to the *k*th case in group *h* using rest *i*; *h* = 1, … *g*, *i* = 1, … *t*, *j* = 1, …, *r*, *k* = 1, …, *c*. Each reader and case is included in only one group.

			test		
group	reader	case	1	⋯	*t*
1	(1)1	(1)1	Y_{1111}	⋯	Y_{1t11}
⋮	⋮	⋮	⋮	⋱	⋮
1	(1)1	(1)*c*	Y_{111c}	⋯	Y_{1t1c}
⋮	⋮	⋮	⋮	⋮	⋮
1	(1)*r*	(1)1	Y_{11r1}	⋯	Y_{1tr1}
⋮	⋮	⋮	⋮	⋱	⋮
1	(1)*r*	(1)*c*	Y_{11rc}	⋯	Y_{1trc}
⋮	⋮	⋮	⋮	⋮	⋮
g	(*g*)1	(*g*)1	Y_{g111}	⋯	Y_{gt11}
⋮	⋮	⋮	⋮	⋱	⋮
g	(*g*)1	(*g*)*c*	Y_{g11c}	⋯	Y_{gt1c}
⋮	⋮	⋮	⋮	⋮	⋮

d)Reader and case crossed and nested within group. Y_{hijk} = rating assigned by the *j*th reader in group *h* to the *k*th case in group *h* using rest *i*; *h* = 1, … *g*, *i* = 1, … *t*, *j* = 1, …, *r*, *k* = 1, …, *c*. Each reader and case is included in only one group.

			test		
g	(1)*r*	(1)1	Y_{g1r1}	⋯	Y_{gtr1}
⋮	⋮	⋮	⋮	⋱	⋮
g	(*g*)*r*	(*g*)*c*	Y_{g1rc}	⋯	Y_{gtrc}

Table 8

Obuchowski-Rockette split-plot (cases nested within test) analysis of Kundel et al [9] data for soft-copy computed radiographs and screen-film radiographs using trapezoid AUC estimation and jackknife covariance estimation for $t = 2$ tests, $r = 4$ readers. The images were from different patients for each type of radiograph, with 95 images in each group (soft-copy computed radiograph: 66 nondiseased, 29 diseased; screen-film radiograph: 68 nondiseased, 27 diseased).

a. Trapezoid AUCs:

	Test	
	1 (Soft-copy computed radiograph)	2 (Screen-film)
Reader (j)	$\hat{\theta}_{1j}$	$\hat{\theta}_{2j}$
1	0.815	0.818
2	0.767	0.836
3	0.831	0.828
4	0.803	0.834
	$\hat{\theta}_{1\cdot} = .804$	$\hat{\theta}_{2\cdot} = .829$

b. ANOVA table:

Source	df	Sum of squares	Mean square
T	1	0.00125969	0.00125969
R	4	0.00076530	0.00025510
T*R	4	0.00164974	0.00054991

c. Fixed-reader covariance estimates computed from jackknife covariance matrix:
$\hat{\sigma}_e^2 = 0.0023651313,\ \hat{\mathrm{Cov}}_2 = 0.0008800774$

d.

$$F_{\mathrm{OR}} = \frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R) + \max\left(r\hat{\mathrm{Cov}}_2,\ 0\right)} = 0.31$$

e. Denominator degrees of freedom:

$$\mathrm{df}_2 = \frac{\left[\mathrm{MS}(T*R) + \max\left(r\hat{\mathrm{Cov}}_2,\ 0\right)\right]^2}{\frac{\left[\mathrm{MS}(T*R)\right]^2}{(t-1)(r-1)}} = 164.4$$

f. P-value for H_0: $\theta_1 = \theta_2$: $p = \Pr\left(F_{(t-1),\,\mathrm{df}_2} \geq F_{\mathrm{OR}}\right) = 0.579$

g. 95% CI for $\theta_2 - \theta_1$: $\hat{\theta}_{2\cdot} - \hat{\theta}_{1\cdot} \pm t_{\mathrm{df}_2}\sqrt{\frac{2}{r}\left\{\mathrm{MS}(t*R) + r\max\left(\widehat{\mathrm{Cov}}_2, 0\right)\right\}} = (-0.064, 0.114)$

h. Single test 95% confidence intervals using only corresponding data. Note:

$$\mathrm{StdErr}^{(i)} = \sqrt{\frac{1}{r}\left\{\mathrm{MS}(R)^{(i)} + r * \max\left(\widehat{\mathrm{Cov}}_2^{(i)}, 0\right)\right\}}$$

i	$\hat{\theta}_i$	$\widehat{\mathrm{Cov}}_2^{(i)}$	**MS(*R*)$^{(i)}$**	**StdErr$^{(i)}$**	$\mathrm{df}_2^{(i)}$	**95% CI**
1(Soft-copy)	0.804	0.000880	0.000735	0.0326	100.4	0.739, 0.867
2(Screen-film)	0.829	0.000881	0.000070	0.0300	7997.2	0.770, 0.888

Table 9

Number of replications, readers, and cases needed to achieve .80 power to detect a .04 AUC difference between soft- and hard-copy radiographs using a factorial study design, based on estimates from the Kundel et al [9] data, an assumed within-reader within-replication correlation of 0.60, and alpha = .05.

replications (*n*)	readers (*r*)	cases (*c*)	power
1	4	585	0.800
1	5	366	0.801
1	6	266	0.800
1	7	210	0.802
1	8	173	0.801
2	4	348	0.800
2	5	218	0.801
2	6	158	0.800
2	7	125	0.802
2	8	103	0.802