Published in final edited form as:
Acad Radiol. 2011 February ; 18(2): 129–142. doi:10.1016/j.acra.2010.09.007.

Power Estimation for Multireader ROC Methods: An Updated and Unified Approach

Stephen L. Hillis
Center for Research in the Implementaion of Innovative Strategies in Practice (CRIISP), Iowa City VA Medical Center, Iowa City, IA

Department of Biostatistics, University of Iowa, Iowa City, IA

Nancy A. Obuchowski
Department of Quantitative Health Sciences/JJN3, and the Imaging Institute, The Cleveland Clinic, Cleveland, OH

Kevin S. Berbaum
Department of Radiology, University of Iowa, Iowa City, IA

Abstract

Rationale and Objectives—We describe a step-by-step procedure for estimating power and sample size for planned multireader receiver operating characteristic (ROC) studies that will be analyzed using either the Dorfman-Berbaum-Metz (DBM) or Obuchowski-Rockette (OR) method. This procedure updates previous approaches by incorporating recent methodological developments and unifies the approaches by allowing inputs to be conjectured parameter values or outputs from either a DBM or OR pilot-study analysis.

Materials and Methods—Power computations are described in a step-by-step procedure and the theoretical basis for the procedure is described. Updates include using the currently recommended denominator degrees of freedom, accounting for different pilot and planned study normal-to-abnormal case ratios, and a new method for computing the OR test-by-reader variance component.

Results—Using a real data set we illustrate how to compute the power for two planned studies, one having the same normal-to-abnormal case ratio as the pilot study and the other having a different ratio. In a simulation study we show that the proposed procedure gives mean power estimates close to the true power.

Conclusions—Application of the updated procedure is straightforward. It is important that pilot data be comparable to the planned study with respect to the modalities, reader expertise, and case selection. Variability of the power estimates warrants further investigation.

Keywords

ROC curve; sample size; power; multireader

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclaimer The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

1. Introduction

Receiver operating characteristic (ROC) curve analysis is a well-established method for evaluating and comparing the performance of diagnostic tests for radiological imaging studies. Throughout we assume that rating data have been collected using the study design where multiple readers (typically radiologists) assign disease-severity or disease-likelihood ratings, using one or more tests, to the same images using either a discrete (e.g., 1, 2, 3, 4, 5) or a quasi-continuous (e.g., 0–100%) scale. From these ratings, ROC curves and corresponding accuracy estimates are computed for each reader and each test, in order to assess how well a test performs or to compare the performance of tests. In such studies there is variability between cases and between readers. Thus it is important that results generalize to both the corresponding case and reader populations; methods that accomplish this goal are commonly referred to as *multireader multicase* (MRMC) methods.

Two popular MRMC methods are those proposed by Dorfman, Berbaum, and Metz (DBM) [1,2] and by Obuchowski and Rockette (OR) [3,4]. For the OR method, power computation using conjectured parameter estimates is discussed by Obuchowski [4,5] and Zhou et al [6]; for the DBM method power computation based on pilot-study estimates or conjectured parameter values is discussed by Hillis and Berbaum [7]. Since the publication of these articles, it has been shown [8] that both the DBM and OR methods can be improved by using a common denominator degrees of freedom, ddf_H, for the F statistic for testing for equality of tests. When both methods use ddf_H, the DBM method can be viewed as an implementation of the OR method using jackknife covariance estimates, with both methods yielding the same conclusions. Furthermore, Reference [9] shows that if the OR method is not based on jackknife covariance estimates, then *quasi pseudovalues* can be generated that give the same results when analyzed by the DBM method. Thus we can consider the DBM and OR procedures to be equivalent. These developments in the DBM procedure and its relationship with the OR procedure are summarized in Reference [10].

Although equivalent results can be obtained using either method, the DBM model is not statistically acceptable since several of its assumptions are not true [8]. Thus the DBM model should be viewed only as a “working” model; although “pretending” that the DBM model is correct generally leads to valid inferences, parameters for the model are difficult to interpret in terms of the model. For these reasons, theoretical justification for results provided in this paper will be based on the OR model.

Our purpose is to describe a step-by-step procedure for computing power (and hence sample size) for either method. This procedure updates previous approaches by incorporating ddf_H, accounting for different pilot- and planned-study normal-to-abnormal case ratios, and incorporating a new method for estimating the OR test-by-reader variance component. The procedure unifies the approaches by allowing both procedures to be based on either pilot data or conjectured parameter values, and yields the same results regardless of whether the inputted values are DBM or OR pilot-data analysis outputs or conjectured parameter values. We describe the procedure for the OR method and then show how this same procedure can also be used with inputs obtained from a DBM analysis. The procedure is illustrated in an example and its performance is evaluated in a simulation study.

2. Materials and Methods

2.1. Design and notation

We assume that rating data have been collected from a test×reader×case factorial study design, where each case undergoes each diagnostic test and the resulting images are evaluated once by each reader. (We use *test* to refer to a diagnostic test, modality, or treatment.) Letting Z_{ijk} denote the rating assigned to the kth case by the jth reader using the ith test, the observed rating data consists of the Z_{ijk}, $i = 1,\ldots,t$, $j = 1,\ldots,r$, $k = 1,\ldots, c$, where t is the number of tests, r the number of readers, and c the number of cases. In addition, each case is classified as diseased or nondiseased according to an available reference standard.

We let $\widehat{\theta}_{ij}$ denote the AUC estimate based on all of the data for the ith test and jth reader; however, more generally $\widehat{\theta}_{ij}$ can be any ROC accuracy estimate, such as the partial AUC, sensitivity for a fixed specificity, or specificity for a fixed sensitivity. We let θ_{ij} denote the corresponding population AUC, defined statistically by $\theta_{ij}=E\left(\widehat{\theta}_{ij}\right)$ for fixed i. That is, for a given test i and case sample size c, θ_{ij} is the expected AUC for a randomly selected reader reading c randomly selected cases.

2.2. The DBM procedure

For the DBM procedure, AUC pseudovalues are computed using the Quenouille-Tukey jackknife [11–13] separately for each reader-test combination. Let Y_{ijk} denote the AUC pseudovalue for test i, reader j, and case k; by definition $Y_{ijk}=c\widehat{\theta}_{ij}-(c-1)\widehat{\theta}_{ij(k)}$, where $\widehat{\theta}_{ij}(k)$ denotes the AUC estimate when data for the kth case are omitted. Treating the Y_{ijk} as the outcomes, the original DBM procedure specified testing for a test effect using a fully-crossed three-factor ANOVA, with test treated as a fixed factor and reader and case as random factors; the original DBM estimate of θ_{ij} was $Y_{ij\cdot}=\frac{1}{c}\sum_{k=1}^{c}Y_{ijk}$, which is the jackknife accuracy estimate corresponding to $\widehat{\theta}_{ij}$. (A subscript replaced by a dot indicates that values are averaged across the missing subscript.) Later, Hillis et al [9] recommended that the DBM method be used with normalized pseudovalues, defined by $Y_{ijk}^{*}=Y_{ijk}+\left(\widehat{\theta}_{ij}-Y_{ij\cdot}\right)$. For normalized pseudovalues the DBM accuracy estimate, given by $Y_{ij\cdot}^{*}$, is equal to $\widehat{\theta}_{ij}$ and hence the analysis is not restricted to jackknife accuracy estimates.

Let MS(T)_Y, MS(T*R)_Y, MS(T*C)_Y, and MS(T*R*C)_Y denote the test, test×reader, test×case, and test×reader×case mean squares for the DBM three-way ANOVA of the pseudovalues. (Here the Y subscript is used to indicate that these mean squares are computed from pseudovalues, in contrast to the OR mean squares discussed in the next section that are computed from reader-level AUCs.) The DBM F statistic for testing the null hypothesis of no test effect is

$$F_{\text{DBM}}=\frac{\text{MS(T)}_Y}{\text{MS}(\text{T}*\text{R})_Y+H\left[\text{MS}(\text{T}*\text{C})_Y-\text{MS}(\text{T}*\text{R}*\text{C})_Y\right]} \quad (1)$$

where the function $H(\cdot)$ is defined by

$$H(x)=\begin{cases} x & \text{if } x>0 \\ 0 & \text{if } x\le 0 \end{cases}$$

Equation 1 is recommended by Hillis et al [9] and differs slightly from the original DBM formulation in that less data-based model reduction is allowed. Hillis and Berbaum [7] used Equation 1 in their power algorithm; although they used raw instead of normalized pseudovalues, the use of normalized pseudovalues does not alter their algorithm.

Hillis [8] showed that the DBM method has improved performance if the following denominator degrees of freedom for F_{DBM} is used:

$$\mathrm{ddf}_H=\frac{\left\{\mathrm{MS(T*R)}_Y+H\left[\mathrm{MS(T*C)}_Y-\mathrm{MS(T*R*C)}_Y\right]\right\}^2}{\left[\mathrm{MS(T*R)}_Y\right]^2/[(t-1)(r-1)]} \tag{2}$$

The updated power procedure that we will present incorporates Equation 2, which was not used by Hillis and Berbaum [7] since it had not yet been proposed. We note that since it was proposed in 2007 by Hillis [8], ddf_H has been incorporated into freely available DBM analysis software [14–16].

2.3. The OR procedure

Obuchowski and Rockette [3] analyze AUC estimates using a test × reader factorial ANOVA model, 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 for each test. Their model, which we refer as the *OR model*, can be written as

$$\widehat{\theta}_{ij}=\mu+\tau_i+R_j+(TR)_{ij}+\epsilon_{ij} \tag{3}$$

$i = 1,\ldots,t$, $j = 1,\ldots,r$, where τ_i denotes the fixed effect of test i, R_j denotes the random effect of reader j, $(TR)_{ij}$ denotes the random test × reader interaction, and ϵ_{ij} is the error term. The R_j and $(TR)_{ij}$ are assumed to be mutually independent and normally distributed with zero means and respective variances σ^2_R, reflecting differences in reader ability, and σ^2_{TR}, reflecting test-by-reader interaction. The ϵ_{ij} are assumed to be normally distributed with zero mean and variance σ^2_ϵ, which represents variability attributable to cases and within-reader variability that describes how a reader interprets the same image in different ways on different occasions. The ϵ_{ij} are independent of the R_j and $(TR)_{ij}$. Equi-covariance of the errors between readers and tests is assumed, resulting in three possible covariances:

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

It follows from model (3) that σ^2_ϵ, 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 [3] suggest the following ordering for the covariances:

$$\mathrm{Cov}_1\geq\mathrm{Cov}_2\geq\mathrm{Cov}_3\geq 0. \tag{4}$$

The OR statistic for testing the null hypothesis of no test effect is given by

$$F_{\mathrm{OR}}=\frac{\mathrm{MS}(T)_{\widehat{\theta}_{ij}}}{\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}+H\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right]} \tag{5}$$

where $\mathrm{MS}(T)_{\widehat{\theta}_{ij}}$ and $\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}$ are the two-way ANONVA test and test-by-reader mean squares; these mean squares are based on the AUC outcomes, in contrast to the DBM mean squares that are based on the case-level pseudovalues. The quantities $\widehat{\mathrm{Cov}_2}$ and $\widehat{\mathrm{Cov}_3}$ denote estimates for Cov_2 and Cov_3, respectively. Note that Equation 5 incorporates the constraint given by Equation 4 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 ROC analysis methods that treat cases as random but readers as fixed, such as jackknifing, bootstrapping, parametric methods, or the method proposed by DeLong et al [17] for trapezoidal-rule (or empirical) AUC estimates [18]. The OR estimates obtained from averaging corresponding fixed-reader AUC variances and covariances are denoted by $\widehat{\sigma_\epsilon^2}$, $\widehat{\mathrm{Cov}_1}$, $\widehat{\mathrm{Cov}_2}$, and $\widehat{\mathrm{Cov}_3}$.

Hillis [8] shows that F_{OR} has an approximate null $F_{t-1,\mathrm{df}2}$ distribution, where

$$\mathrm{df}_2=\frac{\left\{E\left[\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}\right]+r\left(\mathrm{Cov}_2-\mathrm{Cov}_3\right)\right\}^2}{\frac{\left[E\left[\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}\right]\right]^2}{(t-1)(r-1)}} \tag{6}$$

and suggests estimating df_2 by

$$\mathrm{ddf}_{\mathrm{H}}=\frac{\left\{\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}+H\left[r\left(\widehat{\mathrm{Cov}_2}-\widehat{\mathrm{Cov}_3}\right)\right]\right\}^2}{\frac{\left[\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}\right]^2}{(t-1)(r-1)}} \tag{7}$$

Note that the estimate ddf_H replaces the parameters in df_2 by estimates; in particular, the expected test × reader mean square, $E\left[\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}\right]$, is replaced by the observed mean square, $\mathrm{MS}(T*R)_{\widehat{\theta}_{ij}}$, and $r(\mathrm{Cov}_2 - \mathrm{Cov}_3)$ is replaced by $H\left[r\left(\widehat{\mathrm{Cov}_2}-\widehat{\mathrm{Cov}_3}\right)\right]$, which incorporates the model covariance constraints given by Equation 4. He also shows that ddf_H results in improved performance compared to the denominator degrees of freedom, $\mathrm{ddf}_0 = (t - 1)(r - 1)$, originally proposed by Obuchowski and Rockette [3]. A 100 $(1 - \alpha)$% confidence interval for $\theta_i - \theta_{i'}$, $i \neq i'$, is given by $\widehat{\theta_i}. - \widehat{\theta_{i'}}. \pm t_{\alpha/2;\mathrm{ddf}_H}\sqrt{\frac{2}{r}\mathrm{MSden}_{\mathrm{OR}}}$, where $\mathrm{MSden}_{\mathrm{OR}}$ is the denominator of the right-hand-side of Equation 5.

If the null hypothesis of equal tests is not true, then F_{OR} has an approximate $F_{t-1,\mathrm{df}_2;\Delta}$ distribution where the noncentrality parameter is given by

$$\Delta=\frac{r\Sigma_{i=1}^{t}(\theta_i-\theta.)^2}{\sigma_{TR}^2+\sigma_\varepsilon^2-\mathrm{Cov}_1+(r-1)\left(\mathrm{Cov}_2-\mathrm{Cov}_3\right)} \tag{8}$$

and $\theta_i = \mu + \tau i$ is the expected accuracy measure for test *i*. This result is stated by Obuchowski [4] and a detailed proof is provided in Reference [8].

Power estimation for the OR method has been previously described [4,5]. However, these references use the originally proposed denominator degrees of freedom, $\text{ddf}_0 = (t - 1)(r - 1)$, which has been shown to give overly conservative inferences [8]. The updated algorithm for computing power differs from the previously published OR *power* methods in that it uses ddf_H; in addition, it estimates σ_{TR}^2, and hence the noncentrality parameter Δ, differently.

2.3.1. OR and DBM relationships—As previously noted, the DBM procedure can be viewed as an implementation of the OR procedure if the OR covariance estimates are based on jackknife covariance estimates and the DBM procedure uses normalized pseudovalues. In this case, there is a oneto-one correspondence between the parameters and outputs for the two procedures and Equations 2 and 7 yield the same value [9]. The relationships between the OR and DBM outputs are given in Table 1. Thus to use the OR power procedure with DBM output values, we only need to transform the DBM values to their corresponding OR values.

2.3.2. OR method in terms of correlations and the σ_c^2, σ_w^2 parameterization—The OR method can also be notationally described with population correlations $r_i=\text{Cov}_i/\sigma_\varepsilon^2$ replacing corresponding Cov_i, and estimated correlations $\widehat{r_i}=\widehat{\text{Cov}}_i/\widehat{\sigma}_\varepsilon^2$ replacing corresponding $\widehat{\text{Cov}}_i$, $i = 1, 2, 3$. All of the results cited thus far can be equivalently expressed in terms of correlations; thus the choice of which notation to use is not important. An advantage of using correlations is that their interpretation does not depend on sample size. A disadvantage is possible misunderstanding about the definition of the denominator used to compute them. For example, Obuchowski and Rockette [3] write σ_ε^2 as $\sigma_c^2+\sigma_w^2$, where σ_c^2 denotes variability attributable to cases and σ_w^2 denotes within-reader variability, and then define $r_i=\text{Cov}_i/\sigma_c^2$. This definition is convenient to use when only σ_c^2 can be estimated from the data (as discussed in the next paragraph); in this case, one can think of the error terms as partitioned into two parts: $\varepsilon_{ij} = u_{ij} + w_{ij}$ where $\text{var}\left(u_{ij}\right)=\sigma_c^2$, $\text{var}\left(w_{ij}\right)=\sigma_w^2$, the w_{ij} are mutually independent and are independent of the u_{ij}, but the u_{ij} are correlated and have the same covariances as the ε_{ij}. Practically either definition will give similar correlations since σ_w^2 is typically neglible compared to σ_c^2.

We note that formulas for the variance of the AUC or other ROC accuracy measure based on assumed parametric models that ignore within-reader inconsistency will give estimates of σ_c^2 rather than of σ_ε^2. For example, these methods include the AUC variance estimates proposed by Hanley and McNeil [18] and Obuchowski [19]. Basing power estimates on estimates of σ_c^2, obtained from such methods, technically necessitates also estimating σ_w^2 separately from repeated readings, as previously has been suggested by Obuchowski [4,5], or else using a conjectured value of σ_w^2. In contrast, resampling methods such as bootstrapping and jackknifing, as well as the method proposed by DeLong et al [17] yield estimates of $\sigma_\varepsilon^2=\sigma_c^2+\sigma_w^2$. Thus there is no need to estimate σ_w^2 separately for power computation based on repeated readings when these methods are used to estimate the error variance. However, as previously noted, since since σ_w^2 is typically neglible compared to σ_c^2, using σ_c^2 in place of σ_ε^2 in the power computations will make little difference.

2.4. Updated and unified OR/DBM power computation procedure

In this section we present a step-by-step procedure for computing power for either the OR or DBM procedure. The procedure is described for a two-sided test comparing two modalities based either on data from a pilot or previous study, or on conjectured parameter values. We assume that the ratio of normal to diseased cases in the planned study is approximately the same as in the pilot or previous study. Later we discuss how the procedure can be modified for a one-sided test and for the situation where the pilot and planned study normal-to-abnormal case ratios differ.

The steps of the procedure are the following: (1) specify the effect size; (2) transform OR or DBM outputs into OR parameter estimates, or use conjectured OR parameter values; (3) transform the OR parameter values into OR noncentrality parameter and denominator degrees of freedom values for specified case and reader sample sizes; and (4) compute the power based on the estimated OR noncentrality parameter and denominator degrees of freedom. Below we describe the steps in detail. Theoretical details are provided in Appendix A (available online at www.academicradiology.org).

1. *Specify the effect size*. Specify the effect size, denoted by *d*, that the researcher wants to be able to detect with sufficient power. The effect size is the absolute difference of the two population ROC accuracy measures. For example, if the AUC is the outcome of interest, then $d = |\text{AUC}_1 - \text{AUC}_2|$, where AUC_1 and AUC_2 are the population AUC values for the two tests. For a given number of cases *c*, the population AUC is the expected AUC for a randomly selected reader reading *c* randomly selected cases.

2. Transform OR or DBM outputs into OR parameter estimates, or use conjectured OR parameter values. If using outputs from an analysis of pilot data, let *c** denote the number of cases for the pilot study. If using conjectured parameters, let *c** denote the number of cases corresponding to the conjectured value of σ_ε^2. Use step 2a or 2b below, depending on whether an OR or DBM analysis of pilot data was performed, or step 2c if conjectured values are inputted.

   (a) *Using OR outputs*. Let $\overline{\text{MS}}\,(T)$ and $\overline{\text{MS}}\,(T * R)$ denote the test and test × reader mean squares resulting from the OR pilot-data analysis, and let $\hat{\sigma}_\epsilon^2, \widehat{\text{Cov}}_1, \widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$ denote the fixed-reader variance and co-variance estimates. (If correlations are available instead of covariances, then compute the covariances using $\widehat{\text{Cov}}_i = r_i \hat{\sigma}_\varepsilon^2$ or $\widehat{\text{Cov}}_i = r_i \hat{\sigma}_c^2$, $i = 1, 2, 3$, depending on the definition of the correlation as discussed in Section 2.3.2.) Estimate σ_{TR}^2 using

   $$\hat{\sigma}_{TR}^2 = \overline{\text{MS}}\,(T * R) - \hat{\sigma}_\varepsilon^2 + \widehat{\text{Cov}}_1 + H\left(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3\right) \quad (9)$$

   If $\hat{\sigma}_{TR}^2 < 0$ then set $\hat{\sigma}_{TR}^2$ equal to zero or to a positive conjectured value for the remaining steps - see Section 2.5.3 for further discussion of this point.

   (b) *Using DBM outputs*. Compute the OR quantities $\overline{\text{MS}}\,(T)$, $\overline{\text{MS}}\,(T * R)$, $\hat{\sigma}_\epsilon^2, \widehat{\text{Cov}}_1, \widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$ from theDBM mean squares using Table 1, and then proceed with step 2a.

   (c) *Using conjectured inputs*. This step is similar to step 2a, except that $\hat{\sigma}_{TR}^2, \hat{\sigma}_\epsilon^2, \widehat{\text{Cov}}_1, \widehat{\text{Cov}}_2$, and $\widehat{\text{Cov}}_3$ are conjectured OR parameter values rather than estimates from pilot data. When using conjectured inputs, it is typically conceptually easier to think in terms of the correlations and then compute the corresponding covariances. As previously noted, *c** should denote the number of cases corresponding to $\widehat{\sigma}^2_\varepsilon$, which represents the AUC variance due to cases for a given treatment and fixed reader. Zhou, Obuchowski, and McClish [6, pp. 298–304] discuss choosing values for conjectured inputs. One could also first start with conjectured DBM parameter values and then transform them to OR parameter values since there is a one-to-one transformation between the parameters – these relationships are provided in Table 2.

3. Compute the noncentrality parameter and denominator degrees of freedom estimates for specified case and reader sample sizes. Let *r* and *c* denote the number of readers and cases, respectively, for which we want to compute power. Compute

$$\widehat{\Delta}=\frac{\frac{r}{2}d^2}{\widehat{\sigma}^2_{TR}+\frac{c*}{c}\left\{\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}}_1+(r-1)H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right\}} \qquad (10)$$

and

$$\widehat{\mathrm{df}}_2=\frac{\left[\widehat{\sigma}^2_{TR}+\frac{c*}{2}\left(\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}}_1+(r-1)H\left[\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right]\right)\right]^2}{\frac{\left\{\widehat{\sigma}^2_{TR}+\left(\frac{c*}{c}\right)\left[\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}}_1-H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right]\right\}^2}{r-1}}$$

Here $\widehat{\Delta}$ is the estimated noncentrality parameter and $\widehat{\mathrm{df}}_2$ is the estimated denominator degrees of freedom for the distribution of F_{OR} (Eq. 5). The above formulas were derived for $t = 2$ tests. It is easy to show that $\widehat{\mathrm{df}}_2$ has the same value as ddf_H (Eq. 7) for $c = c*$.

4. Compute the power based on the estimated OR noncentrality parameter and denominator degrees of freedom. Let $F_{1,\nu;\delta}$ denote a random variable having a noncentral *F* distribution with degrees of freedom 1 and ν and noncentrality parameter δ, and let $F_{1-\alpha;1,\nu}$ denote the $100(1-\alpha)$th percentile of a central *F* distribution with degrees of freedom 1 and ν. The estimated power for a two-sided test with significance level α is given by

$$\mathrm{power}=\Pr\left(F_{1,\widehat{\mathrm{df}}_2;\widehat{\Delta}}>F_{1-\alpha;1,\widehat{\mathrm{df}}_2}\right)$$

treating $\widehat{\mathrm{df}}_2$ and $\widehat{\Delta}$ as fixed.

2.5. Other considerations

2.5.1. Accounting for different pilot and planned study normal-to-abnormal case ratios—The preceding power-computation procedure is based on the assumption that the abnormal-to-normal case ratios are similar for the pilot and planned studies. This assumption is important since the fixed-reader covariances and variance depend on the abnormal-to-normal case ratio. For the situation where the researcher expects or wants the planned study to have a normal-to-abnormal case ratio that differs considerably from that of the pilot data, we suggest the following ad hoc approach. From the group (normals or abnormals) that will be proportionately more represented in the planned study than in the pilot study, sample with replacement enough cases to achieve the desired balance between the two groups. Combine these cases with the cases from the other group to create a data set with the desired ratio between normal and abnormal cases. Repeat this process to create several (e.g., 10) data sets having the desired normal-to-abnormal case balance. For each of these data sets compute the fixed-reader covariance matrix and corresponding OR covariances. Use the averages of the OR covariances for the power computation. Note that for the power procedure *c** will not be the number of cases in the original pilot study, but rather the number of cases in each of the "new" pilot data sets.

For the power procedure we can use the estimate of σ^2_{TR} obtained from the original pilot data before doing any resampling. To understand why this is appropriate, define

$$\eta_{ij}=\mu+\tau_i+R_j+TR_{ij} \tag{11}$$

From Equation (3) if follows that for given test *i* and fixed reader *j*, $\eta_{ij}=E\left(\widehat{\theta}_{ij}\right)$; this is the expected or mean AUC across the population of cases. Thus η_{ij} is the latent or true AUC for test *i* and reader *j*, which can be loosely interpreted as the AUC that would result if reader *j* read a very large number of cases. It follows that σ^2_{TR} can be interpreted as the interaction variance component for the η_{ij}, and hence the value of this parameter does not depend on the ratio or numbers of normals and abnormals in the sample. We note that alternatively estimating σ^2_{TR} using Equation 9 from each of the 10 generated data sets would not be valid, since Equation 9 assumes that both readers and cases are random units but our generated data sets only treated cases as random.

2.5.2. Comparison with earlier results—As previously noted, the proposed power procedure updates previous DBM and OR power procedures, as described in References [4–7], by incorporating the new degrees of freedom ddf_H suggested by Hillis [8]. In addition, our estimate of σ^2_{TR} (Eq. 9) for the OR method updates the estimate previously proposed in References [4–6]; this previous estimate was a function of the sample variances of the AUCs across readers within each test and the between-test sample correlation of the AUCs. In Appendix B (available online at www.academicradiology.org) we show that this previous estimate actually estimates $\sigma^2_{TR}+\sigma^2_{\varepsilon}-\text{Cov}_1-\text{Cov}_2+\text{Cov}_3$, with $\sigma^2_{\varepsilon}-\text{Cov}_1-\text{Cov}_2+\text{Cov}_3\geq 0$; thus it is likely that the previous estimator tended to overestimate σ^2_{TR}.

Although a previously available SAS macro [20] for computing power based on DBM outputs had taken into account ddf_H, we note that the power algorithm presented in this paper gives somewhat different results when the DBM variance component estimates for σ^2_{TR} and σ^2_{TC} are both zero, due to the way that the covariance constraints are incorporated.

2.5.3. What to do if $\widehat{\sigma}^2_{TR}<0$—It has been our experience that often the pilot estimate of the test × reader interaction variance component σ^2_{TR} is less than zero, as it is in the Example in the next section. Since the typical radiological imaging study has only a few readers, we expect the precision of the σ^2_{TR} estimate will be low, and hence it is not surprising that estimates of σ^2_{TR} will often not be positive, especially if the true value of σ^2_{TR} is close to zero. In such situations one choice is to set the variance component equal to zero in step 2. However, it seems reasonable that in most studies there should be some interaction, suggesting that when the estimate is not positive we may want to conservatively use a positive value for this variance component instead of zero. One way to decide on a reasonable positive value is to consider estimates for σ^2_{TR} from similar studies, keeping in mind, however, that estimates computed as proposed in References [4–6] tend to overestimate σ^2_{TR} as previously discussed.

Alternatively, we can specify σ^2_{TR} by considering its interpretation in terms of the latent AUCs, the η_{ij}, as defined by Equation 11. Specifically, for fixed tests i and i', it is easy to show that

$$\left(\eta_{ij} - \eta_{i'j}\right) - \left(\eta_{ij} - \eta_{i'j'}\right) \sim \mathcal{N}\left(0, 4\sigma^2_{TR}\right)$$

For example, if reader 1 has latent AUC values of .95 and .90 for tests 1 and 2, respectively, and reader 2 has corresponding latent AUC values of .93 and .91, then $\eta_{11} - \eta_{21} - (\eta_{12} - \eta_{22}) = (.95 - .90) - (.93 - .91) = .05 - .02 = .03$. The quantity $\eta_{ij} - \eta_{i'j} - (\eta_{ij'} - \eta_{i'j'})$ can be interpreted as the difference of the two intra-reader latent AUC differences for randomly selected readers j and j'. Thus σ^2_{TR} is equal to one-fourth of the variance of the difference of the intra-reader latent AUC differences for two randomly chosen readers.

Suppose it seems reasonable that for a randomly selected pair of readers the absolute difference of their intra-reader latent AUC differences will be bounded by a specified value l (e.g., $l = .06$) with probability ≥ .95; i.e., $\Pr\left(|\eta_{ij} - \eta_{i'j} - (\eta_{ij}' - \eta_{i'j'})| \le l\right) \ge .95$. Then since the probability is .95 that a normal random variable is within 1.96 standard deviations of its mean, we have

$$l \ge (1.96)\sqrt{\operatorname{var}\left(\eta_{ij} - \eta_{i'j} - \eta_{ij'} + \eta_{i'j'}\right)} = (1.96)\,(2\sigma_{TR})$$

i.e.,

$$\sigma TR \le \frac{l}{3.92} \text{and } \sigma^2_{TR} \le \left(\frac{l}{3.92}\right)^2$$

For $l = .06$ we have $\sigma_{TR} \le \frac{.06}{3.92} = .01531$, or equivalently $\sigma^2_{TR} \le .01531^2 = .00023$. Thus if $\widehat{\sigma}^2_{TR} < 0$ it would be reasonable to set $\widehat{\sigma}^2_{TR} = .00023$ in step 2 if $l = .06$ seems like a reasonable 95% bound. Table 3 presents values of σ^2_{TR} corresponding to various values of l.

2.5.4. One-sided tests—To compute power for a one-sided test, the only change that needs to be made in the power procedure is to set the significance level to twice the nominal level for the planned study. Although this approach noticeably *overestimates* power for very small effect sizes, the *overestimate* will be negligible for a clinically relevant effect size.

3. Results

Throughout this section we assume a .05 significance level.

3.1. Example: Spin echo versus cine MRI for detection of aortic dissection

Our example study [21] compares the relative performance of single spin-echo magnetic resonance imaging (MRI) to cinematic presentation of MRI for the detection of thoracic aortic dissection. There were 45 patients with an aortic dissection and 69 patients without a dissection imaged with both spin-echo and cinematic MRI. Five radiologists independently interpreted all of the images using a five-point ordinal scale: 1 = definitely no aortic dissection, 2 = probably no aortic dissection, 3 = unsure about aortic dissection, 4 = probably aortic dissection, and 5 = definitely aortic dissection.

Suppose that the researcher would like to know what combinations of reader and case sample sizes for a similar study will have at least .80 power to detect an absolute difference of .05 between the modality AUCs. We first show how to determine the power for 8 readers and 240 cases, based on an OR and DBM analysis of the data. Then we present the smallest case sample size for each of several reader sample sizes that yields .80 power.

Situation 1: Similar normal-to-abnormal ratios. The OR analysis of the data is presented in Table 4. Part (a) presents the AUCs corresponding to ROC curves estimated by the PROPROC procedure [22,23]; part (b) the ANOVA table; part (c) the jackknife covariance matrix for the AUCs, treating readers as fixed; part (d) the variance and covariance estimates based on the covariance matrix in part (c); part (e) the correlations, computed using $\widehat{r_i}=\widehat{\mathrm{Cov}_i}/\widehat{\sigma}^2_\varepsilon$; part (f) the OR *F* statistic; and part (g) ddf_H. From part (h) the *p*-value for testing the hypothesis of equal modalities is .092, and from part (i) a 95% confidence interval for the difference of the population AUCs (spin-echo – cinematic) is given by (–0.0073, 0.0921) Thus there is not sufficient evidence that the modalities differ (*p* = .092).

Treating this study as a pilot study, the power computation steps are as follows:

1. *Specify the effect size*. The effect size of interest is *d* = .05.
2. *Transform outputs into OR parameter estimates*. For the pilot data *c** = 114. From Table 4 we have $\overline{\mathrm{MS}}\,(T) = 0.004003$, $\overline{\mathrm{MS}}(T * R) = 0.000623$, $\widehat{\sigma}^2_\varepsilon = 0.001394$, $\widehat{\mathrm{Cov}_1} = 0.000352$, $\widehat{\mathrm{Cov}_2} = 0.000347$, and $\widehat{\mathrm{Cov}_3} = 0.000221$. Substituting these values into Equation (9) yields $\widehat{\sigma}^2_{TR} = -0.000294$. Since $\widehat{\sigma}^2_{TR}<0$ then we have two choices: either set $\widehat{\sigma}^2_{TR}$ equal to zero or to a conjectured positive value for the remaining steps. In our computations below we set it to zero.
3. Compute the noncentrality parameter and denominator degrees of freedom estimates. We want to compute the power for a study with *r* = 8 readers and *c* = 240 cases. We compute

$$\begin{aligned}\widehat{\mathrm{df}_2} &= \frac{\left\{\widehat{\sigma}^2_{TR}+\frac{c*}{c}\left(\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}_1}\right)+\frac{c*}{c}(r-1)H\left(\widehat{\mathrm{Cov}_2}-\widehat{\mathrm{Cov}_3}\right)\right\}^2}{\frac{\left[\widehat{\sigma}^2_{TR}+\left(\frac{c*}{c}\right)\left\{\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}_1}-H\left[\left(\widehat{\mathrm{Cov}_2}-\widehat{\mathrm{Cov}_3}\right)\right]\right\}\right]^2}{r-1}} \\ &= \frac{\left[0+\frac{114}{240}(0.001394-0.000352)+\frac{114}{240}[7(0.000347-0.000221)]\right]^2}{\frac{\left[\frac{114}{240}[0.001394-0.000352-(0.000347-0.000221)]\right]}{7}} \\ &=30.6\end{aligned}$$

and

$$\begin{aligned}\widehat{\Delta} &= \frac{\frac{r}{2}d^2}{\widehat{\sigma}^2_{TR}+\frac{c*}{c}\widehat{\sigma}^2_\varepsilon-\widehat{\mathrm{Cov}_1}+H\left[(r-1)\left(\widehat{\mathrm{Cov}_2}-\widehat{\mathrm{Cov}_3}\right)\right]} \\ &= \frac{\frac{8}{2}(.05)^2}{\left\{0+\frac{114}{240}(0.001394-0.000352)+\frac{114}{240}[7(0.000347-0.000221)]\right\}}=10.98\end{aligned}$$

4. *Compute the power*. The estimated power for *r* = 8, *c* = 240, α = .05 is given by

$$\begin{aligned}\text{power} \quad &= \Pr\left(F_{1,\widehat{\text{df}}_2;\widehat{\Delta}} > F_{1-\alpha;1,\widehat{\text{df}}_2}\right)\\ &= \Pr\left(F_{1,30.6;10.98} > F_{.95;1,30.6}\right) = .89\end{aligned}$$

Recall that this estimate was computed assuming no test × reader interaction, since we have set $\widehat{\sigma}^2_{TR}$. A more conservative approach would be, for example, to set $\sigma^2_{TR} = .0001$ corresponding to the belief that a 95% upper bound on the absolute difference of two intra-reader AUC differences is given by $l = .04$. Using this approach, the power is .86.

Typically the researcher will want to consider different combinations of readers and cases that result in the desired power and then choose the most suitable combination. Reader-case sample size combinations that result in approximately .80 power are presented in the left-hand side of Table 5, using both $\sigma^2_{TR} = 0$ and $\widehat{\sigma}^2_{TR} = .0001$. For example, a few of the reader-case combinations that yield .80 power with $\widehat{\sigma}^2_{TR} = 0$ are 5 readers and 266 cases, 8 readers and 183 cases, or 15 readers and 136 cases. We see that the increase in the number of cases needed based on $\widehat{\sigma}^2_{TR} = .0001$ is most noticeable for $r \leq 5$.

Appendix C (available online at www.academicradiology.org) includes the SAS [24] statements used to compute the power for this example with $r = 8$ and $c = 240$, as well as the statements used to create the left-hand side of Table 5. To produce the Table 5 output, the program loops the statements through various combinations of reader and case sample sizes and outputs the number of cases for which the power is closest but greater than. 80 for each reader sample size. These statements can be easily modified to work in another programming language.

Situation 2: Different normal-to-abnormal ratios. Suppose that the researcher wants to use equal numbers of normal and abnormal images in the planned study in order to increase power. Since there are 45 abnormal and 69 normal cases, we randomly sample with replacement 69 abnormal cases from the original 45. We repeat this process 10 times, combining each generated sample with the 69 normal cases to produce 10 data sets, each containing 69 abnormal and 69 normal cases. Note that the normal cases are the same for each data set, in contrast to the 69 abnormal cases which vary from set to set and which do not necessarily contain all of the original 45 abnormal cases.

For each of these ten data sets we compute the jackknife covariance matrix and then compute $\widehat{\sigma}^2_{\varepsilon}$, $\widehat{\text{Cov}}_1$, $\widehat{\text{Cov}}_2$ and $\widehat{\text{Cov}}_3$. These estimates are shown in Table 6 along with the corresponding means. We use the estimate $\widehat{\sigma}^2_{TR} = 0$ based on the original pilot data before doing any resampling, and well as the more conservative conjectured estimate $\widehat{\sigma}^2_{TR} = .0001$. Using $\widehat{\sigma}^2_{TR} = 0$ and the means from Table 6 as inputs in our power program, with $c^* = 69 + 69 = 138$, we find for $r = 8$ and $c = 240$ that the power has now increased from .89 to .98, showing the advantage of using a normal-to-abnormal ratio equal to 1. The right-hand side of Table 5 shows combinations of readers and case sample sizes that yield .80 power for an equal balance of normal and abnormal cases.

3.1.1. Power computation based on DBM analysis—The DBM analysis of the data based on the PROPROC AUC estimates is presented in Table 7. Part (a) presents the DBM ANOVA table, part (b) the F statistic, part (c) ddf$_H$, and part (d) the p-value. Note that the F statistic, ddf$_H$, and the p-value are the same as for the OR analysis in Table 4; this will always be the case when the OR analysis uses jackknife covariance estimates, as previously discussed. Using the Table 1 relationships, we compute the corresponding OR quantities $\overline{\text{MS}}(T)$, $\overline{\text{MS}}(T * R)$, $\widehat{\sigma}^2_{\epsilon}$, $\widehat{\text{Cov}}_1$, $\widehat{\text{Cov}}_2$, $\widehat{\text{Cov}}_3$ for step 2a from the DBM mean squares. Otherwise the steps are identical. Appendix D (available online at www.academicradiology.org) includes SAS statements that convert the DBM mean squares to the corresponding OR quantities for this example, based on the Table 1 relationships. The SAS output included in Appendix D shows that the resulting OR quantities are the same as those obtained from the OR analysis; thus power results are identical regardless of whether we use the OR or DBM analysis outputs – this will always be the case when the OR analysis uses jackknife covariance estimates and the DBM analysis uses normalized pseudovalues.

3.2. Simulation study

In a simulation study we examine the performance of the proposed power procedure. We use the simulation model of Roe and Metz [25], which provides continuous decision-variable outcomes generated from a binormal model that treats both case and reader as random factors. We use their "HH" model for which the decision-variable values have relatively high within-reader correlations and reader variability (both pure reader and test × reader interaction variance components). We specify the separation between the normal and abnormal case populations such that for one test the median AUC across readers is .855 and for the other test it is .92, resulting in a nominal effect size of .065. Using this model, we simulate 4000 samples for each of nine combinations of three reader-sample sizes (readers = 3, 5, and 10) and three case-sample sizes (cases = 50, 100, and 200) with equal numbers of normal and abnormal cases. Within each simulation, all Monte Carlo readers read the same cases for each of the two tests. For these simulations we set $\widehat{\sigma}^2_{TR} = 0$ if it is negative.

For each sample we perform an OR analysis using the empirical AUC as the accuracy estimate. The mean values of the parameter estimates and AUC differences are displayed in Table 8. The "Power" column indicates the proportion of samples where the null hypothesis of equal tests was rejected at alpha = .05. We make the following observations: (1) The mean σ^2_{TR} values are very similar (range: 1.20 – 1.33) regardless of the number of readers or cases; this is expected, since σ^2_{TR} can be interpreted as the interaction variance component for the latent AUCs, as discussed in Section 2.5.1. (2) The correlations are also very similar (e.g., range of r_1: .36 – .38) across combinations as expected. (3) The covariances and error variance decrease as the number of cases increases, but are similar for similar case sample sizes regardless of the reader sample size. (4) The mean AUC difference is .066, except for one combination; note that this differs from the .065 median AUC difference for the decision variable; (5) The fact that r_2 is roughly a third larger than r_1 should not be taken as evidence that the constraint given by Equation 4 is not realistic, but rather that the simulation model does not properly reflect the typical clinical situation.

We now investigate how well the sample data predict power for a planned study with 10 readers and 200 cases for an effect size of .066. From the last line in Table 8 we estimate the true power to be approximately 0.781, based on 4000 simulated data sets. For the power procedure to be valid, it should give power estimates close to the true power when reliable estimates are available. For each combination we compute the power using the parameter estimates from Table 8. The results are displayed in the "Reliable estimates" column in Table 9. We see that, with the exception of the first combination (3 readers, 50 cases), the power estimated from the reliable estimates is within .039 of the actual power and the mean of these estimates is .744, thus validating the power procedure. Note that this estimate of power is performed only once for each combination using the reliable parameter estimates. The means for the sample power estimates (computed for each sample based on the sample parameter estimates) across the 4000 samples are presented in the "Sample estimates" column; these are closer, within .021 of the actual power, and have an overall mean of .767. The 25th and 75th percentiles and their differences for the sample power estimate distributions are presented in the last three columns. We see, for example, that the middle 50% of the sample power estimates has, on average, a range of .252.

4. Discussion

We have provided a step-by-step procedure for estimating power for planned multireader ROC studies that will be analyzed using either the DBM or OR methods. This procedure updates previous approaches by using the currently recommended denominator degrees of freedom, accounting for Different pilot- and planned-study normal-to-abnormal case ratios, and using a new method for computing the OR test-by-reader variance component.

This procedure, as is true for most power procedures, was derived with the parameter values treated as known. A small simulation study validated the method by showing that power estimates were quite close to the actual power when computed from reliable parameter estimates. In addition, the means of sample power estimates – those based on sample-specific parameter estimates – were even closer to the actual power. However, we emphasize that this was a small simulation study based on only one latent decision-variable model, and that more extensive simulation studies are needed to more fully validate the procedure.

Variability in power estimates increases as the parameter estimates become less precise for any power procedure. Thus it is to be expected that there will be much variability in sample power estimates based on outputs from the typical pilot study that has only a few readers, due to lack of precision for the test × reader variance component estimate. In our simulation study the middle 50% of the sample power estimates had, on average, a range of .252, which we would prefer to be less. A recent simulation investigation [26] of an earlier version of the DBM power method has also noted large variability in sample power estimates. However, variability is probably much less than indicated by simulations when the same readers are used in both the pilot and future study, as is often the case. Nevertheless, the variability issue warrants further investigation. For example, one possible way to reduce the variability would be to use a conjectured value for the test × reader variance component when feasible.

Finally, we note that the pilot study should be comparable *to* the planned study with respect to modalities, reader expertise, and selection of cases in order that the parameter estimates obtained will accurately estimate those of the planned study.

Acknowledgments

We thank Carolyn Van Dyke, MD for sharing her data set for the example. We thank the reviewer for helpful suggestions that clarified the presentation.

Appendix A: Power derivation for the OR procedure

As previously noted, the OR procedure test statistic (Eq. 5) has an approximate $F_{t-1,\mathrm{df}_2;\Delta}$ distribution with df_2 and noncentrality parameter Δ given by Equations 6 and 8, respectively. For t = 2 tests it follows that

$$\Delta = \frac{\frac{r}{2}(\theta_1 - \theta_2)^2}{\sigma^2_{TR} + \sigma^2_{\varepsilon} - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)} \tag{A.1}$$

and

$$\mathrm{df}_2=\frac{[E\,(\mathrm{MS}\,(T*R))+r\,(\mathrm{Cov}_2-\mathrm{Cov}_3)]^2}{\frac{[E(\mathrm{MS}(T*R))]^2}{r-1}} \tag{A.2}$$

It is shown in Reference [8] that

$$E\,(\mathrm{MS}\,(T*R))=\sigma^2_{TR}+\sigma^2_{\varepsilon}-\mathrm{Cov}_1-(\mathrm{Cov}_2-\mathrm{Cov}_3) \tag{A.3}$$

It follows from Equations A.2–A.3 that

$$\sigma^2_{TR}=E\,(\mathrm{MS}\,(T*R))-\sigma^2_{\varepsilon}+\mathrm{Cov}_1+(\mathrm{Cov}_2-\mathrm{Cov}_3) \tag{A.4}$$

and

$$\mathrm{df}_2=\frac{\left[\sigma^2_{TR}+\sigma^2_{\varepsilon}-\mathrm{Cov}_1+(r-1)\,(\mathrm{Cov}_2-\mathrm{Cov}_3)\right]^2}{\frac{\left[\sigma^2_{TR}+\sigma^2_{\varepsilon}-\mathrm{Cov}_1-(\mathrm{Cov}_2-\mathrm{Cov}_3)\right]^2}{r-1}} \tag{A.5}$$

Let *r** and *c** denote reader and case pilot-study sample sizes from which covariance parameter estimates are obtained and *r* and *c* the corresponding sample sizes for which we want to compute power. Based on Equations A.1, A.4 and A.5 we use the following estimates that incorporate the constraint $\mathrm{Cov}_2 \geq \mathrm{Cov}_3$:

$$\begin{aligned}&\widehat{\sigma}^2_{TR}=\mathrm{MS}\,(T*R)-\sigma^2_{\varepsilon}+\widehat{\mathrm{Cov}}_1+H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\\&\widehat{\Delta}=\frac{\frac{r}{2}(\theta_1-\theta_2)^2}{\widehat{\sigma}^2_{TR}+\frac{c^*}{c}\left\{\widehat{\sigma}^2_{\varepsilon}-\widehat{\mathrm{Cov}}_1+(r-1)H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right\}}\end{aligned}$$

and

$$\widehat{\mathrm{df}}_2=\frac{\left\{\widehat{\sigma}^2_{TR}+\frac{c^*}{c}\left[\widehat{\sigma}^2_{\varepsilon}-\widehat{\mathrm{Cov}}_1+(r-1)\,H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right]\right\}^2}{\frac{\left\{\widehat{\sigma}^2_{TR}+\left(\frac{c^*}{c}\right)\left[\widehat{\sigma}^2_{\varepsilon}-\widehat{\mathrm{Cov}}_1-H\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right]\right\}^2}{r-1}}$$

In deriving these estimates we make the reasonable assumption that σ^2_{ε}, $\mathrm{Cov1}_1$, Cov_2, and Cov_3 are inversely proportional to the number of cases for a specified normal-to-abnormal case ratio. Note that if $\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\leq 0$, then

$$\widehat{\mathrm{df}}_2=\frac{\left[\widehat{\sigma}^2_{TR}+\left(\frac{c*}{c}\right)\left(\widehat{\sigma}^2_{\varepsilon}-\widehat{\mathrm{Cov}1}_1\right)\right]^2}{\frac{\left[\widehat{\sigma}^2_{TR}+\left(\frac{c*}{c}\right)\left(\widehat{\sigma}^2_{\varepsilon}-\widehat{\mathrm{Cov}}_1\right)\right]^2}{r-1}}=r-1$$

with $r-1$ being the lower bound on the denominator degrees of freedom.

The power is then estimated by

$$\text{power}=\Pr\left(F_{1,\widehat{\text{df}}_2;\widehat{\Delta}}>F_{1-\alpha;1,\widehat{\text{df}}_2}\right)$$

for a two-sided test with significance level α, treating $\widehat{\text{df}}_2$ and $\widehat{\Delta}$ as constants.

Appendix B: Determination of the parameter estimated by the previously used estimate of σTR2 for the OR procedure

We assume that the pilot data have two tests (t = 2). The estimate for σ^2_{TR} proposed in References [4, 6] is given by

$$\widehat{\sigma}^2_{TR\_O}=\frac{\left(\widehat{\sigma}_1^2+\widehat{\sigma}_2^2\right)}{2}\left[1-\widehat{\rho}\right] \tag{B.1}$$

where σ_1^2 and σ_2^2 are the sample variances of the AUCs and across readers for tests 1 and 2, respectively, and $\widehat{\rho}$ is the within-reader correlation coefficient for the paired data $\left(\widehat{\theta}_{1j},\widehat{\theta}_{2j}\right)$, j = 1,…,r; i.e.,

$$\widehat{\sigma}_i^2=\frac{\Sigma_{j=1}^{r}\left(\widehat{\theta}_{ij}-\widehat{\theta}_{i.}\right)^2}{r-1},\ i=1,2$$

and

$$\widehat{\rho}=\frac{\Sigma_{j=1}^{r}\left(\widehat{\theta}_{1j}-\widehat{\theta}_{1.}\right)\left(\widehat{\theta}_{2j}-\widehat{\theta}_{2.}\right)/\left(r-1\right)}{\sqrt{\widehat{\sigma}_1^2}\sqrt{\widehat{\sigma}_2^2}} \tag{B.2}$$

From the OR model (Eqn. 3) it follows that $\text{var}\left(\widehat{\theta}_{ij}\right)=\sigma^2_R+\sigma^2_{TR}+\sigma^2_{\varepsilon}$ and $\underset{j\neq j'}{\text{cov}}\left(\widehat{\theta}_{ij},\widehat{\theta}'_{ij}\right)=\text{Cov}_2$; it follows that

$$E\left[\frac{\Sigma_{j=1}^{r}\left(\widehat{\theta}_{1j}-\widehat{\theta}_{1.}\right)^2}{r-1}\right]=\text{var}\left(\widehat{\theta}_{ij}\right)-\text{Cov}_2$$

i.e.,

$$E\left(\widehat{\sigma}_i^2\right)=\sigma^2_R+\sigma^2_{TR}+\sigma^2_{\varepsilon}-\text{Cov}_2 \tag{B.3}$$

Furthermore, we show at the end of this section that

$$\frac{\Sigma_{j=1}^{r}\left(\widehat{\theta}_{1j}-\widehat{\theta}_{1\cdot}\right)\left(\widehat{\theta}_{2j}-\widehat{\theta}_{2\cdot}\right)}{r-1}=\frac{\left[\text{MS}\left(R\right)-\text{MS}\left(T*R\right)\right]}{t} \tag{B.4}$$

where MS(R) and MS($T * R$) are the reader and test × reader mean squares resulting from fitting the OR model to the pilot data.

Expectations for the OR mean squares are given by Hillis [8, p. 600]. From these it follows that

$$E\left[\mathrm{MS}\left(R\right)-\mathrm{MS}\left(T*R\right)\right]/t=\sigma_R^2+\mathrm{Cov}_1-\mathrm{Cov}_3 \tag{B.5}$$

From Equations B.4–B.5 it follows that

$$E\left[\frac{\Sigma_{j=1}^r\left(\widehat{\theta}_{1j}-\widehat{\theta}_{1\cdot}\right)\left(\widehat{\theta}_{2j}-\widehat{\theta}_{2\cdot}\right)}{r-1}\right]=\sigma_R^2+\mathrm{Cov}_1-\mathrm{Cov}_3 \tag{B.6}$$

Replacing estimates by their expected value in Equation B.1 using Equations B.2, B.3, and B.6 shows that $\widehat{\sigma}_{TR-O}^2$ estimates the following parameter:

$$\left[\sigma_R^2+\sigma_{TR}^2+\sigma_\varepsilon^2-\mathrm{Cov}_2\right]\left[1-\frac{\left(\sigma_R^2+\mathrm{Cov}_1-\mathrm{Cov}_3\right)}{\sigma_R^2+\sigma_{TR}^2+\sigma_\varepsilon^2-\mathrm{Cov}_2}\right]=\sigma_R^2+\sigma_{TR}^2+\sigma_\varepsilon^2-\mathrm{Cov}_1-(\mathrm{Cov}_2-\mathrm{Cov}_3)$$

The relationship var($\varepsilon_{11}-\varepsilon_{12}-\varepsilon_{21}+\varepsilon_{22}$) ≥ 0 implies that $\sigma_\varepsilon^2-\mathrm{Cov}_1-(\mathrm{Cov}_2-\mathrm{Cov}_3)\geq 0$.

Proof of Equation B.4:

$$\begin{aligned}\mathrm{MS}\left(R\right)-\mathrm{MS}\left(TR\right)&=\frac{t\Sigma_{j=1}^r\left(\widehat{\theta}_{\cdot j}-\widehat{\theta}_{\cdot\cdot}\right)^2}{r-1}-\frac{\Sigma_{i=1}^t\Sigma_{j=1}^r\left(\widehat{\theta}_{ij}-\widehat{\theta}_{i\cdot}+\widehat{\theta}_{\cdot j}-\widehat{\theta}_{\cdot\cdot}\right)^2}{(t-1)(r-1)}\\&=\frac{\left[\left(t\Sigma_{j=1}^r\widehat{\theta}_{\cdot j}^2-tr\widehat{\theta}_{\cdot\cdot}^2\right)-\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2-r\Sigma_{i=1}^t\widehat{\theta}_{i\cdot}^2-t\Sigma_{j=1}^r\widehat{\theta}_{\cdot j}^2+tr\widehat{\theta}_{\cdot\cdot}^2\right)\right]}{r-1}\end{aligned}$$

Case 1: $\widehat{\theta}_{1\cdot}=\widehat{\theta}_{2\cdot}=0$ (hence $\widehat{\theta}_{\cdot\cdot}=0$). It follows that

$$\begin{aligned}\frac{\mathrm{MS}(R)-\mathrm{MS}(TR)}{t}&=\frac{1}{t}\frac{\left(t\Sigma_{j=1}^r\widehat{\theta}_{\cdot j}^2\right)-\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2-t\Sigma_{j=1}^r\widehat{\theta}_{\cdot j}^2\right)}{r-1}\\&=\frac{1}{t}\frac{\left(2t\Sigma_{j=1}^r\widehat{\theta}_{\cdot j}^2\right)-\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2\right)}{r-1}\\&=\frac{1}{t}\frac{\left(2t\Sigma_{j=1}^r\left(\frac{\widehat{\theta}_{1j}+\widehat{\theta}_{2j}}{t}\right)^2\right)-\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2\right)}{r-1}\\&=\frac{1}{t}\frac{\left[\frac{2}{t}\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2+2\quad\Sigma_{j=1}^r\widehat{\theta}_{1j}+\widehat{\theta}_{2j}\right)-\left(\Sigma_{i=1}^t\Sigma_{j=1}^r\widehat{\theta}_{ij}^2\right)\right]}{r-1}\end{aligned}$$

Since t = 2 we have

$$\frac{\mathrm{MS}\left(R\right)-\mathrm{MS}\left(T*R\right)}{t}=\frac{\Sigma_{j=1}^r\widehat{\theta}_{1j}\widehat{\theta}_{2j}}{r-1}$$

and thus Equation B.4 holds for the $\widehat{\theta}_{ij}$.

Case 2: $\widehat{\theta}_{1\cdot} \neq 0$ or $\widehat{\theta}_{2\cdot} \neq 0$. Define $W_{ij} = \widehat{\theta}_{ij} - \widehat{\theta}_{i\cdot}$. Since $W_{1\cdot} = W_{2\cdot} = 0$, then Equation B.4 holds for the W_{ij}. Since it can be shown that $\sum_{j=1}^{r} \left(\widehat{\theta}_{1j} - \widehat{\theta}_{1\cdot}\right)\left(\widehat{\theta}_{2j} - \widehat{\theta}_{2\cdot}\right) = \sum_{j=1}^{r} \left(W_{1j} - W_{1\cdot}\right)\left(W_{2j} - W_{2\cdot}\right)$ and the quantities MS(R), and MS(T * R) computed from the W_{ij} are identical to those computed from the $\widehat{\theta}_{ij}$, then Equation B.4 must also hold for the $\widehat{\theta}_{ij}$.

Appendix C: SAS statements for computing power for the example

a) Computation of power for *r* = 8 readers and *c* = 240 cases

```
data data1; **Enter the OR outputs computed from pilot data**; length study $16;

input study $ c_star mstr var_error cov1 cov2 cov3 var_tr;/*Notes:

mstr = MS(test × reader)

var_tr = OR test × reader variance component

var_error = OR fixed-reader error variance component

c_star = number of cases for pilot data

Either mstr or var_tr must be specified--enter a missing value for the one not specified. If
var_tr is not missing then the program uses the specified var_tr value, regardless of whether
mstr is specified or missing. If var_tr is missing then var_tr is computed as a function of
mstr and other inputs, and if the computed value is negative then it will be set to zero.

*/

cards;

VanDyke 114 .000622731 .001393652 .000351859 .000346505 .000221453 . ; /*

NOTE: to obtain result with the test-by-reader variance component set to .0001, just change
the missing data value above to .0001. That is, substitute the following line:

VanDyke 114 .000622731 .001393652 .000351859 .000346505 .000221453 .0001 */

proc print; title “Pilot study estimates”; run;

data data2; set data1; **Compute power for r = 8, c = 240**;

/* set the following as desired */

alpha = .05; **significance level**;

AUCdiff = .05; **effect size: difference in populations AUCs**;

r = 8; **reader sample size for power estimate**;

c = 240; **case sample size for power estimate**;

/* now estimate var_tr if it was not specified*/
```

```
if var_tr = . then do;

var_tr = mstr − var error + cov1 + max(cov2 − cov3,0);

var_tr = var_tr*(var_tr>0); *constrains var_tr to be nonnegative*; end;

/* now estimate noncentrality parameter (nc) and denominator df (df2)*/

denon = var_tr + (c_star/c)*(var_error−cov1+max((r−1)*(cov2−cov3),0));

nc = r*.5 * AUCdiff**2/denon;

df2 = denon**2/((var_tr+(c_star/c)*(var_error−cov1−max(cov2−cov3,0)))**2/(r−1));

/* now compute power */

F_critical = Finv(1−alpha, 1,df2);

power = 1 −probF(F_critical,1, df2, nc);

proc print; title “Power results”;

var study AUCdiff r c nc df2 power;

run;
```

Output:

```
Pilot study estimates
   study     c_star      mstr        var_error      cov1          cov2          cov3       var_tr
VanDyke       114     .000622731   .001393652   .000351859   .000346505   .000221453       .

Power results
study      AUCdiff    r    c       nc        df2       power
VanDyke       0.05    8   240   10.9812   30.6140   0.89402
```

(b) Computation of reader and case sample sizes needed for power = .80. These results are presented in the left-hand side of Table 3

```
***looped version***;

data data2; set data1;

/* set the following as desired */

alpha = .05; **significance level**;

AUCdiff = .05; **effect size: difference in populations AUCs**;

power_target = .80; **desired power**;

/* now estimate var_tr if it was not specified */

if var_tr = . then do;
```

```
var_tr = mstr − var_error + cov1 + max(cov2 − cov3,0);

var_tr = var_tr*(var_tr>0); *constrains var_tr to be nonnegative*; end;

do r = 3 to 15; **reader sample size for power estimate**;

flag = 0;

do c = 20 to 2000; **candidate case sample sizes for power estimate --change as needed**;

/* now estimate noncentrality parameter(nc)and denominator df (df2)*/

denon = var_tr + (c_star/c)*(var_error − cov1 + max((r−1)*(cov2 − cov3),0));

nc = r*.5 * AUCdiff**2/denon; **nc = noncentrality parameter**;

df2 = denon**2/((var_tr + (c_star/c)*(var_error−cov1−max(cov2−cov3,0)))**2/(r−1));

/* now compute power */

F_critical = Finv(1−alpha, 1,df2); **F_critical = OR critical F value**;

power = 1 −probF(F_critical,1, df2, nc);

if (flag = 0) and (power ge power_target) then do;

output; flag = 1; GOTO HERE;

end;

end;

HERE:;

end;

proc print;

var study AUCdiff r c power; run;
```

Output:

```
Power results
Obs    study     AUCdiff     r     c     power
  1    VanDyke     0.05      3   559   0.80044
  2    VanDyke     0.05      4   343   0.80040
  3    VanDyke     0.05      5   266   0.80142
  4    VanDyke     0.05      6   225   0.80045
  5    VanDyke     0.05      7   200   0.80020
  6    VanDyke     0.05      8   183   0.80007
  7    VanDyke     0.05      9   171   0.80079
  8    VanDyke     0.05     10   162   0.80175
  9    VanDyke     0.05     11   154   0.80028
 10    VanDyke     0.05     12   148   0.80025
```

11	VanDyke	0.05	13	143	0.80010
12	VanDyke	0.05	14	139	0.80055
13	VanDyke	0.05	15	136	0.80214

Appendix D: SAS statements for converting DBM mean squares to OR statistics for the example

```
data OR_statistics;

input t r c mst msr mstr msc mstc msrc mstrc; **DBM mean squares**;

/* Notes:

t, r, and c are number of tests, readers and cases for the data set mst, msr, mstr, msc, mstc,
msrc, and mstrc are the DBM mean squares for test, reader, test × reader, case, test × case,
reader × case, and test × reader × case

*/

/*Now compute corresponding OR mean squares and fixed-reader covariances*/

mst_OR = c**−1 * mst;

msr_OR = c** −1 * msr;

mstr_OR = c**−1 * mstr;

var_error = (t*r*c)**−1 * (msc + (t−1)*mstc + (r−1)*msrc + (t−1)*(r−1)*mstrc);

cov1 = (t*r*c)**−1 *(msc − mstc +(r−1)*(msrc − mstrc));

cov2 = (t*r*c)**−1 * (msc − msrc + (t−1)*(mstc − mstrc));

cov3 = (t*r*c)**−1 * (msc − mstc − msrc + mstrc);

cards;

2 5 114 0.45638557 0.32315642 0.07099138 0.45797697 0.17578816 0.13424103
0.10450847

proc print; title “DBM mean squares”;

var t r c mst msr mstr msc mstc msrc mstrc;

proc print; title “Corresponding OR mean squares and covariances”;

var msr_OR mstr_OR var_error cov1 cov2 cov3;

run;
```

Output:

```
DBM mean squares
t   r   c      mst       msr       mstr       msc       mstc     msrc      mstrc
2   5   114    0.45639   0.32316   0.070991   0.45798   0.4579   0.13424   0.10451
```

```
Corresponding OR mean squares and covariances

    msr_OR      mstr_OR    var_error         cov1          cov2          cov3

 .002834705  .000622731  .001393652  .000351859  .000346505  .000221453
```

References

[1]. 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]

[2]. 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]

[3]. 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.

[4]. 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]

[5]. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Statistics in Medicine 1997;16:1529–1542. [PubMed: 9249923]

[6]. Zhou, X-H.; Obuchowski, NA.; McClish, DK. Statistical methods in diagnostic medicine. Wiley; New York: 2002.

[7]. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Academic Radiology 2004;11:1260–1273. DOI:10.1016/j.acra.2004.08.009. [PubMed: 15561573]

[8]. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine 2007;26:596–619. DOI:10.1002/sim.2532. [PubMed: 16538699]

[9]. 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. DOI:10.1002/sim.2024. [PubMed: 15685718]

[10]. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Academic Radiology 2008;15:647–61. [PubMed: 18423323]

[11]. Quenoille MH. Approximate tests of correlation in time series. Journal of the Royal Statistical Society 1949;11:68–84.Series B

[12]. Quenoille MH. Notes on bias in estimation. Biometrika 1956;43:353–360.

[13]. Tukey JW. Bias and confidence in not quite large samples. Annals of Mathematical Statistics 1958;29:614. abstract.

[14]. Berbaum, KS.; Schartz, KM.; Pesce, LL.; Hillis, SL. DBM MRMC 2.2. (computer software). Available for download from http://perception.radiology.uiowa.edu. Accessed August 1, 2009

[15]. Berbaum, KS.; Metz, CE.; Pesce, LL.; Schartz, KM. DBM MRMC 2.1 User's Guide. (software manual). Available for download from http://perception.radiology.uiowa.edu. Accessed August 1, 2009

[16]. Hillis, SL.; Schartz, KM.; Pesce, LL.; Berbaum, KS.; Metz, CE. DBM MRMC procedure for SAS. (computer software). Available for download from http://perception.radiology.uiowa.edu. Accessed August 1, 2009

[17]. 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]

[18]. 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]

[19]. Obuchowski NA. Computing sample size for receiver operating characteristic studies. Investigative Radiology 1994;29:238–243. [PubMed: 8169102]

[20]. Hillis, SL.; Berbaum, KS. MRMC sample size program user's guide. (software manual). Available for download from http://perception.radiology.uiowa.edu. Accessed August 1, 2009

[21]. Van Dyke, CW.; White, RD.; Obuchowski, NA.; Geisinger, MA.; Lorig, RJ.; Meziane, MA. Cine MRI in the diagnosis of thoracic aortic dissection. 79th RSNA Meetings; Chicago, IL. November 28 - December 3, 1993;

[22]. Pan XC, Metz CE. The ”proper” binormal model: parametric receiver operating characteristic curve estimation with degenerate data. Academic Radiology 1997;4:380–389. [PubMed: 9156236]

[23]. Metz CE, Pan XC. ”Proper” binormal ROC curves: theory and maximum-likelihood estimation. Journal of Mathematical Psychology 1999;43:1–33. [PubMed: 10069933]

[24]. SAS for Windows. Version 9.2. SAS Institute Inc.; Cary, NC, USA: copyright (c) 2002–2008

[25]. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. Academic Radiology 1997;4:298–303. [PubMed: 9110028]

[26]. Chakraborty DP. Prediction accuracy of a sample-size estimation method for ROC Studies. Academic Radiology 2010;17:628–638. [PubMed: 20380980]

Table 1

OR outputs in terms of DBM mean squares from a pilot study. The number of tests, readers, and cases in the study are denoted by *t**, *r**, and *c**, respectively. Adapted and reprinted, with permission, from Hillis et al [10].

OR Output	Equivalent function of DBM mean squares
$\mathrm{MS}(T)_{\hat{\theta}_{ij}}$	$= \frac{1}{c^{*}}\mathrm{MS}(T)$
$\mathrm{MS}(R)_{\hat{\theta}_{ij}}$	$= \frac{1}{c^{*}}\mathrm{MS}(T)$
$\mathrm{MS}(T * R)_{\hat{\theta}_{ij}}$	$= \frac{1}{c^{*}}\mathrm{MS}(T * R)$
$\hat{\sigma}_{\epsilon}^{2}$	$= \left[\frac{1}{t^{*}r^{*}c^{*}}\mathrm{MS}(C) - \left(t^{*} - 1\right)\mathrm{MS}(T * C) + \left(r^{*} - 1\right)\mathrm{MS}(R * C) + \left(t^{*} - 1\right)\left(r^{*} - 1\right)\mathrm{MS}(T * R * C)\right]$
$\widehat{\mathrm{Cov}}_{1}$	$= \frac{1}{t^{*}r^{*}c^{*}}\left[\mathrm{MS}(C) - \mathrm{MS}(T * C) + \left(r^{*} - 1\right)\mathrm{MS}(R * C) - \mathrm{MS}(T * R * C)\right]$
$\widehat{\mathrm{Cov}}_{2}$	$= \frac{1}{t^{*}r^{*}c^{*}}\left[\mathrm{MS}(C) - \mathrm{MS}(R * C) + \left(t^{*} - 1\right)\mathrm{MS}(T * C) - \mathrm{MS}(T * R * C)\right]$
$\widehat{\mathrm{Cov}}_{3}$	$= \frac{1}{t^{*}r^{*}c^{*}}\left[\mathrm{MS}(C) - \mathrm{MS}(T * C) - \mathrm{MS}(R * C) + \mathrm{MS}(T * R * C)\right]$

Table 2

Relationships between OR and DBM variance component and covariance parameters. Notes: c is the number of cases; see Reference [9] for definitions of the DBM variance components. Adapted and reprinted, with permission, from Hillis et al [9, Table III].

OR parameter	Equivalent function of DBM variance components
σ_R^2	$= \sigma_R^2$
σ_{TR}^2	$= \sigma_{TR}^2$
σ_ϵ^2	$= \left(\sigma_C^2 + \sigma_{TC}^2 + \sigma_{RC}^2 + \sigma_{TRC}^2 + \sigma_\epsilon^2\right) / c$
Cov_1	$= \left(\sigma_C^2 + \sigma_{RC}^2\right) / c$
Cov_2	$= \left(\sigma_C^2 + \sigma_{TC}^2\right) / c$
Cov_3	$= \sigma_C^2 / c$

Table 3

Relationship between OR test-by-reader interaction variance component σ^2_{TR} and 95% probability upper bound l on the absolute difference of intra-reader latent AUCs; i.e., Pr $\{|\eta_{ij} - \eta_{i'j} - (\eta_{ij'} - \eta_{i'j'})| \leq l\} \geq .95$, where $\eta_{ij} - \eta_{i'j}$ is the difference in latent AUCs for tests i and i' for randomly chosen reader j and $\eta_{ij} - \eta_{i'j}$ is the corresponding difference for randomly chosen reader j'.

l	σ^2_{TR}
0.01	0.00001
0.02	0.00003
0.03	0.00006
0.04	0.00010
0.05	0.00016
0.06	0.00023
0.07	0.00032
0.08	0.00042
0.09	0.00053
0.1	0.00065

Table 4

Obuchowski-Rockette analysis of Van Dyke et al [21] data. H_0: test AUCs ar equal; $t = 2$ tests; $r = 5$ readers.

(a) PROPROC AUC estimates for cine and spin-echo MRI

	Test	
Reader	**Cine**	**Spin-echo**
1	.934	.952
2	.891	.926
3	.908	.930
4	.977	1.000
5	.841	.943
Mean:	.910	.950

(b) ANOVA table based on PROPROC AUCs

Source	**df**	**Mean square**
T	1	0.004003382
R	4	0.002834705
T*R	4	0.000622731

(c) Jackknife covariance matrix corresponding to PROPROC AUC estimates. C1–C5 = readers 1–5, cine; S1–S5 = readers 1–5, spin echo. Values have been multiplied by 10^4.

	C1	**C2**	**C3**	**C4**	**C5**	**S1**	**S2**	**S3**	**S4**	**S5**
C1	9.54									
C2	7.47	20.35								
C3	8.73	6.64	61.78							
C4	2.24	2.65	1.70	1.48						
C5	5.48	12.26	3.11	2.00	18.07					
S1	3.93	4.26	3.67	0.37	2.62	5.19				
S2	3.28	5.50	3.26	1.07	4.70	2.46	4.94			
S3	4.74	5.59	5.53	1.23	4.40	5.03	3.95	8.03		
S4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
S5	0.85	0.35	3.82	0.07	2.63	0.40	2.98	2.19	0.00	10.00

(d) Covariance estimates

$\hat{\sigma}_{\epsilon}^{2} = .001393652; \widehat{\mathrm{Cov}}_{1} = .000351859; \widehat{\mathrm{Cov}}_{2} = .000346505; \widehat{\mathrm{Cov}}_{3} = .000221453$

(e) Correlation estimates ($r_i = \widehat{\mathrm{Cov}}_i \,/\, \hat{\sigma}_{\varepsilon}^{2}$)

$r_1 = 0.25247$; $r_2 = 0.24863$; $r_3 = 0.15890$

(f) $$F = \frac{\mathrm{MS}(T)}{\mathrm{MS}(T*R) + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right),\ 0\right]} = \frac{.004003382}{.000622731 + 5(000346505 - .000221453)} = 3.21$$

(g) $$\mathrm{ddf}_H = \frac{\left\{\mathrm{MS}(T*R) + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right),\ 0\right]\right\}^2}{\left[\mathrm{MS}(T*R)\right] / \left[(t-1)(r-1)\right]} = 16.065$$

(h) p-value= Pr ($F_{1;16.065} > 3.21$) = .092

(i) 95% CI for $\theta_2 - \theta_1$: $.04 \pm t_{0.25;16.065}\sqrt{\frac{2}{5}\mathrm{MSDen}_{\mathrm{OR}}} = (-0.0073,\ 0.0921)$

Table 5

Combinations of cases and readers having .80 power to detect a .05 AUC difference between spin-echo and cine MRI based on the Van Dyke et al [21] data. The pilot-study estimate of $\widehat{\sigma}^2_{TR}$ was negative. The left-hand-side results are for a planned study that has the same normal-to-abnormal ratio as the pilot study; the right-hand-side results are for a planned study that has equal numbers of normal and abnormal cases.

	normal/abnormal ratio = 69/45 = 1.53				normal/abnormal ratio = 1			
	$\hat{\sigma}^2_{TR} = 0$		$\hat{\sigma}^2_{TR} = .0001$		$\hat{\sigma}^2_{TR} = 0$		$\hat{\sigma}^2_{TR} = .0001$	
Readers	**Cases**	**Power**	**Cases**	**Power**	**Cases**	**Power**	**Cases**	**Power**
3	559	0.800	1898	0.800	374	0.800	1282	0.800
4	343	0.800	491	0.800	229	0.800	328	0.800
5	266	0.801	330	0.801	177	0.801	220	0.801
6	225	0.800	263	0.800	150	0.801	176	0.802
7	200	0.800	227	0.801	133	0.800	151	0.801
8	183	0.800	203	0.801	122	0.801	135	0.801
9	171	0.801	187	0.802	114	0.802	124	0.801
10	162	0.802	174	0.800	108	0.803	116	0.802
11	154	0.800	165	0.801	103	0.803	110	0.803
12	148	0.800	158	0.802	99	0.803	105	0.803
13	143	0.800	151	0.800	95	0.801	101	0.803
14	139	0.801	146	0.800	93	0.804	97	0.801
15	136	0.802	142	0.801	90	0.801	94	0.800

Table 6

OR variance component estimates from ten randomly generated Van Dyke [20] data sets having 69 normal and 69 abnormal images. Each data set contains 69 resampled abnormal images combined with the original 69 normal images.

Sample	$\hat{\sigma}_{\varepsilon}^{2}$	$\widehat{\mathrm{Cov}}_1$	$\widehat{\mathrm{Cov}}_2$	$\widehat{\mathrm{Cov}}_3$
1	0.000512	0.000204	0.000181	0.000125
2	0.000416	0.000019	0.000112	0.000073
3	0.001173	0.000118	0.000169	0.000138
4	0.001121	0.000078	0.000129	0.000074
5	0.000545	0.000153	0.000242	0.000106
6	0.000629	0.000316	0.000224	0.000189
7	0.000634	0.000155	0.000225	0.000107
8	0.001117	0.000135	0.000204	0.000116
9	0.000608	0.000176	0.000222	0.000145
10	0.000470	0.000130	0.000136	0.000089
mean:	0.000723	0.000148	0.000184	0.000116

Table 7

Dorfman-Berbaum-Metz (DBM) analysis of Van Dyke et al [21] data. H_0: test AUCs are equal; t = 2 tests; r = 5 readers.

(a) ANOVA table based on normalized jackknife AUC pseudovalues			
Source	**df**	**SS**	**MS**
T	1	0.45638557	0.45638557
R	4	1.29262569	0.32315642
T*R	4	0.28396550	0.07099138
C	113	51.75139760	0.45797697
T*C	113	19.86406163	0.17578816
R*C	452	60.67694615	0.13424103
T*R*C	452	47.23783039	0.10450847

(b) $$F = \frac{\mathrm{MS}(T)}{\mathrm{Ms}(T*R) + \max[\mathrm{MS}(T*C) - \mathrm{MS}(T*R*C),\ 0]} = \frac{0.45638557}{0.07099138 + 0.17578816 - 0.10450847} = 3.21$$

(c) $$\mathrm{ddf}_H = \frac{\{\mathrm{Ms}(T*R) + \max[\mathrm{MS}(T*C) - \mathrm{MS}(T*R*C),\ 0]\}^2}{[\mathrm{MS}(T*R)]^2 / [(t-1)(r-1)]} = 16.065$$

(d) p-value= Pr ($F_{4;16.065} > 3.21$) = .092

Table 8

Simulation study results: mean parameter estimates based on the 4000 simulated samples for each combination of reader and case sample sizes. Notes: AUC_1–AUC_2 is the mean difference of the test 1 and test 2 empirical AUC estimates; *r* =number of readers; *c* =number of cases;σ^2_{TR}, σ^2_{ε}, Cov_1, Cov_2, and Cov_3 values are multiplied by 1000.

				Mean parameter estimates							
r	**c**	**AUC_2–AUC_1**	**Power**	σ^2_{TR}	σ^2_{ε}	**Cov_1**	**Cov_2**	**Cov_3**	r_1	r_2	r_3
3	50	0.066	0.189	1.330	2.350	0.873	1.116	0.475	0.358	0.461	0.189
3	100	0.066	0.276	1.282	1.123	0.427	0.549	0.234	0.372	0.480	0.200
3	200	0.066	0.343	1.294	0.547	0.210	0.270	0.115	0.378	0.488	0.204
5	50	0.065	0.256	1.251	2.335	0.861	1.106	0.469	0.357	0.461	0.190
5	100	0.066	0.407	1.257	1.126	0.426	0.551	0.233	0.372	0.482	0.200
5	200	0.066	0.525	1.250	0.549	0.211	0.271	0.115	0.381	0.490	0.206
10	50	0.066	0.355	1.196	2.349	0.867	1.112	0.470	0.360	0.462	0.191
10	100	0.066	0.571	1.260	1.119	0.423	0.543	0.229	0.373	0.479	0.200
10	200	0.066	0.781	1.257	0.550	0.272	0.382	0.115	0.382	0.491	0.207

Table 9

Simulation results: power estimates for 10 readers, 200 cases, and effect size (AUC difference) = .066. Notes: the “Actual power” estimate 0.781 is from the last line of Table 8; the “Reliable estimates” column contains power estimates based on the mean parameter estimates from Table 8; the “Sample estimates” column contains the mean of the sample power estimates across the 4000 simulated samples; P25 and P75 are the 25th and 75th percentiles of the sample power estimates; r =number of readers; c =number of cases.

			Power estimated from				
r	**c**	**Actual power**	**Reliable estimates**	**Sample estimates**	**P25**	**P75**	**P75−P25**
3	50	0.781	0.729	0.773	0.639	0.955	.316
3	100	0.781	0.742	0.776	0.658	0.935	.277
3	200	0.781	0.745	0.772	0.653	0.918	.265
5	50	0.781	0.742	0.768	0.635	0.930	.295
5	100	0.781	0.744	0.766	0.655	0.906	.251
5	200	0.781	0.751	0.767	0.666	0.892	.226
10	50	0.781	0.749	0.765	0.649	0.903	.254
10	100	0.781	0.747	0.760	0.662	0.868	.206
10	200	0.781	0.749	0.760	0.675	0.856	.181
mean:			0.744	0.767			.252