Published in final edited form as:
Acad Radiol. 2008 May ; 15(5): 647–661. doi:10.1016/j.acra.2007.12.015.

Recent Developments in the Dorfman-Berbaum-Metz Procedure for Multireader ROC Study Analysis

Stephen L. Hillis, Ph.D.,
Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP) Iowa City VA Medical Center, Iowa City, IA, U.S.A. Department of Biostatistics, University of Iowa, Iowa City, IA, U.S.A

Kevin S. Berbaum, Ph.D., and
Department of Radiology, University of Iowa, Iowa City, IA, U.S.A

Charles E. Metz, Ph.D.
Department of Radiology, University of Chicago Medical Center, Chicago, IL, U.S.A

Abstract

Rationale and Objectives—The Dorfman-Berbaum-Metz (DBM) method has been one of the most popular methods for analyzing multireader receiver operating characteristic (ROC) studies since it was proposed in 1992. Despite its popularity, the original procedure has several drawbacks: it is limited to jackknife accuracy estimates, it is substantially conservative, and it is not based on a satisfactory conceptual or theoretical model. Recently, solutions to these problems have been presented in three papers. Our purpose is to summarize and provide an overview of these recent developments.

Materials and Methods—We present and discuss the recently proposed solutions for the various drawbacks of the original DBM method.

Results—We compare the solutions in a simulation study and find that they result in improved performance for the DBM procedure. We also compare the solutions using two real data studies and find that the modified DBM procedure that incorporates these solutions yields more significant results and clearer interpretations of the variance component parameters than the original DBM procedure.

Conclusions—We recommend using the modified DBM procedure that incorporates the recent developments.

Keywords

receiver operating characteristic (ROC) curve; DBM; diagnostic radiology; jackknife; area under the curve (AUC)

Corresponding author information: Stephen L. Hillis, Ph.D. Senior Biostatistician, Center for Research in the Implementation of Innovative Strategies in Practice (CRIISP), Department of Veterans Affairs (VA) Iowa City Medical Center (152), 601 Highway 6 West, Iowa City, IA 52246-2208, Ph: 319-338-0581 x7680, E-mail: steve-hillis@uiowa.edu.

Introduction

There are several different statistical methods for analyzing multireader receiver operating characteristic (ROC) studies, with the Dorfman-Berbaum-Metz (DBM) method [1–3] being one of the most frequently used methods. The DBM method involves an analysis of variance (ANOVA) of pseudovalues computed with the Quenouille-Tukey jackknife [4–6]. The basic data for the analysis are pseudovalues corresponding to test-reader ROC accuracy measures, such as the area under the ROC curve (AUC), computed by jackknifing cases separately for each test-reader combination. Throughout we use the term *test* to refer to a diagnostic test, modality, or treatment. A mixed-effects ANOVA is performed on the pseudovalues to test the null hypothesis that the average accuracy of readers is the same for all of the diagnostic tests studied. Accuracy can be characterized using any accuracy measure, such as sensitivity, specificity, area under the ROC curve, partial area under the ROC curve, sensitivity at a fixed specificity, or specificity at a fixed sensitivity. Furthermore, these measures of accuracy can be estimated parametrically, semiparametrically or nonparametrically; the DBM method accuracy estimates are the corresponding jackknife estimates.

Although the DBM method may be the most frequently used analysis method for multireader ROC studies since it was proposed in 1992, having been used in over 100 published studies [7], the original procedure has several drawbacks: it requires that the analysis be based on jackknife accuracy estimates, it is substantially conservative, and it is not based on a satisfactory conceptual or theoretical model. Recently, solutions to these problems have been presented in three papers [8–10]. We summarize these recent developments and compare the solutions in a simulation study and in two examples.

Materials and Methods

Original DBM Method

The DBM method is typically used with the test×reader × case factorial study design where each case (i.e., patient) undergoes each of several diagnostic tests and the resulting images are interpreted once by each reader. Throughout this paper, we assume that the data have been collected using this factorial design. The competing modalities can be compared using the DBM method; in particular, the null hypothesis of no test effect can be tested and confidence intervals for test differences can be computed. Results generalize to both the population of cases and the population of readers. To simplify the narration here, we assume that the outcome is AUC.

For the original DBM method, AUC pseudovalues are computed using the Quenouille-Tukey jackknife separately for each test-reader combination as described in Dorfman et al [1]. Let Y_{ijk} denote the AUC pseudovalue for test i, reader j, and case k; by definition $Y_{ijk} = c\hat{\theta}_{ij} - (c-1)\hat{\theta}_{ij(k)}$, where c denotes the number of cases, $\hat{\theta}_{ij}$ denotes the AUC estimate based on all of the data for the ith test and jth reader, and $\hat{\theta}_{ij(k)}$, denotes the AUC estimate based on the same data but with data for the kth case removed. Thus, in effect, Y_{ijk} represents the contribution of the kth case to the accuracy estimate for the ith test and jth reader, $\hat{\theta}_{ij}$. Then using the Y_{ijk} as the data to be evaluated by conventional statistical analysis, the DBM procedure tests 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. A "jackknife estimate" of AUC for the ith test and jth reader is given by the mean of the corresponding pseudovalues:

$$\overline{Y}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c} Y_{ijk}. \quad (1)$$

We refer to $\hat{\theta}_{ij}$ as the *original* AUC estimate, $\overline{Y}_{ij\cdot}$ as the *jackknife* AUC estimate, and the Y_{ijk} as the *raw pseudovalues*.

The analysis model is expressed by

$$Y_{ijk}=\mu+\tau_i+R_j+C_k+(\tau R)_{ij}+(\tau C)_{ik}+(RC)_{jk}+(\tau RC)_{ijk}+\varepsilon_{ijk}, \quad (2)$$

$i=1,\ldots,t$; $j=1,\ldots,r$; $k=1,\ldots,c$; where τ_i denotes the fixed effect of test i, R_j denotes the random effect of reader j, C_k denotes the random effect of case k, the multiple symbols in parentheses denote interactions, and ε_{ijk} is the error term. The interaction terms are all random effects. The random effects are assumed to be mutually independent and normally distributed with zero means and respective variances $\sigma_R^2, \sigma_C^2, \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2, \sigma_{\tau RC}^2$ and σ_ε^2. Since there are no replications, $\sigma_{\tau RC}^2$ and σ_ε^2 are inseparable.

The DBM F statistic for testing for a test effect is the conventional mixed-model ANOVA F statistic based on the pseudovalues. Letting MS(T), MS(T*R), MS(T*C), and MS(T*R*C) denote the mean squares corresponding to the test, test×reader, test×case and test×reader × case effects, respectively, the F statistic for testing for a test effect for model (2) is given by

$$F=\frac{\mathrm{MS(T)}}{\mathrm{MS(T*R)+MS(T*C)-MS(T*R*C)}}. \quad (3)$$

Under the null hypothesis of no test effect, F has an approximate $F_{\mathrm{df}_1,\mathrm{df}_2}$ distribution, where $\mathrm{df}_1 = t-1$ and df_2 is the Satterthwaite [11,12] degrees of freedom approximation given by

$$\mathrm{df}_2=\frac{[\mathrm{MS(T*R)+MS(T*C)-MS(T*R*C)}]^2}{\frac{\mathrm{MS(T*R)}^2}{(t-1)(r-1)}+\frac{\mathrm{MS(T*C)}^2}{(t-1)(c-1)}+\frac{\mathrm{MS(T*R*C)}^2}{(t-1)(r-1)(c-1)}}. \quad (4)$$

In the original DBM formulation, extensive model-based simplification is performed to prevent the F statistic (3) from becoming negative (due to a negative denominator). Specifically, model (2) is simplified by omitting (or equivalently, setting to zero) the test×reader and the test×case variance components if the corresponding ANOVA estimates are not positive. For the simplified model the appropriate F statistic and denominator degrees of freedom (ddf) are used; the appropriate F statistic for each simplified model contains only one mean square in the denominator and hence cannot be negative. Thus equations (3–4) are used only when both of the variance component estimates are positive.

The test×reader and the test×case variance component ANOVA estimates are

$$\begin{aligned}\widehat{\sigma}_{\tau R}^2&=\tfrac{1}{c}[\mathrm{MS(T*R)-MS(T*R*C)}]\\ \widehat{\sigma}_{\tau C}^2&=\tfrac{1}{r}[\mathrm{MS(T*C)-MS(T*R*C)}].\end{aligned} \quad (5)$$

Taking into account possible model simplification, the F statistic and ddf for the original DBM method are given by

$$F_{\text{orig}} = \begin{cases} \frac{\text{MS(T)}}{\text{MS(T*R)+MS(T*C)-MS(T*R*C)}} & \widehat{\sigma}^2_{\tau R}>0, \widehat{\sigma}^2_{\tau C}>0 \\ \text{MS(T)/MS(T} * \text{R)} & \widehat{\sigma}^2_{\tau R}>0, \widehat{\sigma}^2_{\tau C} \le 0 \\ \text{MS(T)/MS(T} * \text{C)} & \widehat{\sigma}^2_{\tau R} \le 0, \widehat{\sigma}^2_{\tau C}>0 \\ \text{MS(T)/MS(T} * \text{R} * \text{C)} & \widehat{\sigma}^2_{\tau R} \le 0, \widehat{\sigma}^2_{\tau C} \le 0 \end{cases} \quad (6)$$

and

$$\text{ddf}_{\text{orig}} = \begin{cases} \text{equation (4)} & \widehat{\sigma}^2_{\tau R}>0, \widehat{\sigma}^2_{\tau C}>0 \\ (t-1)(r-1) & \widehat{\sigma}^2_{\tau R}>0, \widehat{\sigma}^2_{\tau C} \le 0 \\ (t-1)(c-1) & \widehat{\sigma}^2_{\tau R} \le 0, \widehat{\sigma}^2_{\tau C}>0 \\ (t-1)(r-1)(c-1) & \widehat{\sigma}^2_{\tau R} \le 0, \widehat{\sigma}^2_{\tau C} \le 0 \end{cases} . \quad (7)$$

The numerator degrees of freedom for F in equation (6) is $t-1$. We refer to this approach, using F_{orig} and ddf_{orig}, as *original DBM*. Note that the conditions in equations (6) and (7) can also be written in terms of the mean squares; e.g., $\widehat{\sigma}^2_{\tau R}>0, \widehat{\sigma}^2_{\tau C}>0$ is equivalent to MS(T*R) > MS (T*R*C), MS(T*C) > MS(T*R*C).

Problem 1: DBM is limited to jackknife accuracy estimates

One problem with original DBM is that it requires that the analysis be based on jackknife AUC estimates. Although it is possible for the jackknife AUC estimator to perform better that the corresponding original AUC estimator, clearly it would be preferable to have the flexibility to base the analysis on *either* the jackknife or original accuracy estimator, especially if (as is typically the case) it has not been shown that the jackknife AUC estimator performs as well as the original AUC estimator. For trapezoidal-rule (trapezoid) AUC estimates [13] this is not a problem, since the trapezoid and corresponding jackknife AUC estimates are equal [8].

Hillis et al [8] provide a solution to this problem by showing that the DBM method can be based on *normalized pseudovalues* Y^*_{ijk}, defined by $Y^*_{ijk}=Y_{ijk}+(\widehat{\theta}_{ij} - \overline{Y}_{ij\cdot})$. That is, the normalized pseudovalue for patient k, reader j, and test i is equal to the sum of the raw pseudovalue Y_{ijk} and the difference between the ijth test-reader original and jackknife AUC estimates. The estimate for θ_{ij} based on the normalized pseudovalues, given by $\overline{Y}^*_{ij\cdot}=\frac{1}{c}\sum_{k=1}^{c} Y^*_{ijk}$, is equal to the original AUC estimate $\hat{\theta}_{ij}$. Thus, the DBM procedure with normalized pseudovalues yields single test and test-difference confidence intervals centered on the original accuracy estimates and their differences, averaged across readers.

Problem 2: DBM is substantially conservative

Another problem with original DBM is that it is substantially conservative. Dorfman et al [3] conclude from simulations that the DBM method provides a “moderately conservative statistical test of modality differences,” with the degree of conservatism greatest with very large ROC areas and decreasing as the number of cases increases. Using the Roe and Metz [2] simulation structure, Hillis and Berbaum [14] report that, using semiparametric estimation with either normalized or raw pseudovalues, the average type I error across 144 combinations of reader-sample size, case-sample size, AUC, and variance components is.036, considerably lower than the nominal.05 significance level. The downside of a conservative test is that power is diminished compared to the same test with the critical value adjusted to yield significance levels closer to the nominal level.

In simulations Hillis [10] shows that the DBM procedure attains a type I error much closer to the nominal level when two modifications are incorporated: (1) less data-based model simplification is performed, and (2) a different ddf formula is used. We now discuss these two modifications.

Less data-based model simplification—Hillis et al [8] propose that, similar to original DBM, the test×case variance component be omitted if its ANOVA estimate is not positive; however, they stipulate that the test×reader variance component should never be omitted, even when its estimate is zero or negative. We refer to this approach as *new model simplification*. Like original DBM, new model simplification ensures that the *F* test statistic will not be negative. However, an important advantage of new model simplification is that it results in a less conservative test, with the type I error rate considerably closer to the nominal level [9]. Another advantage is that this approach avoids making inferences under the unrealistic assumption that differences between tests are the same for all readers in the population, which is implied when the test×reader variance component is omitted [14].

Using new model simplification, the *F* statistic for testing the null hypothesis of no test effect is the same as that given by equation (3) when $\widehat{\sigma}^2_{\tau C}>0$, whereas it is set equal to MS(T)/MS (T*R) when $\widehat{\sigma}^2_{\tau C} \le 0$. We denote this *F* statistic using new model simplification by F_{DBM}. Thus,

$$F_{\mathrm{DBM}} = \begin{cases} \dfrac{\mathrm{MS(T)}}{\mathrm{MS(T*R)+MS(T*C)-MS(T*R*C)}} & \widehat{\sigma}^2_{\tau C}>0 \\ \mathrm{MS(T)/MS(T*R)} & \widehat{\sigma}^2_{\tau C} \le 0 \end{cases} . \tag{8}$$

Since $\widehat{\sigma}^2_{\tau C} \le 0$ is equivalent to MS(T*C)− MS(T*R*C) ≤ 0, this *F* statistic can be succinctly written in the following form that takes model simplification into account:

$$F_{\mathrm{DBM}} = \frac{\mathrm{MS(T)}}{\mathrm{MS(T*R)+\max[\,MS(T*C)-MS(T*R*C),0]}}. \tag{9}$$

The corresponding conventional ANOVA ddf is given by

$$\mathrm{ddf_D} = \begin{cases} \text{equation (4)} & \widehat{\sigma}^2_{\tau C}>0 \\ (t-1)(r-1) & \widehat{\sigma}^2_{\tau C} \le 0 \end{cases} . \tag{10}$$

Thus, new model simplification uses F_{DBM} and $\mathrm{ddf_D}$.

In Appendix A we derive the following relationships: (1) if $\widehat{\sigma}^2_{\tau R}>0$ then $F_{\mathrm{DBM}} = F_{\mathrm{orig}}$ and $\mathrm{ddf_D} = \mathrm{ddf_{orig}}$; and (2) if $\widehat{\sigma}^2_{\tau R} \le 0$ then $F_{\mathrm{DBM}} \ge F_{\mathrm{orig}}$ but $\mathrm{ddf_D} < \mathrm{ddf_{orig}}$. However, we have found that typically the larger *F* statistic under new model simplification, when $\widehat{\sigma}^2_{\tau R}<0$, will result in a more significant conclusion (smaller *p*-value), compared to that obtained using original DBM, even though the ddf is smaller under new model simplification. In this way new model simplification produces a less conservative test.

New denominator degrees of freedom—Hillis [10] proposes a new ddf given by

$$\mathrm{ddf_H} = \begin{cases} \frac{\{\mathrm{MS(T*R)+MS(T*C)-MS(T*R*C)}\}^2}{\mathrm{MS(T*R)^2/[\,(t-1)(r-1)]}} & \widehat{\sigma}^2_{\tau C} > 0 \\ (t-1)(r-1) & \widehat{\sigma}^2_{\tau C} \le 0 \end{cases}. \tag{11}$$

Equation (11) can be written more compactly in the form

$$\mathrm{ddf_H} = \frac{\{\mathrm{MS(T*R)+\max[\,MS(T*C)-MS(T*R*C),0]}\}^2}{\frac{\mathrm{MS(T*R)^2}}{(t-1)(r-1)}}. \tag{12}$$

The quantity $\mathrm{ddf_H}$ is derived by assuming that new model simplification is used – that is, it is to be used with F_{DBM} (9). We refer to this approach, using F_{DBM} and $\mathrm{ddf_H}$, as *new model simplification plus* ddf_H.

In Appendix A we show that $\mathrm{ddf_H} > \mathrm{ddf_D}$ if $\widehat{\sigma}^2_{\tau C} > 0$, whereas $\mathrm{ddf_H} = \mathrm{ddf_D}$ if $\widehat{\sigma}^2_{\tau C} \le 0$. Since new model simplification and new model simplification plus $\mathrm{ddf_H}$ both use F_{DBM}, it follows that new model simplification plus $\mathrm{ddf_H}$ results in a lower *p*-value when $\widehat{\sigma}^2_{\tau C} > 0$ and the same *p*-value when $\widehat{\sigma}^2_{\tau C} \le 0$; hence, it is less conservative than new model simplification.

Table 1 presents a summary of the three different DBM approaches – original DBM, new model simplification, and new model simplification plus $\mathrm{ddf_H}$ – and Table 2 presents their relationships.

Problem 3: DBM model is unsatisfactory conceptually and theoretically

The original DBM procedure does not provide a satisfactory conceptual model since the the model parameters are expressed in terms of pseudovalues rather than AUC values. The model is also unsatisfactory theoretically since it assumes that the pseudovalues are independent and normally distributed -- but they are neither. Thus, desirable statistical properties of the DBM procedure do not directly follow from the model assumptions, since the assumptions are not true; rather, the validity of the model must be determined through simulation studies.

Hillis et al [8] provide a solution to this problem by showing that the DBM procedure is equivalent to another procedure that is based on an acceptable conceptual and theoretical model. Specifically, they show that the DBM model can be viewed as a "working" model that produces the same inferences as obtained using the test×reader ANOVA model with correlated errors proposed by Obuchowski and Rockette (OR) [15,16]. The OR model is given by

$$\widehat{\theta}_{ij} = \tilde{\mu} + \tilde{\tau}_i + R_j + (\tau R)_{ij} + \varepsilon_{ij}, \tag{13}$$

$i=1,\ldots,t$; $j=1,\ldots,r$; where $\hat{\theta}_{ij}$ is the AUC estimate (or other accuracy estimate) for the ith test and jth reader, 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 having mean zero and variance $\tilde{\sigma}^2_{\varepsilon}$. The random effects R_j and $(\tau R)_{ij}$ are assumed independent and normally distributed with zero means and variances $\tilde{\sigma}^2_{R}$ and $\tilde{\sigma}^2_{\tau R}$, respectively, and are assumed independent of the ε_{ij}. We use the tilde symbol "~" to distinguish OR model parameters from analogous DBM model parameters. Since the same cases are read by each reader using each

test, the error terms are not assumed to be independent. Instead, equi-covariance of the errors between readers and tests is assumed, resulting in three possible covariances given by

$$\mathrm{Cov}\left(\varepsilon_{ij},\varepsilon_{i'j'}\right)=\begin{cases}\mathrm{Cov}_1 & i\neq i',\ j=j' \text{ (different test, same reader)}\\ \mathrm{Cov}_2 & i=i',\ j\neq j' \text{ (same test, different reader)}\\ \mathrm{Cov}_3 & i\neq i',\ j\neq j' \text{ (different test, different reader)}\end{cases}. \tag{14}$$

Obuchowski and Rockette [15] suggest the following ordering: $\mathrm{Cov}_1 \geq \mathrm{Cov}_2 \geq \mathrm{Cov}_3$.

Conditional on the reader and test×reader effects (that is, treating readers as fixed), it follows from model (13) that Cov_1, Cov_2, and Cov_3 are also the corresponding covariances of the AUC estimates; for example, Cov_2 is the covariance between the AUCs for two fixed readers using the same test, while Cov_3 is the covariance between the AUCs for two fixed readers using different modalities.

The OR *F* statistic for testing for a test difference is given by

$$F_{\mathrm{OR}}=\frac{\mathrm{MS(T)}_{\hat{\theta}_{ij}}}{\mathrm{MS(T*R)}_{\hat{\theta}_{ij}}+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]}, \tag{15}$$

where $\mathrm{MS(T)}_{\hat{\theta}_{ij}}$ and $\mathrm{MS(T*R)}_{\hat{\theta}_{ij}}$ are the test and test×reader mean squares corresponding to the OR model (13), and where $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are covariance estimates; the subscript "$\hat{\theta}_{ij}$" is used here to indicate that the mean squares are computed from the AUCs rather than the pseudovalues. The quantities $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are estimated by averaging corresponding covariance estimates for pairs of AUCs, estimated using covariance estimation methods that treat readers as fixed. For example, $\widehat{\mathrm{Cov}}_2=\frac{2}{tr(r-1)}\sum_{i=1}^{t}\sum_{j<j'}\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij},\hat{\theta}_{ij'}\right)$, where $\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij},\hat{\theta}_{ij'}\right)$ is an estimate of the covariance between AUCs for fixed readers *j* and *j*′ using test *i*, estimated using a fixed reader method such as bootstrapping or jackknifing.

The DBM and OR procedures are related as follows [8]. Note that the jackknife procedure provides both AUC point estimates, defined by equation (1), and covariance and variance estimates for the AUCs, as discussed in Reference [8]. The DBM and OR *F* statistics, F_{DBM} and F_{OR} defined by equations (9) and (15), are equal if $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are jackknife covariance estimates and normalized pseudovalues are used with the DBM procedure. This relationship does not require any particular estimation method for the $\hat{\theta}_{ij}$ in equation (13). On the other hand, if raw pseudovalues are used, then the relationship still holds if, additionally, the $\hat{\theta}_{ij}$ in equation (13) are jackknife estimates. More generally, for any given AUC estimation method and any given method of estimating Cov_2 and Cov_3, $F_{\mathrm{DBM}} = F_{\mathrm{OR}}$ if the DBM procedure is used with *quasi pseudovalues*, as defined in Reference [8]. These conditions which ensure that $F_{\mathrm{DBM}} = F_{\mathrm{OR}}$ are summarized in Table 3. The appropriate ddf to use with either the DBM or OR procedure is $\mathrm{ddf}_{\mathrm{H}}$, given by equation (12) for the DBM procedure. In terms of the OR procedure mean squares, Reference [10] shows that $\mathrm{ddf}_{\mathrm{H}}$ is given by

$$\mathrm{ddf}_{\mathrm{H}}=\frac{\left\{\mathrm{MS(T*R)}+\max\left[r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right),0\right]\right\}^2}{\frac{\mathrm{MS(T*R)}^2}{(t-1)(r-1)}}. \tag{16}$$

Under any of the conditions described above that result in $F_{DBM} = F_{OR}$, the same value for ddf_H is obtained using either equation (12) or (16), and there is a one-to-one correspondence between the DBM and OR computed quantities, as shown in Table 4.

The OR model is a satisfactory conceptual model since it is expressed in terms of meaningful reader-level accuracy outcomes (e.g., AUC values). In addition, the model assumptions are reasonable. The assumed independence of the reader effects follows from the independent selection of readers, and the assumption of independent test×reader interactions and equi-covariant errors allows for a fairly general covariance structure. Normality for the error terms is reasonable since typically there are many cases for each reader, and normality for the reader and test×reader effects is a typical assumption for generalizing from a sample to a population when we do not know the exact population distribution. Of course, these assumptions may not always hold, and topics for future research include the robustness of the DBM and OR procedures to violations of these assumptions and generalization of the procedures to accommodate less restrictive assumptions.

The equivalence of the DBM and OR procedures allows for interpretation of the DBM parameters in terms of the meaningful OR parameters. Table 5 shows the relationships between the DBM and OR parameters. We see that the DBM parameters μ, τ_i, σ_R^2, and $\sigma_{\tau R}^2$ have the same interpretation as the analogous OR parameters $\tilde{\mu}$, $\tilde{\tau}_i$, $\tilde{\sigma}_R^2$ and $\tilde{\sigma}_{\tau R}^2$, while σ_C^2, $\sigma_{\tau C}^2$, σ_{RC}^2 and $\sigma_{\tau RC}^2+\sigma_\varepsilon^2$ are equal to linear functions of $\tilde{\sigma}_\varepsilon^2$, Cov_1, Cov_2 and Cov_3, and vice versa. For example, we see from Table 5 that $\sigma_{\tau C}^2 = c(Cov_2 - Cov_3)$; hence, setting $\sigma_{\tau C}^2 = 0$, as is done with new model simplification when $\widehat{\sigma}_{\tau C}^2 \leq 0$, is equivalent to assuming that $Cov_2 = Cov_3$, which is a reasonable assumption. On the other hand, we see that setting $\sigma_{\tau R}^2 = 0$, as is done with original DBM when $\widehat{\sigma}_{\tau R}^2 \leq 0$, is equivalent to assuming that the test×reader variance component of the OR model ($\tilde{\sigma}_{\tau R}^2$) is zero, implying that differences between tests are the same for all readers in the population. As mentioned earlier, this is an unreasonable assumption and is one reason why we no longer recommend original DBM.

Other examples of interpreting functions of OR parameters are the following. The expected accuracy measure across readers for the *i*th test is given by $\mu + \tau_i$; the variance of the *inherent* (or latent) reader accuracy measure is given by $\tilde{\sigma}_R^2+\tilde{\sigma}_{\tau R}^2$, with $\tilde{\sigma}_R^2$ denoting the component due to the main effect of readers and $\tilde{\sigma}_{\tau R}^2$ the component due to test×reader interaction; the variance of the reader accuracy measure *estimate* is given by $\tilde{\sigma}_R^2+\tilde{\sigma}_{\tau R}^2+\tilde{\sigma}_\varepsilon^2$; and the measurement error variance that is attributable to cases and within-reader variability that describes how a reader interprets the same image in different ways on different occasions is given by $\tilde{\sigma}_\varepsilon^2$. The interpretations of Cov_1, Cov_2 and Cov_3 have been discussed earlier. Various correlations are functions of the parameters. For example, define $\rho_{BR} = Cov_2/(\tilde{\sigma}_R^2+\tilde{\sigma}_{\tau R}^2+\tilde{\sigma}_\varepsilon^2)$ and $\rho_{BR|readers} = Cov_2/\tilde{\sigma}_\varepsilon^2$; then ρ_{BR} is the correlation between AUC estimates for two different readers using the same test, and $\rho_{BR|readers}$ is the analogous correlation but treating readers as fixed. See Appendix B for derivations of these last two correlations.

Formulas for computing the DBM variance components are presented in Table 6. Estimates for the OR variance components and covariances result from using Table 5 with the DBM variance components replaced by their estimates.

Summary of related papers

The relationship between the DBM and OR methods is described by Hillis et al [8]. They generalize the DBM method, using new model simplification, to include the use of normalized and quasi pseudovalues and determine the conditions under which the DBM and OR methods produce equal test statistics. They also show how the DBM method can be used when readers are treated as fixed and show the relationship between the DBM and OR methods for fixed readers. Hillis and Berbaum [9] show empirically that new model simplification performs better than original DBM, as well as showing that use of normalized pseudovalues has little effect on the type I error compared to raw pseudovalues. Hillis [10] derives ddf_H for both the DBM and OR procedures and empirically shows that new model simplification plus ddf_H performs better than new model simplification. Hillis and Berbaum [14] show how to compute the power for the DBM method using new model simplification; updated power software using new model simplification plus ddf_H can be downloaded from http://perception.radiology.uiowa.edu

Results

Simulation Study

In a simulation study we examined the performance of the three DBM approaches –original DBM, new model simplification, and new model simplification plus ddf_H – with respect to the empirical type I error rate for testing the null hypothesis of no test effect. The simulation model of Roe and Metz [2] provided continuous decision-variable outcomes generated from a conventional binormal model that treats both cases and readers as random. We used this simulation model to create discrete rating data by computer simulation. The discrete rating data, taking integer values from one to five, were created by transforming the continuous outcomes using the cutpoints reported by Dorfman et al [3]. The combinations of reader and case sample sizes, AUC values, and variance components were the same as those used in Roe and Metz [2] and Dorfman et al [3]. Briefly, rating data were simulated for 144 combinations of three reader-sample sizes (readers = 3, 5, and 10); four case sample sizes (10+/90−, 25+/25−, 50+/50−, and 100+/100−, where "+" indicates a diseased case and "−" indicates a normal case); three AUC values (AUC = 0.702, 0.855, and 0.961) that describe the separation between the normal and diseased case populations, averaged across readers; and four combinations of reader and case variance components. Two thousand samples were generated for each of the 144 combinations; within each simulation, all Monte Carlo readers read the same cases for each of two equal tests.

The data from each simulated sample were analyzed by all three approaches. Both maximum likelihood (semiparametric) estimation assuming a latent binormal model [17,18] and the trapezoidal-rule (nonparametric) method were used to estimate AUC from the 5-category discrete rating data. Analyses that employed semiparametric AUC estimation were performed using both raw and normalized pseudovalues, while for nonparametric AUC estimation no distinction was made since raw and normalized pseudovalues produce the same AUC estimates. For each of the 144 combinations, the empirical type I error rate was taken as the proportion of samples for which the null hypothesis was rejected at the alpha = 0.05 level. Data simulation was performed using the IML procedure in SAS [19]. The semiparametric AUC pseudovalues were computed using a dynamic link library (DLL), written in Fortran 90 by Don Dorfman and Kevin Schartz, that was accessed from within the IML procedure; this DLL, as well as a SAS macro that performs the different analyses used in this paper, can be downloaded from http://perception.radiology.uiowa.edu.

From the results, summarized in Tables 7 and 8, we draw the following conclusions. (1) New model simplification plus ddf_H has the mean empirical type I error rate closest to the nominal.

05 level: 0.051 (raw pseudovalues) and 0.049 (normalized pseudovalues) for semiparametric estimation, and 0.053 for nonparametric estimation. (2) Original DBM has the most conservative type I error rates: 0.036 (raw and normalized pseudovalues) for semiparametric estimation and 0.041 for nonparametric estimation. (3) New model simplification gives type I error rates midway between those obtained from the other two approaches. (4) With semiparametric estimation, the mean type I error rates for raw and normalized pseudovalues differ only slightly for each approach. (5) New model simplification confidence intervals can be extremely wide, due to a small proportion of proportion of samples where ddf_D approaches zero [10]. We note that new model simplification plus ddf_H does not have this problem, since ddf_H is bounded below by $(t-1)(r-1)$. (6) For semiparametric estimation using either original DBM or new model simplification plus ddf_H, normalized pseudovalue confidence interval widths are 4% smaller, on average, than those for raw pseudovalues, For new model simplification the confidence interval widths are 40% smaller, although here outliers are affecting the results as noted above. These results suggest that the original AUC estimator has more precision and power for semiparametric estimation than the jackknife AUC estimator.

Example 1: Spin-Echo versus CINE MRI for Detection of Aortic Dissection

The data for this example were provided by Carolyn Van Dyke, MD, who had obtained them in a study [20] that compared the relative performance of single Spin-Echo Magnetic Resonance Imaging (SE MRI) and CINE MRI in detecting thoracic aortic dissection. There were 45 patients with an aortic dissection and 69 patients without a dissection imaged with both SE MRI and CINE MRI. Five radiologists independently interpreted all of the images using a 5-point ordinal scale.

Table 9 presents the analysis results for raw and normalized pseudovalues obtained with semiparametric AUC estimation. We note that the jackknife and original semiparametric AUC estimates are similar, so there is little difference in the population estimates: the test AUC estimates based on the raw pseudovalues are.920 for CINE and.951 for Spin Echo, whereas the estimates based on normalized pseudovalues are.911 for CINE and.952 for Spin Echo.

Since $\widehat{\sigma}^2_{\tau R}>0$ for both types of pseudovalues, both original DBM and new model simplification yield the same results. For the normalized pseudovalues, $F_{orig} = F_{DBM} = 2.619$, $ddf_{orig} = ddf_{DBM} = 10.31$ and $p = 0.1358$ in assessing the difference in AUC. (We note that results for this and the following example differ slightly from those in References [8,9,14] because we have used an updated AUC algorithm). From equation (12) we have $ddf_H = 10.99$, resulting in $p = 0.1339$ with new model simplification plus ddf_H. Hence, the latter approach produces a slightly more significant result, illustrating a point made earlier: if $\widehat{\sigma}^2_{\tau C}>0$, then new model simplification plus ddf_H will yield a more significant result than new model simplification, since $ddf_H > ddf_{DBM}$. We note that the raw pseudovalues analysis produced less significant results, with $p = .2579$ for new model simplification and $p = .2563$ for new model simplification plus ddf_H.

Table 10 presents the DBM and OR variance components obtained on the basis of normalized pseudovalues. The DBM variance components were computed using the equations in Table 6, whereas the OR variance components and covariances were computed by replacing the DBM variance components in Table 5 with their estimates. The OR parameter estimates allow us to make statements such as the following about the variability in the reader-level AUC outcomes. The estimated variance of the inherent reader accuracy measures is $\widetilde{\sigma}^2_R+\widetilde{\sigma}^2_{\tau R}=0.000713+0.000316=0.001029$; thus, we estimate that, with probability.95, the inherent (or latent) AUC of a randomly selected reader lies within $1.96\sqrt{0.001029}=.063$ of the population test AUC. The estimated variance of the observed reader accuracy measures is

$\widehat{\sigma}_R^2+\widehat{\sigma}_{\tau R}^2+\widehat{\sigma}_\varepsilon^2=0.001029+0.001069=0.002098$. The estimated measurement error variance due to cases and within-reader variability is $\widehat{\sigma}_\varepsilon^2=0.001069$. The estimated correlation between observed AUC values for a randomly selected reader reading the same cases in different modalities is given by $\widehat{\rho}_{BR}=\widehat{\mathrm{Cov}}_2/\left(\widehat{\sigma}_R^2+\widehat{\sigma}_{\tau R}^2+\widehat{\sigma}_\varepsilon^2\right)=0.000320/0.002098=0.153$, and the analogous correlation for a given (or fixed) reader is $\widehat{\rho}_{BR|R,\tau R}=\widehat{\mathrm{Cov}}_2/\widehat{\sigma}_\varepsilon^2=0.000320/0.001069=0.300$.

Example 2: Picture archiving communication system versus plain film interpretation of neonatal examinations

Franken et al [21] compared the diagnostic accuracy of interpreting clinical neonatal radiographs using a picture archiving and communication system (PACS) workstation versus plain film. The case sample consisted of 100 chest or abdominal radiographs (67 abnormal and 33 normal). The readers were four radiologists with considerable experience in interpreting neonatal examinations. The readers indicated whether each patient had normal or abnormal findings and their degree of confidence in this judgment using a five-point ordinal scale.

Table 11 presents the ANOVA tables for the raw and normalized pseudovalues using semiparametric AUC estimation. For either type of pseudovalue we have MS(T*R) < MS (T*R*C) and MS(T*C) < MS(T*R*C); thus $\widehat{\sigma}_{\tau R}^2<0$ and $\widehat{\sigma}_{\tau C}^2<0$ from equation (5). Hence for original DBM we assume $\sigma_{\tau R}^2=\sigma_{\tau C}^2=0$ and use MS(T*R*C) as the denominator for F_{orig} with $\mathrm{ddf}_{\mathrm{orig}}=(t-1)(r-1)(c-1)=297$; in contrast, for new model simplification and new model simplification plus ddf_H we only assume $\sigma_{\tau C}^2=0$ and use MS(T*R) as the denominator for F_{DBM} with $\mathrm{ddf}_D=\mathrm{ddf}_H=(t-1)(r-1)=3$. Using the normalized pseudovalues with original DBM yields $F_{\mathrm{orig}}=0.796$, $\mathrm{ddf}_{\mathrm{orig}}=297$ and $p=0.3729$, while new model simplification and new model simplification plus ddf_H yield $F_{\mathrm{DBM}}=8.888$, $\mathrm{ddf}_D=\mathrm{ddf}_H=3$ and $p=0.0585$. The raw pseudovalues analysis produces less significant results, with $p=0.0647$ for both new model simplification and new model simplification plus ddf_H.

Discussion

We have summarized recently proposed solutions for the various drawbacks of the original DBM method and examined the performance of these solutions in a simulation study. The solutions include using normalized pseudovalues which allow DBM results to be based on either the original or the jackknife accuracy estimates; using less data-based model reduction and ddf_H to make DBM less conservative with a type I error rate much closer to the nominal level; and showing that the DBM model can be viewed as a “working” model that produces the same inferences as obtained using the acceptable conceptual and theoretical OR model. This last solution is especially important, since it establishes a solid theoretical justification for using DBM, allows us to make meaningful statements about the variability and covariances of the accuracy estimates by computing OR model parameter estimates from the DBM model parameter estimates, and allows for generalization in future research. Thus we recommend the revised DBM procedure (“new model simplification plus ddf_H”) that incorporates these recent developments. Stand-alone software as well as a SAS macro that incorporates these modifications are available to the public [22–24].

The DBM and OR approaches complement each other. We can think of each approach as consisting of a *model* and a *procedure*, where *procedure* denotes the computational algorithm steps and *model* denotes the statistical model used to motivate the procedure and justify

inferences. The OR *model* is conceptually and theoretically more acceptable. However, the DBM *procedure* is easier to implement, because after computing the pseudovalues (for each test-reader combination) the *F* statistic is easily obtained by subjecting the pseudovalues to a conventional 3-way ANOVA analysis. Furthermore, the DBM *model*, though not statistically acceptable, makes the DBM *procedure* easier to initially comprehend, especially for users without an extensive statistical background.

Finally, we note that the choice between using the original or corresponding jackknife AUC estimator should depend on which estimator has superior performance properties. For the trapezoidal method AUC this is not an issue, since the original and jackknife estimates are equal; however, for semiparametric estimation our simulation study and examples (both examples had a smaller *p* value using normalized pseudovalues) suggest that the original estimator has higher precision and power.

Acknowledgments

The authors thank Carolyn Van Dyke, M.D. for sharing her data set. This research was supported by the National Institutes of Health, grant R01EB000863. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

Grant support: This research was supported by the National Institutes of Health, grant R01EB000863.

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. 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]
3. 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]
4. Quenoille MH. Approximate tests of correlation in time series. Journal of the Royal Statistical Society, Series B 1949;11:68–84.
5. Quenoille MH. Notes on bias in estimation. Biometrika 1956;43:353–360.
6. Tukey JW. Bias and confidence in not quite large samples (abstract). Annals of Mathematical Statistics 1958;29:614.
7. Berbaum KS. God, like the devil, is in the details. Academic Radiology 2006;13:1311–1316. [PubMed: 17070448]
8. 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;24:1579–1607.10.1002/sim.2024 [PubMed: 15685718]
9. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Academic Radiology 2005;12:1534–1542.10.1016/j.acra.2005.07.012 [PubMed: 16321742]
10. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine 2007;26:596–619.10.1002/sim.2532 [PubMed: 16538699]
11. Satterthwaite FE. Synthesis of variance. Psychometrika 1941;6:309–316.
12. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometric Bulletin 1946;2:110–114.
13. 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]

14. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Academic Radiology 2004;11:1260–1273.10.1016/j.acra.2004.08.009 [PubMed: 15561573]
15. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. Communications in Statistics-Simulation and Computation 1995;24:285–308.
16. 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]
17. Dorfman DD, Alf E Jr. Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals: rating method data. Journal of Mathematical Psychology 1969;6:487–496.
18. Dorfman, DD.; RSCORE, II. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. Swets, JA.; Pickett, RM., editors. Academic Press; San Diego, CA: 1982. p. 212-232.
19. The SAS System for Windows, Version 9.1. SAS Institute Inc; Cary, NC: 2002.
20. 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. 1993.
21. Franken EA Jr, Berbaum KS, Marley SM, Smith WL, Sato Y, Kao SC, Milam SG. Evaluation of a digital workstation for interpreting neonatal examinations: a receiver operating characteristic study. Invest Radiol 1992;27:732–737. [PubMed: 1399457]
22. Berbaum, KS.; Schartz, KM.; Pesce, LL.; Hillis, SL. DBM MRMC 2.1 (Computer software). 2006. Available for download from http://perception.radiology.uiowa.edu
23. Berbaum, KS.; Metz, CE.; Pesce, LL.; Schartz, KM. DBM MRMC 2.1 User's Guide (Software manual). 2006. Available for download from http://perception.radiology.uiowa.edu
24. Hillis, SL.; Schartz, KM.; Pesce, LL.; Berbaum, KS. DBM MRMC 2.1 for SAS (Computer software). 2007. Available for download from http://perception.radiology.uiowa.edu

Appendix

APPENDIX A

In this section we derive the relationships given in Table 2 between F_{orig} and F_{DBM}, as defined by equations (6) and (8), respectively, and between ddf_{orig}, ddf_{D}, and ddf_{H}, as defined by equations (7), (10), and (11), respectively. We do this for the four possible situations corresponding to the test-by-reader and test-by-case variance component estimates being either positive or nonpositive. We make the reasonable assumptions that none of the mean squares are zero (and hence must be positive) and that the number of cases exceeds two ($c>2$).

First we derive the relationship between ddf_{D} and ddf_{H}. If $\widehat{\sigma}^2_{\tau C}>0$ then

$$\text{ddf}_{\text{D}}=\frac{[\text{MS}(\text{T}*\text{R})+\text{MS}(\text{T}*\text{C})-\text{MS}(\text{T}*\text{R}*\text{C})]^2}{\frac{\text{MS}(\text{T}*\text{R})^2}{(t-1)(r-1)}+\underbrace{\frac{\text{MS}(\text{T}*\text{C})^2}{(t-1)(c-1)}+\frac{\text{MS}(\text{T}*\text{R}*\text{C})^2}{(t-1)(r-1)(c-1)}}_{>0}}$$
$$<\frac{\{\text{MS}(\text{T}*\text{R})+\text{MS}(\text{T}*\text{C})-\text{MS}(\text{T}*\text{R}*\text{C})\}^2}{\text{MS}(\text{T}*\text{R})^2/[(t-1)(r-1)]}=\text{ddf}_{\text{H}}.$$

If $\widehat{\sigma}^2_{\tau C}\leq 0$ then

$$\text{ddf}_{\text{D}}=(t-1)(r-1)=\text{ddf}_{\text{H}}.$$

Thus $ddf_D < ddf_H$ if $\widehat{\sigma}^2_{\tau C} > 0$ and $ddf_D = ddf_H$ if $\widehat{\sigma}^2_{\tau C} \leq 0$. These relationships hold regardless of the value of $\widehat{\sigma}^2_{\tau R}$. Now we consider each of the four situations separately for the other relationships.

Situation 1

$\widehat{\sigma}^2_{\tau R} > 0, \widehat{\sigma}^2_{\tau C} > 0$. For this situation we have

$$F_{\text{orig}} = F_{\text{DBM}} = \frac{\text{MS(T)}}{\text{MS(T*R)} + \text{MS(T*C)} - \text{MS(T*R*C)}}$$
$$\text{ddf}_{\text{orig}} = \text{ddf}_{\text{D}} = \frac{[\text{MS(T*R)} + \text{MS(T*C)} - \text{MS(T*R*C)}]^2}{\frac{\text{MS(T*R)}^2}{(t-1)(r-1)} + \frac{\text{MS(T*C)}^2}{(t-1)(c-1)} + \frac{\text{MS(T*R*C)}^2}{(t-1)(r-1)(c-1)}}.$$

Situation 2

$\widehat{\sigma}^2_{\tau R} \leq 0, \widehat{\sigma}^2_{\tau C} > 0$. From equation (5) we have $c\widehat{\sigma}^2_{\tau R} = \text{MS(T} * \text{R)} - \text{MS(T} * \text{R} * \text{C)}$. Hence

$$F_{\text{orig}} = \frac{\text{MS(T)}}{\text{MS(T} * \text{C)}} \leq \frac{\text{MS(T)}}{\underbrace{\text{MS(T} * \text{R)} - \text{MS(T} * \text{R} * \text{C)}}_{c\widehat{\sigma}^2_{\tau R} \leq 0} + \text{MS(T} * \text{C)}} = F_{\text{DBM}},$$

with $F_{orig} = F_{DBM}$ if and only if $\widehat{\sigma}^2_{\tau R} = 0$. Also,

$$\text{ddf}_{\text{D}} = \frac{\left[\underbrace{\text{MS(T} * \text{R)} - \text{MS(T} * \text{R} * \text{C)}}_{c\widehat{\sigma}^2_{\tau R} \leq 0} + \text{MS(T*C)}\right]^2}{\frac{\text{MS(T*R)}^2}{(t-1)(r-1)} + \frac{\text{MS(T*C)}^2}{(t-1)(c-1)} + \frac{\text{MS(T*R*C)}^2}{(t-1)(r-1)(c-1)}}$$
$$\leq \frac{\text{MS(T*C)}^2}{\frac{\text{MS(T*R)}^2}{(t-1)(r-1)} + \frac{\text{MS(T*C)}^2}{(t-1)(c-1)} + \frac{\text{MS(T*R*C)}^2}{(t-1)(r-1)(c-1)}}$$
$$< \frac{\text{MS(T*C)}^2}{\frac{\text{MS(T*C)}^2}{(t-1)(c-1)}} = (t-1)(c-1) = \text{ddf}_{\text{orig}}.$$

That is, $ddf_D < ddf_{orig}$. In the proof we have utilized the relationship MS(T*R) −MS(T*R*C) + MS(T*C)>0, since from equation (5) we have $\text{MS(T} * \text{C)} - \text{MS(T} * \text{R} * \text{C)} = r\widehat{\sigma}^2_{\tau C} > 0$.

Situation 3

$\widehat{\sigma}^2_{\tau R} > 0, \widehat{\sigma}^2_{\tau C} \leq 0$. For this situation we have

$$F_{\text{orig}} = F_{\text{DBM}} = \frac{\text{MS(T)}}{\text{MS(T*R)}}$$
$$\text{ddf}_{\text{orig}} = \text{ddf}_{\text{D}} = (t-1)(r-1)$$

Situation 4

$\widehat{\sigma}^2_{\tau R} \leq 0, \widehat{\sigma}^2_{\tau C} \leq 0$. From equation (5) it follows that MS(T*R)≤MS(T*R*C), with equality if and only if $\widehat{\sigma}^2_{\tau R}=0$. Thus

$$F_{\text{orig}}=\frac{\text{MS(T)}}{\text{MS(T}*\text{R}*\text{C)}} \leq \frac{\text{MS(T)}}{\text{MS(T}*\text{R)}}=F_{\text{DBM}},$$

with equality if and only if $\widehat{\sigma}^2_{\tau R}=0$. Also,

$$\text{ddf}_{\text{orig}}=(t-1)(r-1)(c-1)>(t-1)(r-1)=\text{ddf}_{\text{DBM}}.$$

Note that we require the assumption that $c>2$ for this last relationship.

APPENDIX B

In this section we show how to derive AUC correlations assuming the OR model (13). Let $\widehat{\text{AUC}}_{ij}$ and $\widehat{\text{AUC}}_{i'j'}$ denote two AUC estimates, with the first subscript denoting test and the second reader. Their correlation is defined by

$$\text{Corr}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{i'j'}\right)=\frac{\text{Cov}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{i'j'}\right)}{\sqrt{\text{Var}\left(\widehat{\text{AUC}}_{ij}\right)\text{Var}\left(\widehat{\text{AUC}}_{i'j'}\right)}}$$

where $\text{Cov}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{i'j'}\right)$ is the covariance. To find the covariance and variances, we write $\widehat{\text{AUC}}_{ij}$ and $\widehat{\text{AUC}}_{i'j'}$ as functions of random and fixed effects using the OR model (13). It follows from well known statistical properties that the variance for each AUC estimate is the sum of the OR model variance components corresponding to the random effects, and $\text{Cov}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{i'j'}\right)$ is the sum of the variance components corresponding to the reader or test×reader random effects that the AUC estimates have in common (i.e., they have the same subscript values for each AUC estimate), plus the covariance between the error terms.

For example, the between-reader correlation between AUC estimates for two different readers using the same test is given by

$$\rho_{\text{BR}}=\text{Corr}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{ij'}\right)=\frac{\text{Cov}\left(\widehat{\text{AUC}}_{ij}, \widehat{\text{AUC}}_{ij'}\right)}{\sqrt{\text{Var}\left(\widehat{\text{AUC}}_{ij}\right)\text{Var}\left(\widehat{\text{AUC}}_{ij'}\right)}} \quad (17)$$

where $j \neq j'$. From equation (13), with $\widehat{\text{AUC}}_{ij}$ taking the place of $\hat{\theta}_{ij}$, we have

$$\widehat{\mathrm{AUC}}_{ij} = \tilde{\mu} + \tilde{\tau}_i + R_j + (\tau R)_{ij} + \varepsilon_{ij}$$
$$\widehat{\mathrm{AUC}}_{ij'} = \tilde{\mu} + \tilde{\tau}_i + R_{j'} + (\tau R)_{ij'} + \varepsilon_{ij'}. \quad (18)$$

Each AUC estimate has the same variance, equal to the sum of all of the variance components corresponding to the random effects; that is,

$$\mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij}\right) = \mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij'}\right) = \tilde{\sigma}_R^2 + \tilde{\sigma}_{\tau R}^2 + \tilde{\sigma}_\varepsilon^2.$$

Examination of equations (18) shows that the AUCs do not have any reader or test×reader random effects in common since $j \neq j'$. Thus the covariance is equal to Cov_2, the covariance between the error terms for different readers using the same test:

$$\mathrm{Cov}\left(\widehat{\mathrm{AUC}}_{ij}, \widehat{\mathrm{AUC}}_{ij'}\right) = \mathrm{Cov}_2. \quad (19)$$

It follows from equations (17), (18) and (19) that

$$\rho_{\mathrm{BR}} = \frac{\mathrm{Cov}_2}{\tilde{\sigma}_R^2 + \tilde{\sigma}_{\tau R}^2 + \tilde{\sigma}_\varepsilon^2}.$$

Now we derive the between-reader correlation between AUC estimates for two different readers using the same test, but this time *treating readers as fixed*. In this case the correlation is a measure of the association between the deviation of one reader's AUC estimate from that reader's underlying AUC, due to case variation and reader error, with the deviation of the other reader's AUC estimate from that reader's underlying AUC. In contrast, ρ_{BR} is a measure of association between deviations of *randomly chosen* readers' AUC estimates from the *reader population* AUC.

To derive this correlation we treat the reader and test×reader effects as fixed in model (13) by conditioning on them; thus these effects do not have corresponding variance components, but rather are treated like constants. We denote this correlation by $\rho_{\mathrm{BR|readers}}$ to indicate that it is for two fixed readers. The correlation is defined as before, except now the covariance and variances are conditional on the reader and test×reader random effects:

$$\rho_{\mathrm{BR|readers}} = \mathrm{Corr}\left(\widehat{\mathrm{AUC}}_{ij}, \widehat{\mathrm{AUC}}_{ij'} | R_i, (\tau R)_{ij}, (\tau R)_{ij'}\right)$$
$$= \frac{\mathrm{Cov}\left(\widehat{\mathrm{AUC}}_{ij}, \widehat{\mathrm{AUC}}_{ij'} | R_i, (\tau R)_{ij}, (\tau R)_{ij'}\right)}{\sqrt{\mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij} | R_i, (\tau R)_{ij}, (\tau R)_{ij'}\right) \mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij'} | R_i, (\tau R)_{ij}, (\tau R)_{ij'}\right)}} \quad (20)$$

When we condition on the reader and test×reader random effects, the only random effects in equations (18) are the error terms. Thus each AUC has the same variance, equal to $\tilde{\sigma}_\varepsilon^2$:

$$\mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij} | R_i, (\tau R)_{ij}\right) = \mathrm{Var}\left(\widehat{\mathrm{AUC}}_{ij'} | R_i, (\tau R)_{ij}\right) = \tilde{\sigma}_\varepsilon^2. \quad (21)$$

Similarly, the covariance is equal to Cov_2, the covariance between the error terms:

$$\mathrm{Cov}\left(\widehat{\mathrm{AUC}}_{ij}, \widehat{\mathrm{AUC}}_{ij'} | R_i, (\tau R)_{ij}, (\tau R)_{ij'}\right) = \mathrm{Cov}_2 \tag{22}$$

It follows from equations (20), (21) and (22) that

$$\rho_{\mathrm{BR|readers}} = \frac{\mathrm{Cov}_2}{\tilde{\sigma}^2_{\varepsilon}}.$$

These correlations can be written in terms of the DBM model parameters using the relationships in Table 5. For example, since $\mathrm{Cov}_2 = (\sigma^2_C + \sigma^2_{\tau C})$ and $\tilde{\sigma}^2_{\varepsilon} = \sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_{tRC} + \sigma^2_{\varepsilon}$, where $\sigma^2_C, \sigma^2_{\tau C}, \sigma^2_{RC}, \sigma^2_{\tau RC}$ and σ^2_{ε} denote the DBM model variance components, then $\rho_{\mathrm{BR|readers}} = (\sigma^2_C + \sigma^2_{\tau C})/(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_{\tau RC} + \sigma^2_{\varepsilon})$ in terms of the DBM variance components. This last expression is also given in equation (4) of Reference [2].

Table 1

Summary of the different DBM approaches

a) Original DBM

F_{orig}	ddf_{orig}	condition
$\frac{MS(T)}{MS(T \times R) + MS(T \times C) - MS(T \times R \times C)}$	Equation (4)	$\hat{\sigma}^2_{\tau R} > 0,\ \hat{\sigma}^2_{\tau C} > 0$
MS(T)/MS(T*R)	$(t-1)(r-1)$	$\hat{\sigma}^2_{\tau R} > 0,\ \hat{\sigma}^2_{\tau C} \le 0$
MS(T)/MS(T*C)	$(t-1)(c-1)$	$\hat{\sigma}^2_{\tau R} \le 0,\ \hat{\sigma}^2_{\tau C} > 0$
MS(T)/MS(T*R*C)	$(t-1)(r-1)(c-1)$	$\hat{\sigma}^2_{\tau R} \le 0,\ \hat{\sigma}^2_{\tau C} \le 0$

b) New model simplification

$$F_{DBM} = \frac{MS(T)}{MS(T \times R) + \max[MS(T \times C) - MS(T \times R \times C),\ 0]}$$

$$ddf_D = \begin{cases} \text{equation (3)} & \hat{\sigma}^2_{\tau C} > 0 \\ (t-1)(r-1) & \hat{\sigma}^2_{\tau C} \le 0 \end{cases}$$

c) New model simplification plus ddf_H

$$F_{DBM} = \frac{MS(T)}{MS(T \times R) + \max[MS(T \times C) - MS(T \times R \times C),\ 0]} \text{[same as in (b)]}$$

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

These approaches can be used with raw, normalized, or quasi pseudovalues. See Table 6 for computational formulas for $\tilde{\sigma}^2_{\tau R}$ and $\tilde{\sigma}^2_{\tau C}$.

Table 2

Relationships between the DBM *F* statistics and between the DBM denominator degrees of freedom.

$\sigma^2_{\tau R}$	$\sigma^2_{\tau C}$	***F* relationship**	**Ddf relationship**
>0	>0	$F_{orig} = F_{DBM}$	$ddf_{orig} = ddf_D < ddf_H$
≤0	>0	$F_{orig} \leq F_{DBM}$ (equality iff $\sigma^2_{\tau R} = 0$)	$ddf_D < ddf_{orig}$, $ddf_D < ddf_H$
>0	≤0	$F_{orig} = F_{DBM}$	$ddf_{orig} = ddf_D = ddf_H$
≤0	≤0	$F_{orig} \leq F_{DBM}$ (equality iff $\sigma^2_{\tau R} = 0$)	$ddf_D = ddf_H < ddf_{orig}$

These relationships are derived in Appendix A. Iff: if and only if.

Table 3

Conditions which result in $F_{DBM} = F_{OR}$ as defined by equations (9) and (15)

1	Normalized pseudovalues are used with DBM and $\widetilde{\sigma}_{\varepsilon}^2, \widehat{\mathrm{Cov}}_1, \widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are jackknife variance and covariance estimates.
	or
2	Raw pseudovalues are used with DBM, $\widetilde{\sigma}_{\varepsilon}^2, \widehat{\mathrm{Cov}}_1, \widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are jackknife variance and covariance estimates, and $\hat{\theta}_{ij}$ are jackknife accuracy estimates.
	or
3	Quasi pseudovalues are used with DBM.

Note: any one of the above conditions results in $F_{DBM} = F_{OR}$.

Table 4

Relationship between DBM and OR computed quantities.

OR computed quantity	Equivalent function of DBM computed quantities
$MS(T)_{\hat{\theta}_{ij}}$	$= \frac{1}{c}MS(T)$
$MS(R)_{\hat{\theta}_{ij}}$	$= \frac{1}{c}MS(R)$
$MS(T*R)_{\hat{\theta}_{ij}}$	$= \frac{1}{c}MS(T \times R)$
$\hat{\sigma}^2_{\varepsilon}$	$= \frac{1}{trc}[MS(C) + (t-1)MS(T \times C) + (r-1)MS(R \times C) + (t-1)(r-1)MS(T \times R \times C)]$
$\widehat{Cov}_1$	$= \frac{1}{trc}\{MS(C) - MS(T \times C) + (r-1)[MS(R \times C) - MS(T \times R \times C)]\}$
$\widehat{Cov}_2$	$= \frac{1}{trc}\{MS(C) - MS(R \times C) + (t-1)[MS(T \times C) - MS(T \times R \times C)]\}$
$\widehat{Cov}_3$	$= \frac{1}{trc}[MS(C) - MS(T \times C) - MS(R \times C) + MS(T \times R \times C)]$
DBM computed quantity	Equivalent function of OR computed quantities
MS(T)	$=cMS(T)_{\hat{\theta}_{ij}}$
MS(R)	$=cMS(R)_{\hat{\theta}_{ij}}$
MS(T*R)	$=cMS(T*R)_{\hat{\theta}_{ij}}$
MS(C)	$= c[\hat{\sigma}^2_{\varepsilon} - (t-1)\widehat{Cov}_1 + (r-1)\widehat{Cov}_2 + (t-1)(r-1)\widehat{Cov}_3)]$
MS(T*C)	$= c[\hat{\sigma}^2_{\varepsilon} - \widehat{Cov}_1 + (r-1)(\widehat{Cov}_2 - \widehat{Cov}_3)]$
MS(R*C)	$= c[\hat{\sigma}^2_{\varepsilon} + (t-1)\widehat{Cov}_1 - \widehat{Cov}_2 - (t-1)\widehat{Cov}_3)]$
MS(T*R*C)	$= c[\hat{\sigma}^2_{\varepsilon} - \widehat{Cov}_1 - \widehat{Cov}_2 + \widehat{Cov}_3]$

These relationships assume one of the three conditions given in Table 3.

Table 5

Relationship between DBM and OR model parameters

OR model parameter	Equivalent function of DBM model parameters
$\tilde{\mu}$	$=\mu$
$\tilde{\tau}_i$	$=\tau_i$
$\hat{\sigma}^2_R$	$=\sigma^2_R$
$\hat{\sigma}^2_{\tau R}$	$=\sigma^2_{\tau R}$
$\hat{\sigma}^2_{\varepsilon}$	$=(\sigma^2_C+\sigma^2_{\tau C}+\sigma^2_{RC}+\sigma^2_{\tau RC}+\sigma^2_{\varepsilon})/c$
Cov_1	$=(\sigma^2_C+\sigma^2_{RC})/c$
Cov_2	$=(\sigma^2_C+\sigma^2_{\tau C})/c$
Cov_3	$=\sigma^2_C/c$
DBM model parameter	Equivalent function of OR model parameters
μ	$=\tilde{\mu}$
τ_i	$=\tilde{\tau}_i$
σ^2_R	$=\hat{\sigma}^2_R$
$\sigma^2_{\tau R}$	$=\hat{\sigma}^2_{\tau R}$
σ^2_C	$=c\text{Cov}_3$
$\sigma^2_{\tau C}$	$=c(\text{Cov}_2-\text{Cov}_3)$
σ^2_{RC}	$=c(\text{Cov}_1-\text{Cov}_3)$
$\sigma^2_{\tau RC}+\sigma^2_{\varepsilon}$	$=c(\hat{\sigma}^2_{\varepsilon}-\text{Cov}_1-\text{Cov}_2+\text{Cov}_3)$

These relationships assume that the constraints for the OR model parameters are those implied by the DBM model: $\sigma^2_{\varepsilon}\geq \text{Cov}_1+\text{Cov}_2-\text{Cov}_3$, $\text{Cov}_1\geq\text{Cov}_3$, $\text{Cov}_2\geq\text{Cov}_3$, and $\text{Cov}_3\geq 0$. They also assume the same linear constraint for the τ_i (e.g., $\Sigma\tau_i=0$) for both models and that either (1) normalized or quasi pseudovalues are used; or (2) if raw pseudovalues are used, then the OR model outcome is the jackknife accuracy estimate.

Note: Adapted and reprinted, with permission, from Reference [8]

Table 6

ANOVA estimates for DBM variance components

DBM model parameter	Estimate
σ_R^2	$\frac{1}{tc}[\mathrm{MS(R)} - \mathrm{MS(T \times R)} - \mathrm{MS(R \times C)} + \mathrm{MS(T \times R \times C)}]$
σ_C^2	$\frac{1}{tr}[\mathrm{MS(C)} - \mathrm{MS(T \times C)} - \mathrm{MS(R \times C)} + \mathrm{MS(T \times R \times C)}]$
$\sigma_{\tau R}^2$	$\frac{1}{c}[\mathrm{MS(T \times R)} - \mathrm{MS(T \times R \times C)}]$
$\sigma_{\tau C}^2$	$\frac{1}{r}[\mathrm{MS(T \times C)} - \mathrm{MS(T \times R \times C)}]$
σ_{RC}^2	$\frac{1}{t}[\mathrm{MS(R \times C)} - \mathrm{MS(T \times R \times C)}]$
$\sigma_{\tau RC}^2 + \sigma_\varepsilon^2$	MS(T*R*C)

Note: These estimates, except for the last, can be negative.

Table 7

Semiparametric estimation results of the simulation study for discrete rating data.

		Type I error rates						
Approach	**Pseudovalues**	**N**	**Mean**	**Min**	**Max**	**Range**	**SD**	**CI width mean**
Original	raw	144	0.036	0.009	0.063	0.054	0.0124	0.196
	normalized	144	0.036	0.011	0.062	0.052	0.0111	0.188
New	raw	144	0.042	0.011	0.070	0.060	0.0123	4.05E+121
	normalized	144	0.043	0.017	0.067	0.050	0.0108	2.74E+121
New plus ddf_H	raw	144	0.049	0.016	0.075	0.060	0.0124	0.192
	normalized	144	0.051	0.025	0.077	0.052	0.0105	0.184

Original: original DBM; New: new model simplification; New plus ddf_H: new model simplification plus ddf_H; Min: minimum; Max: maximum; SD: standard deviation; CI width: width of a 95% confidence interval for the difference of the AUC estimates.

Table 8

Nonparametric estimation results of the simulation study for discrete rating data.

	Type I error rates						
Approach	**N**	**Mean**	**Min**	**Max**	**Range**	**SD**	**CI width mean**
original	144	0.041	0.014	0.069	0.055	0.0098	0.177
new	144	0.046	0.024	0.072	0.049	0.0100	4.55E+121
new plus ddf_H	144	0.053	0.029	0.079	0.050	0.0097	0.174

No distinction is made between raw and normalized pseudovalues since the trapezoid estimate is the same for either type of pseudovalues. Original: original DBM; new: new model simplification; new plus ddf_H: new model simplification plus ddf_H; min: minimum; max: maximum; SD: standard deviation; CI width: width of a 95% confidence interval for the difference of the AUC estimates.

Table 9

DBM procedure analyses for Van Dyke et al [20] data

Semiparametric and corresponding jackknife AUC estimates:

	test			
	1 (CINE)		2 (Spin Echo)	
reader (j)	$\hat{\theta}_{1j}$ (semiparametric)	$Y_{1j\cdot}$ (jackknife)	$\hat{\theta}_{2j}$ (semiparametric)	$Y_{2j\cdot}$ (jackknife)
1	0.933	0.947	0.951	0.950
2	0.890	0.909	0.935	0.933
3	0.929	0.929	0.928	0.928
4	0.970	0.981	1.000	0.999
5	0.833	0.836	0.945	0.943
	$\hat{\theta}_{1\cdot}$ =.911	$Y_{1\cdot\cdot}$ =.920	$\hat{\theta}_{2\cdot}$ =.952	$Y_{2\cdot\cdot}$ =.951

ANOVA table:

Source	ddf	Raw pseudovalue mean square	Normalized pseudovalue mean square
T	1	0.264166	0.468996
R	4	0.315637	0.297310
C	113	0.392538	0.392538
T×R	4	0.112560	0.108062
T×C	113	0.143095	0.143095
R×C	452	0.098771	0.098771
T×R×C	452	0.072068	0.072068

T: tests; R: readers; C: cases.

Raw pseudovalues results:

[a] Original DBM: $F_{\text{orig}} = 1.439$, $\text{ddf}_{\text{orig}} = 10.03$, $p = 0.2579$

[b] New model simplification: $F_{\text{DBM}} = 1.439$, $\text{ddf}_{\text{D}} = 10.03$, $p = 0.2579$

[c] New model simplification plus ddf_{H}: $F_{\text{DBM}} = 1.439$, $\text{ddf}_{\text{H}} = 10.64$, $p = 0.2563$

Normalized pseudovalues results:

[a] Original DBM: $F_{\text{orig}} = 2.619$, $\text{ddf}_{\text{orig}} = 10.31$, $p = 0.1358$

[b] New model simplification: $F_{\text{DBM}} = 2.619$, $\text{ddf}_{\text{D}} = 10.31$, $p = 0.1358$

[c] New model simplification plus ddf_{H}: $F_{\text{DBM}} = 2.619$, $\text{ddf}_{\text{H}} = 10.99$, $p = 0.1339$

Table 10

Variance component estimates for Van Dyke et al [20] data based on normalized pseudovalues

DBM		OR	
Variance component	**Estimate**	**Variance component**	**Estimate**
σ_R^2	0.000713	σ^2_R	0.000713
$\sigma_{\tau R}^2$	0.000316	$\sigma^2_{\tau R}$	0.000316
σ_C^2	0.022274	Cov_1	0.000313
$\sigma_{\tau C}^2$	0.014205	Cov_2	0.000320
σ_{RC}^2	0.013351	Cov_3	0.000195
$\sigma_{\tau RC}^2 + \sigma_\varepsilon^2$	0.072068	σ^2_ε	0.001069

Table 11

DBM procedure analyses for Franken et al [21] data.

ANOVA table:

Source	ddf	Raw pseudovalue Mean square	Normalized pseudovalue Mean square
T	1	0.063574	0.066606
R	3	0.088782	0.097686
C	99	0.547734	0.547734
T×R	3	0.007781	0.007494
T×C	99	0.078071	0.078071
R×C	297	0.127582	0.127582
T×R×C	297	0.083643	0.083643

T: tests; R: readers; C: cases.

Raw pseudovalues results:

[a] Original DBM: $F_{orig} = 0.760$, $ddf_{orig} = 297$, $p = 0.3840$

[b] New model simplification: $F_{DBM} = 8.171$, $ddf_D = 3$, $p = 0.0647$

[c] New model simplification plus ddf_H: $F_{DBM} = 8.171$, $ddf_H = 3$, $p = 0.0647$

Normalized pseudovalues results:

[a] Original DBM: $F_{orig} = 0.796$, $ddf_{orig} = 297$, $p = 0.3729$

[b] New model simplification: $F_{DBM} = 8.888$, $ddf_D = 3$, $p = 0.0585$

[c] New model simplification plus ddf_H: $F_{DBM} = 8.888$, $ddf_H = 3$, $p = 0.0585$