Published in final edited form as:
Stat Med. 2018 June 15; 37(13): 2067–2093. doi:10.1002/sim.7616.

Relationship between Roe and Metz simulation model for multireader diagnostic data and Obuchowski-Rockette model parameters

Stephen L. Hillis, Ph.D
Departments of Radiology and Biostatistics, The University of Iowa, 3710 Medical Laboratories, 200 Hawkins Drive, Iowa City, IA 52242-1077, U.S.A

Abstract

For the typical diagnostic-radiology study design, each case (i.e., patient) undergoes several diagnostic tests (or modalities) and the resulting images are interpreted by several readers. Often each reader is asked to assign a confidence-of-disease rating to each case for each test and the diagnostic tests are compared with respect to reader performance outcomes that are functions of the reader ROC curves, such as the area under the ROC curve. These reader-performance outcomes are frequently analyzed using the Obuchowski and Rockette method, which allows conclusions to generalize to both the reader and case populations. The simulation model proposed by Roe and Metz in 1997 emulates confidence-of-disease data collected from such studies and has been an important tool for empirically evaluating various reader-performance analysis methods. However, because the Roe and Metz model parameters are expressed in terms of a continuous decision variable rather than in terms of reader performance outcomes, it has not been possible to evaluate the realism of the RM model. I derive the relationships between the Roe-Metz and Obuchowski-Rockette model parameters for the empirical AUC reader-performance outcome. These relationships make it possible to evaluate the realism of the RM parameter models and to assess the performance of Obuchowski-Rockette parameter estimates. An example illustrates the application of the relationships for assessing the performance of a proposed upper one-sided confidence bound for the Obuchowski-Rockette test-by-reader variance component, which is useful for sample size estimation.

Keywords

Receiver operating characteristic (ROC) curve; Roe and Metz model; Obuchowski-Rockette; diagnostic radiology

Correspondence to: Stephen L. Hillis.

Disclaimer
The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

1. Introduction

For the typical diagnostic-radiology study design, each case (i.e., patient) undergoes each of several diagnostic tests (or modalities) and the resulting images are interpreted by each of several readers. Often each reader is asked to assign a confidence-of-disease rating to each case for each test, based on the corresponding image or set of images, and a response-operating-characteristic (ROC) curve for each reader is estimated from the case-level ratings. The resulting data are called multireader multicase (MRMC) data. The diagnostic tests are then compared with respect to reader-performance outcomes that are functions of the reader ROC curves. A commonly used reader-performance summary outcome is the area under the ROC curve (AUC).

Roe and Metz (RM) [1] proposed a model for the simulation of MRMC data that emulate confidence-of-disease data collected from such studies. Studies that have used this model for evaluating various MRMC analysis and sample size methods include References [2]–[18]. To account for the MRMC study design, the RM model generates continuous decision variable (DV) data according to a binormal model for which the nondiseased and diseased DV variances are equal for each test-reader combination, while allowing the differences between the diseased and nondiseased DV means to vary across test-reader combinations. Hillis [12] and [14] proposed extensions of the RM model that allow the diseased and nondiseased DV distributions for each reader to have unequal diseased and nondiseased variances, while keeping intact other characteristics of the original RM model, and recently Gallas and Hillis [15] proposed a more general extension.

The usefulness of a simulation model depends on the degree to which the simulated data are similar to data encountered in practice. Although Roe and Metz state that they used “correlation estimates found in actual ROC analyses” estimated with the CORROC algorithm [19], which assumes a bivariate binormal model, these correlations are defined in terms of the DV distribution. In particular, Roe and Metz do not analytically relate these DV correlations and other parameters in their model to parameters estimated by the frequently used MRMC analysis methods proposed by Obuchowski and Rockette (OR) [20] and Dorfman, Berbaum, and Metz [21]. Knowledge of these relationships would be useful for assessing both the realism of the RM model and the performance of the OR MRMC analysis method.

In this paper I determine analytic relationships between the OR- and RM-model parameter values for empirical AUC reader-performance outcomes computed from RM-model simulated data. The relationships are for the situation where the goal is to compare two tests. Using these relationships, I evaluate the realism of the RM parameter models and assess the performance of OR parameter estimates. An example illustrates the application of the relationships for assessing the performance of a proposed upper one-sided confidence bound for the OR test×reader variance component, which is useful for sample size estimation.

2. Obuchowski-Rockette method

2.1. Design and definitions

Throughout I assume that each case is subjected to each of two tests and is assigned a confidence-of-disease rating by each of *r* readers. In addition, each case is classified as diseased or nondiseased according to an available reference standard. Let $\hat{\theta}_{ij}$ denote the AUC estimate (or other reader-performance outcome) for the *j*th reader reading a randomly selected case sample consisting of n_- nondiseased and n_+ diseased cases using test *i* with $i = 1, 2$ and $j = 1, \ldots, r$. Each reader reads the same cases using each test. To simplify the narration, I often implicitly assume that the reader performance outcome is AUC.

For test *i* and fixed reader *j*, I refer to the expected AUC estimate as the *reader-specific expected AUC* and denote it by ξ_{ij}; i.e.,

$$\xi_{ij} = E\left(\hat{\theta}_{ij} | \text{fixed reader } j\right) \quad (1)$$

Conceptually, ξ_{ij} is the mean of AUC estimates that would result from reader *j* reading many randomly selected case samples. I will use the notation $\xi_{ij}^{(\text{empirical})}$ when I am assuming that $\hat{\theta}_{ij}$ is the empirical AUC estimate in (1), denoted by $\hat{\theta}_{ij}^{(\text{empirical})}$; otherwise, the notation $\hat{\theta}_{ij}$ and ξ_{ij} will be used for results that do not depend on the specific type of estimator. I will use the notation AUC_{ij} to denote the *reader-specific true AUC* for test *i* and fixed reader *j*, which is equal to the probability that a randomly chosen diseased case will receive a higher confidence-of-disease rating than a randomly chosen nondiseased case, plus one-half of the probability of a tied score [22, 23]. Depending on the type of AUC estimator, ξ_{ij} may or may not be equal to AUC*ij*. For example, if the confidence-of-disease ratings are continuous, then $E\left(\hat{\theta}_{ij}^{(\text{empirical})} | \text{fixed reader } j\right) = \text{AUC}_{ij}$ [22, 23], i.e., $\xi_{ij}^{(\text{empirical})} = \text{AUC}_{ij}$. For a simple example where ξ_{ij} is not equal to AUC_{ij}, consider the situation where the rating distributions are normally distributed for fixed reader *j*. It then follows from maximum-likelihood theory that the maximum-likelihood binormal ROC curve AUC estimator is a consistent estimator of AUC_{ij}. However, in general it is not unbiased, i.e., $\xi_{ij}^{(\text{binormal})} \neq \text{AUC}_{ij}$.

2.2. Obuchowski-Rockette (OR) model

Obuchowski and Rockette [20] proposed a test × reader factorial ANOVA model for the AUC estimates, but unlike a conventional ANOVA model, the errors are assumed to be correlated to account for correlation due to each reader evaluating the same cases. Their model, which I refer to as the *OR model*, is given by

$$\hat{\theta}_{ij} = \mu_{\text{OR}} + \tau_{i:\text{OR}} + R_{j:\text{OR}} + (\tau R)_{ij:\text{OR}} + \varepsilon_{ij:\text{OR}} \quad (2)$$

where μ_{OR} is the intercept term, $\tau_{i:\text{OR}}$ denotes the fixed effect of test *i*, $R_{j:\text{OR}}$ denotes the random effect of reader *j*, $(\tau R)_{ij:\text{OR}}$ denotes the random test × reader interaction, and $\varepsilon_{ij:\text{OR}}$

is the error term. The $R_{j:\text{OR}}$ and $(\tau R)_{ij:\text{OR}}$ are assumed to be mutually independent and normally distributed with zero means and respective variances $\sigma^2_{R:\text{OR}}$ and $\sigma^2_{TR:\text{OR}}$. (I include "OR" in effect and variance component subscripts to distinguish OR effects and variance components from similarly notated RM-model quantities.) The $\varepsilon_{ij:\text{OR}}$ are assumed to be normally distributed with mean zero and variance $\sigma^2_{\varepsilon:\text{OR}}$ and are assumed independent of the $R_{j:\text{OR}}$ and $(\tau R)_{ij:\text{OR}}$. Three possible error covariances are assumed:

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

It follows from (2) that $\sigma^2_{\varepsilon:\text{OR}}$ is equal to the variance of AUC for a single fixed reader and test and Cov_1, Cov_2, and Cov_3 are the AUC covariances for a single fixed reader and two tests, two different fixed readers and one test, and two different fixed readers using two different tests, respectively. These error variance-covariance parameters are typically estimated by averaging corresponding estimates computed using a fixed-reader method, such as the jackknife [24, 25], bootstrap [26], or the method proposed by DeLong et al [27] (for empirical AUC estimates). The $\varepsilon_{ij:\text{OR}}$ can be interpreted as AUC measurement error attributable to the random selection of cases and within-reader variability that describes how a fixed reader interprets the same image in different ways on different occasions. The OR model can alternatively be described with population correlations $r_i = \text{Cov}/\sigma^2_{\varepsilon:\text{OR}}$ replacing corresponding Cov_i.

For fixed reader j using test i, it follows from (2) that the reader-specific expected AUC is given by

$$\xi_{ij} = E\left(\hat{\theta}_{ij} | R_{j:\text{OR}}, (\tau R)_{ij:\text{OR}}\right)$$

Here I have conditioned on effects involving reader because reader is treated as fixed. It follows that

$$\xi_{ij} = \mu_{\text{OR}} + \tau_{i:\text{OR}} + R_{j:\text{OR}} + (\tau R)_{ij:\text{OR}}$$

For determining formulas that relate $\sigma^2_{R:\text{OR}}$, $\sigma^2_{TR:\text{OR}}$, and $\mu_{\text{OR}} + \tau_{i:\text{OR}}$ to RM-model parameters, it is useful to express these OR parameters in terms of the distribution of the reader-specific expected AUCs. In particular, it is easy to show that, across the reader population (i.e., now treating readers as random),

$$\sigma^2_{R:\text{OR}} = \underset{i \neq i'}{\text{cov}}(\xi_{ij}, \xi_{i'j}) \quad (3)$$

$$\sigma^2_{TR:\mathrm{OR}} = \underset{i \neq i'}{.5\mathrm{var}}(\xi_{ij} - \xi_{i'j}) \quad (4)$$

and

$$\mu_{\mathrm{OR}} + \tau_{i:\mathrm{OR}} = E(\xi_{ij}) \quad (5)$$

The OR estimate of the between-test AUC difference is given by $\hat{\theta}_{1\cdot} - \hat{\theta}_{2\cdot}$. (Here and elsewhere, a subscript replaced by a dot indicates an average across corresponding subscript values; e.g., $\hat{\theta}_{i\cdot} = \frac{1}{r}\sum_{j=1}^{r}\hat{\theta}_{ij}$.) The OR model implies that the variance of this estimate is given by

$$\mathrm{var}\left(\hat{\theta}_{1\cdot} - \hat{\theta}_{2\cdot}\right) = \frac{2}{r}[\sigma^2_{TR:\mathrm{OR}} + \sigma^2_{\varepsilon:\mathrm{OR}} - \mathrm{Cov}_1 + (r-1)(\mathrm{Cov}_2 - \mathrm{Cov}_3)]$$

The OR estimate of this variance is given by

$$\widehat{\mathrm{var}}\left(\hat{\theta}_{i\cdot} - \hat{\theta}_{i'\cdot}\right) = \frac{2}{r}\left\{\mathrm{MS(T*R)} + \max\left[r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right), 0\right]\right\} \quad (6)$$

where $\mathrm{MS(T*R)} = \frac{1}{(r-1)}\sum_{i=1}^{2}\sum_{j=1}^{r}\left(\hat{\theta}_{ij} - \hat{\theta}_{\cdot j} - \hat{\theta}_{i\cdot} + \hat{\theta}_{\cdot\cdot}\right)^2$ is the test×reader interaction mean square and $\widehat{\mathrm{Cov}}_2$ and $\widehat{\mathrm{Cov}}_3$ are estimates of Cov_2 and Cov_3, respectively.

2.3. Relationship with Dorfman-Berbaum-Metz (DBM) method

The method proposed by Dorfman, Berbaum and Metz (DBM) [2, 21] involves a conventional test×reader×case analysis of jackknife pseudovalues. Hillis and colleagues [7, 9, 10] suggested modifications that improved the performance for OR and DBM and which have been incorporated in current software programs. For the improved versions, DBM is equivalent to OR using jackknife error variance-covariance estimates. A disadvantage of the DBM model is that some of its parameters are difficult to interpret. Because the OR parameters have clear interpretations, I only derive relationships between them and the RM-model parameters.

2.4. Performance of the OR method

The performance of the OR method, modified as suggested by Hillis and colleagues [7, 9, 10], has been assessed in several simulations studies [9, 10, 12, 17, 18]. For these studies data were simulated using the RM simulation model. For the first four studies the outcome is AUC, and for the other study the outcome is binary agreement. With respect to AUC, I find the Hillis [12] study to be the most informative because it plots the empirical type I error rates for each of the possible RM model simulation configurations. For the simulation study the numbers of abnormal and normal cases were equal (25, 50 or 100) and the median AUC values were 0.702, 0.856, and 0.961. Based on results from that study I recommended that more than 3 readers be used to avoid slightly liberal type I error rates (e.g., type I error rates of 0.07 when the nominal rate is 0.05). Five and ten readers gave good results, and thus it was my general recommendation to have at least 5 readers. If fewer readers are used, I suggested that a fixed-readers OR analysis [9] be used and that results be considered preliminary or exploratory. I conjecture that the problem with only a few readers is attributable to the reader-averaged AUC not sufficiently approximating a normal distribution, which is required for correct inference.

2.5. OR extensions and sample size estimation

Although this paper focuses only on the factorial study design having two tests with empirical AUC as the outcome, the OR model has been extended to include other balanced study designs, including various split-plot designs [17, 28, 29, 30]. Furthermore, the outcome for the OR method is not limited to a function of the ROC curve, but more generally can be any reader performance measure, such as a region-of-interest (ROI) [31], localization ROC (LROC) [32, 33, 34], or free-response ROC (FROC) [35, 36] outcome. For example, recently Chen et al [18] assess the performance of the OR method when binary agreement is the reader performance measure. Sample size estimation based on pilot data or conjectured parameter estimates is discussed by Hillis et al [11]. Using this approach, Obuchowski and Hillis [37] present sample-size tables for computer-aided detection studies using an ROI outcome. Publicly available software [38] for performing OR sample size estimation based on Reference [11] is available at http://perception/radiology.uiowa.edu/.

3. Roe and Metz (RM) model formulation

3.1. Introduction and general concepts

The RM model treats cases and readers as random samples from their respective populations, which allows for validation of MRMC analysis methods that generalize to both the case and reader populations. Let x_{ijkt} denote the DV value for test i, reader j, case k, and truth state t ($t = -$ for a normal case, + for a diseased case); i.e., x_{ijkt} represents the reader's degree of confidence that the case is diseased. I show in Section 3.6 (see equation 15) that for test i and fixed reader j the original RM model proposed by Roe and Metz [1] and the extensions of it discussed in Sections 3.4 and 3.6 provide simulated values of x_{ijkt} that comprise two random samples of nondiseased (x_{ijk-}, $k = 1, \ldots, n_-$) and diseased (x_{ijk+},$k = 1, \ldots n_+$) normally distributed continuous DV realizations.

Our interest is in deriving the relationships between the parameters of the RM and OR models when $\hat{\theta}_{ij}^{(\text{empirical})}$ is the outcome for the OR model computed from RM-model simulated x_{ijkt}. It is well known [22, 23] that for fixed test i and fixed reader j, $\hat{\theta}_{ij}^{(\text{empirical})}$ is equal to $\sum_{k=1}^{n_-}\sum_{k'=1}^{n_+} s\left(X_{ijk+} - X_{ijk'-}\right)/(n\_n_+)$, where $s(x) = 1$ if $x > 0$, $= .5$ if $x = 0$ and zero otherwise. (Note that because x_{ijkt} is continuous the probability of a tie is zero, thus equivalently we can use the definition $s(x) = 1$ if $x \geq 0$ and zero otherwise.) Furthermore, $\hat{\theta}_{ij}^{(\mathrm{empirical})}$ is an unbiased estimator of the reader-specific true AUC, given by

$$\mathrm{AUC}_{ij} = \Pr(X_{ijk+} > X_{ijk'-})$$

Thus $\xi_{ij}^{(\mathrm{empirical})} \equiv E\left(\hat{\theta}_{ij}^{(\mathrm{empirical})} \mid \text{fixed reader } j\right) = \mathrm{AUC}_{ij}$; i.e., the reader-specific expected empirical AUC is equal to the reader-specific true AUC:

$$\xi_{ij}^{(\mathrm{empirical})} = \mathrm{AUC}_{ij} \quad (7)$$

As previously noted in Section 2.1., relationship (7) does not in general hold for other AUC estimators, such as the maximum-likelihood binormal AUC estimator.

The symbol A_z will be used to denote the median reader-specific true AUC across the reader population for a particular test. It follows from (7) that A_z is also the median reader-specific expected empirical AUC across the reader population. That is,

$$\mathrm{A}_z = \underset{j}{\mathrm{median}}(\mathrm{AUC}_{ij}) = \underset{j}{\mathrm{median}}\left(\xi_{ij}^{(\mathrm{empirical})}\right) \quad (8)$$

Here I have assumed both tests have the same A_z values, as is the case for null simulations.

3.2. Original RM model

The original RM model proposed by Roe and Metz [1] is a mixed four-factor (test, reader, case, and truth) ANOVA model for X with case nested within truth; test, reader, and truth crossed; test and truth treated as fixed factors; and reader and case treated as random factors. Using their notation, the model is given by

$$X_{ijkt} = \mu_t + \tau_{it} + R_{jt} + C_{kt} + (\tau R)_{ijt} + (\tau C)_{ikt} + (RC)_{jkt} + (\tau RC)_{ijkt} + Ei_{ijkt} \quad (9)$$

$i = 1, 2$; $j = 1, \ldots, r$; $k = 1, \ldots, n_t$; $t = -, +$. Here μ_t is the effect of truth state t, τ_{it} is the interaction effect of test i and truth state t, R_{jt} is the interaction effect of reader j and truth state t, C_{kt} is the effect of case k nested within truth state t, the multiple symbols in parentheses denote interactions, and E_{ijkt} is the error term. All effects are random except for μ_t and τ_{it}. The random effects are mutually independent and normally distributed with zero means. Roe and Metz denote the corresponding variance components by σ_R^2, σ_C^2, $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$ σ_{RC}^2, $\sigma_{\tau RC}^2$ and σ_E^2. They note that $\sigma_{\tau RC}^2$ and σ_E^2 cannot be estimated separately for this model with no replications, and hence define

$$\sigma_\varepsilon^2 = \sigma_{\tau RC}^2 + \sigma_E^2$$

Although not mentioned by Roe and Metz, the omission of test, reader, and test-by-reader effects that do not depend on truth is justified by the invariance of the ROC curve to location shifts; thus inclusion of these terms would not change the ROC curve for a given reader.

Note that interactions with truth are denoted only by a t subscript in (9); in particular, the conventional-notation parenthetical truth symbol is omitted. Letting γ represent the truth factor, a more conventional notation of the model that explicitly shows the interactions of test and reader with truth and the nesting of case within truth is given by

$$X_{ijkt} = \gamma_t + (\tau\gamma)_{it} + (R\gamma)_{jt} + C_{k(t)} + (\tau R\gamma)_{ijt} + (\tau C)_{ik(t)} + (RC)_{jk(t)} + (\tau RC)_{ijk(t)} + E_{ijkt}$$

Although this model is mathematically equivalent to model (9), I prefer it because its notation differentiates between interaction and nesting effects, thus making its relationship to the typical diagnostic radiologic study design more obvious. However, for notational consistency with previous work I use the RM notation, as given by (9).

3.3. Original RM-model input values

Table 1 displays a corrected and revised listing of the original RM null-simulation input values. Note that $\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2 = 1$, which implies (see Section 3.6) that the fixed-reader nondiseased and diseased DV distributions have unit variances. For their null simulations, Roe and Metz set

$$\mu_- = \tau_{i\,-} = \tau_{i\,+} = 0, i = 1, 2 \qquad (10)$$

which implies (see Section 3.6.1) that μ_+ is the median and mean separation between the diseased and nondiseased distributions across the reader population for both tests and $A_z = \Phi\left(\mu_+/\sqrt{2}\right)$

The correlations in Table 1 are defined by $\rho_{\mathrm{WR}} = (\sigma_C^2 + \sigma_{RC}^2)/(\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2)$, $\rho_{\mathrm{BR1}} = (\sigma_C^2 + \sigma_{\tau C}^2)/(\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2)$, and $\rho_{\mathrm{BR2}} = (\sigma_C^2)/(\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2)$. For fixed readers, ρ_{WR} is the within-reader correlation between DV values for one reader and both tests: corr(X_{1jkt}, X_{2jkt}); ρ_{BR1} is the between-reader correlation between DV values for two readers and one test: corr(X_{ijkt}, $X_{ij'kt}$), $j \neq j'$; and ρ_{BR2} is the correlation between DV values for two different readers using different tests: corr (X_{1jkt}, $X_{2j'kt}$), $j \neq j'$. Note that the two DV values for each correlation are for the same randomly selected case, which can be nondiseased or diseased. We see that ρ_{WR}, ρ_{BR1} and ρ_{BR2} are correlations for the DV that are analogous to the r_1, r_2 and r_3 error correlations for the OR model error terms.

Table 1 differs from the original RM table in the following ways. First, in the original table the within-reader correlation was incorrectly computed using the formula $\rho_{\mathrm{WR}} = (\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2)/(\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2)$, which is actually the within-reader replication correlation (correlation between repeated DV values for one reader and one test).

This mistake has not been previously noted in the literature (although Dr. Charles Metz acknowledged this error in a personal communication, January 5, 2009). Values computed using the incorrect ρ_{WR} formula, which match those in the original RM table, are shown in parentheses in Table 1, with the correct values shown in the following column. Second, ρ_{BR1} in Table 1 is labeled as ρ_{BR} in the original RM table. Third, ρ_{BR2} in Table 1 is not included in the original RM table or discussed by Roe and Metz. Finally, Table 1 has twelve lines, instead of eight lines as is the original RM table, with the first six lines in Table 1 corresponding to the first two lines in the original RM table. These last two modifications were made to make it easier to compare the RM inputs with corresponding OR parameter values.

3.4. Constrained unequal-variance RM model

In practice, estimated binormal nondiseased and diseased distribution variances for a fixed reader are often different, with diseased subjects typically having more variable test results. Thus to better emulate real data, Hillis [12] modified the original RM model by allowing variance components involving case to depend on truth, with variance components involving diseased cases set equal to those involving normal cases multiplied by the factor $1/b^2$, $b > 0$. Specifically, the model is given by (9) with variance components (using an obvious notation) denoted by $\sigma_R^2, \sigma_{\tau R}^2, \sigma_{C(-)}^2, \sigma_{\tau C(-)}^2, \sigma_{RC(-)}^2, \sigma_{\varepsilon(-)}^2, \sigma_{C(+)}^2, \sigma_{\tau C(+)}^2, \sigma_{RC(+)}^2$ and $\sigma_{\varepsilon(+)}^2$, with $\sigma_{C(+)}^2 = \frac{1}{b^2}\sigma_{C(-)}^2, \sigma_{\tau C(+)}^2 = \frac{1}{b^2}\sigma_{\tau C(-)}^2, \sigma_{RC(+)}^2 = \frac{1}{b^2}\sigma_{RC(-)}^2, \sigma_{\varepsilon(+)}^2 = \frac{1}{b^2}\sigma_{\varepsilon(-)}^2$. I refer to this as the *constrained unequal-variance RM model.*

Table 2 displays the null simulation input values for the constrained unequal-variance model presented by Hillis [12], for which constraints (10) are imposed. In Section 3.6.1 I show for both tests that μ_+ is the median and mean separation between the diseased and nondiseased distributions across the reader population and $\mathrm{A}_z = \Phi\left(\mu_+/\sqrt{1 + b^{-2}}\right)$ Input values for A_z, $\sigma_{C(-)}^2, \sigma_{\tau C(-)}^2, \sigma_{RC(-)}^2$, and $\sigma_{\varepsilon(-)}^2$ are the same as for the original RM model. For a nondiseased case $\rho_{\mathrm{WR}(-)} = \left(\sigma_{C(-)}^2 + \sigma_{RC(-)}^2\right)/\left(\sigma_{C(-)}^2 + \sigma_{\tau C(-)}^2 + \sigma_{RC(-)}^2 + \sigma_{\varepsilon(-)}^2\right)$ and for a diseased case $\rho_{\mathrm{WR}(+)} = \left(\sigma_{C(+)}^2 + \sigma_{RC+}^2\right)/\left(\sigma_{C(+)}^2 + \sigma_{\tau C(+)}^2 + \sigma_{RC(+)}^2 + \sigma_{\varepsilon(+)}^2\right)$, but because the variance components differ by a factor of $1/b^2$, it follows that $\rho_{\mathrm{WR}(-)} = \rho_{\mathrm{WR}(+)} = \rho_{\mathrm{WR}}$, where ρ_{WR} is the corresponding Roe and Metz value. Similary, ρ_{BR1} and ρ_{BR2} do not depend on disease status and are equal to the corresponding Roe and Metz values.

For a fixed reader, the mean-to-sigma ratio [39, p. 96], which I denote by $\tilde{r}$, is the difference of the DV diseased and nondiseased distribution means divided by the difference of the respective standard deviations. Thus $\tilde{r} = \infty$ for the original RM model because its diseased and nondiseased standard deviations are equal. Green and Swets [39] note that $\tilde{r} \approx 4$ describes the relationship between binormal latent means and standard deviations for a variety of experiments. Accordingly, values of μ_+ and b in Table 2 were chosen by Hillis [12] such that the median $\tilde{r}$ value across readers is 4.5 (4.5 was used instead of 4.0 to avoid having more than 2.5% of the ROC curves be noticeably improper, i.e., having $|\tilde{r}| < 2.0$ [40]). Following Roe and Metz, the σ_R^2 and $\sigma_{\tau R}^2$ parameter values in Table 2 were chosen by

Hillis [12] to be equal and to result in variability of the reader-specific true AUCs similar to that resulting from the original RM-model inputs.

A somewhat similar modified RM model proposed by Abbey [14] constrains the error variances by multiplying the variance components involving case by one of two weights, one for diseased and nondiseased cases, such that the squares of the two weights sum to two. For his model, input values are chosen such that the median $\tilde{r}$ value across readers is 4.0.

3.5. Generalized RM model

Gallas and Hillis [15] modified the original RM model by allowing all of the variance components to depend on truth, with variance components corresponding to effects involving test also allowed to depend on test, without any constraints. Thus this model is more general than the original and constrained unequal-variance RM models. I refer to this as the *generalized RM model.*

3.6. Model assumed for deriving results: Unconstrained unequal-variance RM model

For deriving the relationships between the RM and OR model parameters, I allow the variance components for RM-model effects that involve case to depend on truth without imposing any constraints. Specifically, the model is given by (9) with variance components denoted by $\sigma^2_R, \sigma^2_{\tau R}, \sigma^2_{C(-)}, \sigma^2_{\tau C(-)}, \sigma^2_{RC(-)}, \sigma^2_{\varepsilon(-)}, \sigma^2_{C(+)}, \sigma^2_{\tau C(+)}, \sigma^2_{RC(+)}$ and $\sigma^2_{\varepsilon(+)}$. This model includes both the original and constrained unequal-variance RM models as special cases, but is less general than the generalized RM model. I refer to this as the *unconstrained unequal-variance RM model.* Although formulas that determine the relationships between the OR and RM-model parameter values can be derived assuming the generalized RM model, the unconstrained unequal-variance model results in simpler derivations and formulas which should satisfy the needs of most researchers. Without loss of generality I set

$$\mu_- = \tau_{1-} = \tau_{2-} = 0 \quad (11)$$

For interpreting the unconstrained unequal-variance RM model for fixed reader j using test i (hence R_{jt} and $(\tau R)_{ijt}$ are fixed), it is useful to write (9) in the form

$$X_{ijkt} = \tilde{\mu}^{(ij)}_t + \tilde{\varepsilon}^{(ij)}_{kt}$$

where

$$\tilde{\mu}^{(ij)}_t \equiv \mu_t + \tau_{it} + R_{jt} + (\tau R)_{ijt} \quad (12)$$

$$\tilde{\varepsilon}^{(ij)}_{kt} \equiv C_{kt} + (\tau C)_{ikt} + (RC)_{jkt} + (\tau RC)_{ijkt} + E_{ijkt}$$

For fixed reader *j* using test *i*, it follows that $\tilde{\mu}_t^{(ij)}$ is a constant and the $\tilde{\varepsilon}_{kt}^{(ij)}$ are independent, with

$$\tilde{\varepsilon}_{kt}^{(ij)} \sim N\left(0, \sigma^2_{\text{fixed}(t)}\right), t = -, +$$

where

$$\sigma^2_{\text{fixed}(-)} = \sigma^2_{\text{C}(-)} + \sigma^2_{\tau\text{C}(-)} + \sigma^2_{\text{RC}(-)} + \sigma^2_{\varepsilon(-)} \quad (13)$$

$$\sigma^2_{\text{fixed}(+)} = \sigma^2_{\text{C}(+)} + \sigma^2_{\tau\text{C}(+)} + \sigma^2_{\text{RC}(+)} + \sigma^2_{\varepsilon(+)} \quad (14)$$

Thus for fixed reader *j* and test *i*,

$$X_{ijkt} \sim N\left(\tilde{\mu}_t^{(ij)}, \sigma^2_{\text{fixed}(t)}\right) \quad (15)$$

and the X_{ijkt} are independent, $k = 1, \ldots, n_t$; $t = -, +$. It follows that X_{ijkt} has normal nondiseased and diseased distributions having respective means $\tilde{\mu}_-^{(ij)}$ and $\tilde{\mu}_+^{(ij)}$ and variances $\sigma^2_{\text{fixed}(-)}$ and $\sigma^2_{\text{fixed}(+)}$. Thus for test *i* and fixed reader *j* the simulated values comprise two random samples of nondiseased (x_{ijk-}, $k = 1, \ldots, n_-$) and diseased (x_{ijk+}, $k = 1, \ldots, n_+$) normally distributed DV realizations.

Define

$$W_{ij} = R_{j+} - R_{j-} + (\tau R)_{ij+} - (\tau R)_{ij-}$$

For fixed reader *j* using test *i*, it follows from (11), (12) and (15) that

$$X_{ijk+} - X_{ijk'-} \sim N\left(\mu_+ + \tau_{i+} + W_{ij}, \sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)}\right) \quad (16)$$

Across readers the W_{ij} are independent with

$$W_{ij} \sim N(0, 2(\sigma^2_{\text{R}} + \sigma^2_{TR})) \quad (17)$$

It follows from (15), (16) and (17) that the true AUC for fixed reader *j* using test *i* is

$$\mathrm{AUC}_{ij} = \Pr(X_{ijk+} > X_{ijk'-} \mid W_{ij}) = \Phi\left(\frac{\mu_{+} + \tau_{i+} + W_{ij}}{\sqrt{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}}\right) \quad (18)$$

$\mu_{+} + \tau_{i+}$ is the median and mean separation between the diseased and nondiseased distributions, and

$$\mathrm{A}_z = \Phi\left(\frac{\mu_{+} + \tau_{i+}}{\sqrt{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}}\right) \quad (19)$$

is the median reader-specific true AUC across readers for test *i*. It follows from (8) that the reader-specific expected empirical AUC is also given by the right side of (18), i.e.,

$$\xi_{ij}^{(\mathrm{empirical})} = \Phi\left(\frac{\mu_{+} + \tau_{i+} + W_{ij}}{\sqrt{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}}\right) \quad (20)$$

3.6.1. Special cases: original and constrained unequal-variance RM models—

In Table 1, $\sigma^2_{\mathrm{fixed}(-)} = \sigma^2_{\mathrm{fixed}(+)} = 1$ for the original RM model, and in Table 2, $\sigma^2_{\mathrm{fixed}(-)} = 1$ and $\sigma^2_{\mathrm{fixed}(+)} = b^{-2}\sigma^2_{\mathrm{fixed}(-)} = b^{-2}$ for the constrained unequal-variance model. Thus for fixed reader *j* and test *i* it follows from (15) that the original RM-model nondiseased and diseased DV distributions have unit variances, and the constrained unequal-variance RM model nondiseased and diseased DV distributions have variances 1 and b^{-2}, respectively. In addition, because $\tau_{1+} = \tau_{2+} = 0$ in Tables 1 and 2, it follows from (19) that $\mathrm{A}_z = \Phi(\mu/\sqrt{2})$ for the original RM model and $\mathrm{A}_z = \Phi\left(\mu_{+}/\sqrt{1 + b^{-2}}\right)$ for the constrained unequal-variance RM model.

4. Formulas for OR parameters, mean test-difference variance and expected test×reader mean square in terms of RM parameters

4.1. Derivations

In Appendices A and B (available in the online Supporting Materials) I derive formulas that express OR parameters in terms of RM parameters, assuming that the reader-performance outcome to be analyzed is the empirical AUC computed from DV data simulated from the unconstrained unequal-variance RM model discussed in Section 3.6. I refer to these as *relationship formulas*. The derivations involve straightforward calculations of expectations, variances, and covariances, treating either cases or readers or both as random. Because the empirical AUC is a *U*-statistic, its expectation, variance and covariance can be expressed as linear combinations of various DV moments (e.g., see [41, 42] for applications of *U*-statistics theory involving empirical AUC), making it possible to derive analytic relationship formulas. Although *U*-statistics theory could have been utilized in Appendix A (e.g., (A8) follows from *U*-statistics theory), results are proved directly, thereby eliminating the need for familiarity with *U*-statistics theory.

My approach is as follows: (1) For the OR model, $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3 are equal to a variance or covariance of the AUC estimates, treating reader as a fixed factor, as discussed in Section 2.2. Relationship formulas are derived by applying these definitions to the empirical AUC estimates. (2) For the OR model, $\sigma^2_{R:\mathrm{OR}}$, $\sigma^2_{TR:\mathrm{OR}}$, and $\mu_{\mathrm{OR}} + \boldsymbol{\tau}_{i:\mathrm{OR}}$ can be expressed in terms of the reader-specific expected AUC distribution using (3–5). Relationship formulas are derived by applying these definitions to the reader-specific expected empirical AUCs, defined by (20).

The derived relationship formulas are presented in Table 3. The OR model implies that $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3 are constant across readers and tests, but for empirical AUC estimates computed from RM-model simulated data these OR parameters are actually functions of the random reader and test×reader effects. Thus in Appendix A the expected values of $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3 are derived, treating readers as random. *It is these expected values that are included in* Table 3 and which are treated as the true parameter values when computing bias. Because the expected parameter values can differ by test for Cov_2 and $\sigma^2_{\varepsilon:\mathrm{OR}}$ if the test effects are not equal ($\boldsymbol{\tau}_{1+} \neq \boldsymbol{\tau}_{2+}$), the derived formulas are the average of the test 1 and 2 expected values, denoted by $\overline{E(\mathrm{Cov}_2)} \equiv \frac{1}{2}\sum_{i=1}^{2} E(\mathrm{Cov}_2 | \mathrm{test} = i)$ and $\overline{E(\sigma^2_{\varepsilon:\mathrm{OR}})} \equiv \frac{1}{2}\sum_{i=1}^{2} E(\sigma^2_{\varepsilon:\mathrm{OR}} | \mathrm{test} = i)$

Because $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3 are not constant across readers and tests, it follows that empirical AUC estimates computed from RM-model simulated data do not exactly follow the OR model. However, simulations [9, 10, 12, 17] have shown the OR model to have excellent performance for testing the hypothesis of a test difference when applied to both empirical and parametric AUC estimates computed from continuous and categorized rating data simulated using the original RM model, suggesting that the OR model is a suitable approximation for the distribution of the resulting AUC estimates. To gain further insight into this apparent robustness quality, in Supplementary Appendix C (available in the online Supporting Materials) I derive relationship formulas for the variance of the reader-averaged between-test empirical AUC difference, denoted by var $\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)$, and for the expected values of the OR reader and test×reader mean squares, denoted by E[MS(R)] and E[MS(T*R)], respectively. These relationship formulas are also included in Table 3.

4.2. OR parameter values corresponding to RM-model null-simulation parameter values

Table 4 presents the values of $\mu_{\mathrm{OR}} + \boldsymbol{\tau}_{1:\mathrm{OR}}$, $\mu_{\mathrm{OR}} + \boldsymbol{\tau}_{2:\mathrm{OR}}$, $\sigma^2_{R:\mathrm{OR}}$ and $\sigma^2_{TR:\mathrm{OR}}$, computed using the relationship formulas in Table 3, that correspond to the original RM and constrained unequal-variance RM-model null-simulation values presented in Tables 1 and 2, respectively. Numeric computations for the first line of Table 4a are presented in Table 5a. I note the following: (1) Case and reader sample sizes do not enter into the computation of these OR quantities. (2) The quantity $\mu_{\text{OR}} + \tau_{i:\text{OR}}$ is always slightly less than A_z. This is because, from (5) and (8), $\mu_{\text{OR}} + \tau_{i:\text{OR}}$ and A_z are the expected and median reader-specific expected empirical AUC, respectively, and it follows from $\mu_+ + \tau_{i+} > 0.5$, (17) and (20) that the reader-specific expected AUC distribution is skewed left. (3) The values for $\sigma^2_{R:\text{OR}}$ and $\sigma^2_{TR:\text{OR}}$ are between one and two orders of magnitude smaller than the corresponding RM parameters σ^2_R and $\sigma^2_{\tau R}$. (4) The OR parameter values are essentially the same for both models, showing that the models do not differ with respect to OR parameters that can be expressed in terms of the reader-specific expected empirical AUCs. This is not surprising since, as noted in Section 3.4, σ^2_R and σ^2_{TR} values for the constrained unequal-variance RM model were chosen specifically to result in similar variability of the reader-specific true AUCs.

Tables 6 and 7 present the values of Cov_1, Cov_2, Cov_3 $\sigma^2_{\varepsilon:\text{OR}}$ and correlations $r_1, r_2, r_3 (r_i = \text{Cov}_i / \sigma^2_{\varepsilon:\text{OR}})$ that correspond to the original and constrained unequal-variance RM-model parameter values. Numeric computations for the first line of Table 6 are presented in Table 5b. Results are given for three different case sample sizes: 10+/90−, 25+/25−, and 100+/100−, where, e.g., "10+/90−" indicates 10 diseased and 90 nondiseased cases. (For brevity, sample size 50+/50−is omitted, but conclusions from it are similar.) I note the following: (1) Reader sample size does not enter into the computation of these OR quantities. (2) In both tables the error covariances and variance depend on the sample sizes, but the correlations are relatively stable across different sample sizes for the same combination of RM parameter values, differing by no more than .02. (3) In both tables Cov_1, Cov_2, Cov_3 and their respective correlations have the same ordering as the RM correlations ρ_{WR}, ρ_{BR1}, and ρ_{BR2}. These relationships hold in general; the proof of this result is easy to show, using the covariance formulas in Table 3 and the relationship $\rho_{\text{BVN}}(x, y; \rho_1) < F_{\text{BVN}}(x, y; \rho_2)$ if $\rho_1 < \rho_2$ [43], where $F(\cdot, \cdot; \rho)$ is the standardized bivariate normal distribution function with correlation ρ. (4) Unlike the OR parameter values in Table 4 which were similar for both RM models, the OR parameter values in Table 7 differ from those in Table 6. Averaged across the 48 simulation combinations (the 36 values in Tables 6 and 7 plus the 12 values corresponding to sample size 50+/50−), for the constrained unequal-variance RM model $\sigma^2_{\varepsilon:\text{OR}}$, Cov_1, Cov_2, and Cov_3 were 18.4%, 10.8%, 10.9%, and 9.99% higher, respectively, and r_1, r_2, and r_3 were 5.6%, 5.5%, and 6.7% percent lower, respectively, than corresponding values for the original RM model. Because the reader-specific expected AUCs have similar distributions, as shown by the OR parameter values being essentially the same in Tables 4a and 4b, the difference in the magnitudes of $\sigma^2_{\varepsilon:\text{OR}}$, Cov_1, Cov_2, and Cov_3 for the two RM models can be attributed to the difference in the shapes of the ROC curves, i.e., to the difference in the mean-to-sigma ratios.

4.3. Implications for sample size estimation

From Tables 6 and 7 it is easy to see that the error variance and covariances for 100+/100− case sample size are approximately 25% of the values for 25+/25−. Similarly, the values for

50+/50− (not shown) are approximately 50% of those for 25+25−. These results suggest that for sample size estimation we assume that the error variance and covariances are approximately inversely proportional to case sample size when the abnormal-to-normal ratio is held constant. Although this assumption is not strictly correct (if it was correct the correlations would have been exactly the same for 25+/25− and 100+/100−), our results indicate that it is closely approximated.

Accordingly, sample size estimation for the OR model [11] proceeds as follows. From a pilot study one obtains estimates of the OR parameters. For estimating the power of a future study having the same abnormal-to-normal case ratio, the pilot-study error and covariance estimates are recomputed for the future study under the assumption that they are approximately inversely proportional to case sample size. In contrast, the σ_{TR}^2 and σ_R^2 estimates are not recomputed because, as shown in Table 4, they are not dependent on case or reader sample sizes.

4.4. Realism of the original Roe and Metz model parameter values

In Table 1, $\rho_{\mathrm{WR}} < \rho_{\mathrm{BR1}}$ for six of the twelve simulation parameter combinations. As a result, $r_1 < r_2$ for these same six combinations in Tables 6 and 7, with $-0.10 \le r_1 - r_2 \le -0.04$. However, Obuchowski and Rockette [20] suggest, based on clinical considerations, that $r_1 > r_2$; furthermore, for 20 reader studies reported by Rockette et al [44] the estimated differences $\left(\hat{r}_1 - \hat{r}_2\right)$ all exceed .09. Thus the original and constrained unequal-variance RM models, using the input values in Tables 1 and 2, do not realistically model the relationship between ρ_{WR} and ρ_{BR1}. This appears to be due to the mistake in computing ρ_{WR}, discussed in the Section 3.3, since we see from Table 1 that the incorrectly computed ρ_{WR} ("ρ_{WR} (incorrect)" column) always exceeds ρ_{BR1}, suggesting that Roe and Metz had intended for ρ_{WR} to exceed ρ_{BR1}.

The original RM model assumes that the diseased and nondiseased DV distributions have the same variance for each reader, resulting in binormal ROC curves that are symmetric about the negative 45° diagonal. However, as previously noted, for real data the conditional DV distributions will usually have different variances, with the variance of the diseased distribution typically being larger, resulting in an ROC curve that is not symmetric about the negative 45° diagonal. Although the Hillis [12] constrained unequal-variance RM-model parameter values were chosen to produce more typical reader ROC curves by having the median $\tilde{r}$ value equal to 4.5, its DV correlations are the same as those for the original RM model, which is why it also does not model the relationship between ρ_{WR} and ρ_{BR1} realistically. Similar comments apply to the Abbey [14] constrained unequal-variance RM model, which has a median r value of 4.0.

The realism of the $\sigma_{TR:\mathrm{OR}}^2$ and $\sigma_{R:\mathrm{OR}}^2$ values, as well as the magnitudes of the correlations, also needs to be assessed. In future research, I will make recommendations for a recalibrated RM simulation model having parameter values based on OR parameter estimates from several studies and having reader ROC curves that have typical mean-to-sigma ratios.

5. Simulation study: OR bias assessment and validation of Table 3 formulas

Continuous MRMC confidence-of-disease data were simulated using the original and constrained unequal-variance RM models for two purposes: (1) to demonstrate the usefulness of the relationship formulas for assessing bias of the OR estimators, and (2) to empirically validate the relationship formulas in Table 3.

5.1. Simulation study details

For each of the twelve parameter combinations in Tables 1 and 2, 30,000 samples were simulated for each of 9 reader-case sample size combinations resulting from pairing 3 reader-sample sizes (readers = 3, 5, and 10) with 3 case-sample sizes (25+/25−, 50+/50−, 100+/100−), resulting in a total of 12·9 = 108 different simulation settings. Samples were also simulated for the nonnull constrained unequal-variance RM model having the same inputs as in Table 2, except with $\tau_{1+} = -0.3$ and $\tau_{2+} = 0.3$. Because conclusions were similar for the three RM models, results are only reported for the nonnull constrained unequal-variance RM model. Data simulation and analysis were performed using programs written in SAS [45].

Empirical AUC reader-performance outcomes were analyzed using the OR method updated by the modifications suggested by Hillis and colleagues [7, 9, 10]. The DeLong et al [27] method was used to estimate $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3, with estimates denoted by $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}$, $\widehat{\mathrm{Cov}}_1$, $\widehat{\mathrm{Cov}}_2$, and $\widehat{\mathrm{Cov}}_3$, respectively. The quantities $\sigma^2_{R:\mathrm{OR}}$, $\sigma^2_{TR:\mathrm{OR}}$ and $\mathrm{var}\left(\widehat{\mathrm{AUC}}_{1\cdot}-\widehat{\mathrm{AUC}}_{2\cdot}\right)$ were estimated by $\hat{\sigma}^2_{R:\mathrm{OR}}=\frac{1}{2}[\mathrm{MS}(\mathrm{R})-\mathrm{MS}(\mathrm{T}*\mathrm{R})]-\widehat{\mathrm{Cov}}_1+\widehat{\mathrm{Cov}}_3$, $\hat{\sigma}^2_{TR:\mathrm{OR}}=\mathrm{MS}(\mathrm{T}*\mathrm{R})-\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}+\widehat{\mathrm{Cov}}_1+\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3$ and $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot}-\widehat{\mathrm{AUC}}_{2\cdot}\right)=\frac{2}{r}\left[\mathrm{MS}(\mathrm{T}*\mathrm{R})+r\left(\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3\right)\right]$, respectively. These are the conventional OR estimates, except that $\widehat{\mathrm{Cov}}_2-\widehat{\mathrm{Cov}}_3$ was not constrained to be nonnegative in the $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot}-\widehat{\mathrm{AUC}}_{2\cdot}\right)$ formula, as in (6), in order to simplify the explanation below for bias.

5.2. Bias results

Figure 1 presents bias results for $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}$, $\widehat{\mathrm{Cov}}_1$, $\widehat{\mathrm{Cov}}_2$, and $\widehat{\mathrm{Cov}}_3$ for the nonnull constrained unequal-variance model using 10 readers, resulting in 36 simulation combinations. The 3 pairs of (test 1, test 2) A_z values are (0.632, 0.765), (0.812, 0.892), and (0.948, 0.972), as indicated in Fig. 1. $\mathrm{Bias}\times 100/\sigma^2_{\varepsilon:\mathrm{OR}}$ is plotted for each combination, where bias = (mean OR estimate − estimand value), with the estimand and $\sigma^2_{\varepsilon:\mathrm{OR}}$ values computed using the Table 3 relationship formulas. Thus Fig. 1 displays the bias of each estimator expressed as a percent of the corresponding $\sigma^2_{\varepsilon:\mathrm{OR}}$ value, which allows for meaningful comparisons across the four estimators for a given sample size, as well as across the different case sample sizes for a given estimator, since $\sigma^2_{\varepsilon:\mathrm{OR}}$ increases as sample size increases. Also included in Fig. 1 are corresponding 95% intervals having endpoints $100\times\left(\mathrm{bias}\pm z_{.025}s/\sqrt{n}\right)/\sigma^2_{\varepsilon:\mathrm{OR}}$ where $z_{.025}$ is the

97.5th percentile of a standard normal distribution, s is the empirical standard deviation of the OR estimate, and $n = 30,000$.

I use the following approach to evaluate if an estimator is biased. For a given estimator, let Y denote the number of 95% confidence intervals that do not include zero. Under the null hypothesis of no bias for any of the 36 simulation combinations, Y follows a binomial distribution with 36 independent trials and probability of success 0.05. Hence we can test

$$\begin{aligned}\mathrm{H}_0 &: \text{ no bias for any combination}\\ \mathrm{H}_1 &: \text{ bias for at least one combination}\end{aligned}$$

by rejecting H_0 if the number of confidence intervals not including zero is large. The corresponding p-value is the probability of the number of confidence intervals not including zero being at least as large as the observed number.

From Fig. 1 I make the following observations: (1) For $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}$, all of the confidence intervals are above zero, indicating a positive bias ($p < 0.0000$). Median, minimum and maximum bias values are 1.44%, 0.5% and 4.3% of σ^2_ε, with bias appearing to increase with decreasing case sample size. (2) For $\widehat{\mathrm{Cov}}_1$ 7 confidence intervals are above zero and none below, indicating a positive bias ($p = 0.0018$). Median, minimum and maximum bias values are 0.04%, −0.08% and 0.34% of σ^2_ε, with bias appearing to increase with decreasing case sample size. (3) For $\widehat{\mathrm{Cov}}_2$ 13 confidence intervals are above zero and none below, indicating a positive bias ($p = 0.0000$). Median, minimum and maximum bias values are 0.06%, −0.07% and 0.65% of σ^2_ε, with bias appearing to increase with decreasing case sample size. (4) For $\widehat{\mathrm{Cov}}_3$, all confidence intervals include zero; thus there is not sufficient evidence to conclude the estimator is biased ($p = 1$). Median, minimum and maximum bias values are 0.01%, −0.09% and 0.19% of σ^2_ε.

The DeLong approach is based on the method of structural components proposed by Sen [46], which provides consistent estimates of the elements of the variance-covariance matrix of a vector of U-statistics. Delong et al [27, p. 840] note that their method "… turns out to be equivalent to jackknifing, but is conceptually simpler when dealing with U-statistics." Thus the positive bias observed in Fig. 1 for the DeLong error variance and Cov_1 and Cov_2 estimates is not surprising considering that, according to Efron [47, p. 21], "the jackknife variance estimate tends to be conservative in the sense that its expectation is greater than the true variance." Bandos et al [48] notes that the DeLong approach is equivalent to the two-sample jackknife approach [49, p. 2095], in contrast to the one-sample jackknife which is what is typically used with OR. For example, the freely available MRMC software *OR-DBM MRMC* [50] offers the user the choice of the jackknife, DeLong, or bootstrap option for computing the OR error variance and covariances for the empirical AUC, where the jackknife option refers to the one-sample jackknife.

Figure 2 presents bias results for $\hat{\sigma}^2_{R:\mathrm{OR}}$, $\hat{\sigma}^2_{TR:\mathrm{OR}}$, $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_1. - \widehat{\mathrm{AUC}}_2.\right)$ and MS(T*R) for the nonnull constrained unequal-variance model using 10 readers, using the same simulation setup as for Fig. 1. Percent bias and its corresponding 95% confidence interval are plotted for each combination, where percent bias = 100×(mean OR estimate − estimand value)/(estimand value), with E[MS(T*R)] being the estimand for MS(T*R). Although E[MS(T*R)] is not a parameter of interest, MS(T*R) is included to validate its expectation formula in Table 3 and to illustrate what confidence-interval results look like for an unbiased estimator.

From Figure 2 I make the following observations: (1) For $\hat{\sigma}^2_{R:\mathrm{OR}}$, 4 confidence intervals do not include zero, with 2 above and 2 below zero; thus there is not sufficient evidence to conclude $\hat{\sigma}^2_{R:\mathrm{OR}}$ is biased ($p = 0.10$). Median, minimum and maximum bias values are 0.30%, −10.3% and 2.1%. (2) For $\hat{\sigma}^2_{TR:\mathrm{OR}}$, 26 confidence intervals are below zero and none above, indicating a negative bias ($p < 0.0000$). Median, minimum and maximum bias values are −4.21%, −61.6% and 0.4%, with bias appearing to increase in magnitude with decreasing case sample size. (3) For $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_1. - \widehat{\mathrm{AUC}}_2.\right)$, 23 confidence intervals are above zero and none below, indicating a positive bias ($p < 0.0000$). Median, minimum and maximum percent bias values are 0.24%, −0.30% and 1.43%, with bias appearing to increase with decreasing case sample size. (4) For MS(T*R), only 1 confidence interval does not include zero; thus there is not sufficient evidence to conclude bias ($p = 0.84$), as expected. Median, minimum and maximum bias values are 0.04%, −0.48%, and 1.10%.

The results in Fig. 2 can be explained by the fact that any bias in $\hat{\sigma}^2_{R:\mathrm{OR}}$, $\hat{\sigma}^2_{TR:\mathrm{OR}}$ or $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_1. - \widehat{\mathrm{AUC}}_2.\right)$ is attributable to bias in the error covariance and variance estimates. This can be shown by comparing the formulas for the estimators, $\hat{\sigma}^2_R = \frac{1}{2}[\mathrm{MS(R)} - \mathrm{MS(T*R)}] - \widehat{\mathrm{Cov}}_1 + \widehat{\mathrm{Cov}}_3$, $\sigma^2_{TR:\mathrm{OR}} = \mathrm{MS(T*R)} - \hat{\sigma}^2_{\varepsilon} + \widehat{\mathrm{Cov}}_1 + \widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3$ and $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_1. - \widehat{\mathrm{AUC}}_2.\right) = \frac{2}{r}\left[\mathrm{MS(T*R)} + r\left(\widehat{\mathrm{Cov}}_2 - \widehat{\mathrm{Cov}}_3\right)\right]$, to analogous formulas for the true values given by (C21), (C20) and (C19): $\sigma^2_{R:\mathrm{OR}} = \frac{1}{2}\left\{E[\mathrm{MS(R)}] - E[\mathrm{MS(T*R)}]\right\} - E(\mathrm{Cov}_1) + E(\mathrm{Cov}_3)$, $\sigma^2_{TR:\mathrm{OR}} = E(\mathrm{MS(R)}) - \overline{E(\sigma^2_{\varepsilon:\mathrm{OR}})} + E(\mathrm{Cov}_1) + \overline{E(\mathrm{Cov}_2)} - E(\mathrm{Cov}_3)$ and $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_1. - \widehat{\mathrm{AUC}}_2.\right) = \frac{2}{r}\left\{\mathrm{E[MS(T*R)} + r\left(\overline{E(\mathrm{Cov}_2)} - E(\mathrm{Cov}_3)\right)\right\}$. Note that for these three estimators the true bias values result from replacing the mean squares by their expected values and the error variance and covariance estimates by their true values.

It follows that the biases of the three estimators are given by

$$\mathrm{bias}\left(\hat{\sigma}^2_R\right) = \mathrm{bias}\left(\widehat{\mathrm{Cov}}_3\right) - \mathrm{bias}\left(\widehat{\mathrm{Cov}}_1\right) \quad (21)$$

$$\mathrm{bias}(\hat{\sigma}^2_{TR:\mathrm{OR}}) = \mathrm{bias}\left(\widehat{\mathrm{Cov}}_1\right) + \mathrm{bias}\left(\widehat{\mathrm{Cov}}_2\right) - \mathrm{bias}\left(\widehat{\mathrm{Cov}}_3\right) - \mathrm{bias}\left(\hat{\sigma}^2_{\varepsilon}\right) \quad (22)$$

$$\mathrm{bias}\left[\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)\right] = 2\left[\mathrm{bias}\left(\widehat{\mathrm{Cov}}_2\right) - \mathrm{bias}\left(\widehat{\mathrm{Cov}}_3\right)\right] \quad (23)$$

From these equations the statistically significant negative bias of $\hat{\sigma}^2_{TR:\mathrm{OR}}$ and the positive bias of $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)$ observed in Fig. 2 can be explained as follows. From Fig. 1 we see that the bias of $\hat{\sigma}^2_{\varepsilon}$ exceeds that of each the three covariances by more than a factor of 10; thus it follows from (22) that $\hat{\sigma}^2_{TR:\mathrm{OR}}$ will be negatively biased. From Fig. 1 we see that $\widehat{\mathrm{Cov}}_2$ tends to be more positively biased then Cov_3; thus it follows from (23) that $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)$ tend to be positively biased. Figure 3 displays the observed bias of $\hat{\sigma}^2_{TR:\mathrm{OR}}$ and $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)$, computed from the simulation study, plotted against the predicted bias given by the right side of (22) and (23), respectively, but with bias estimates of the error and covariance estimates replacing the true biases. If (22) and (23) are correct, then we expect to see a regression line with intercept and slope approximately equal to zero and unity, respectively, which is approximately what we see in Fig. 3. On the other hand, because the bias of $\widehat{\mathrm{Cov}}_1$ tends to be greater than that of $\widehat{\mathrm{Cov}}_3$, it follows from (21) that $\hat{\sigma}^2_R$ should tend to be negatively biased, yet there was not sufficient evidence of bias based on the confidence intervals in Fig. 2. Further investigation, however, showed that the bias, given by the right side of (21), is quite small compared to the variability in the distribution of $\hat{\sigma}^2_R$, which explains why the resulting confidence intervals do not include zero.

Conclusions from plots analogous to Figs. 1–2 were similar when the number of readers were 3 and 5 for the nonnull constrained unequal-variance model. These plots are presented in Figures S1.1, S1.2, S2.1, and S2.2 (available in the online Supporting Materials). Conclusions based on simulated data from the original RM and null constrained equal-variance RM model were similar to those for the constrained nonnull RM model.

5.3. An unbiased method for estimating error variance and covariance estimates

It follows from (21–23) that for the empirical AUC outcome, use of unbiased error variance and covariance estimates with the OR method would result in unbiased estimates for $\sigma^2_{TR:\mathrm{OR}}$, σ^2_R and $\mathrm{var}\left(\widehat{\mathrm{AUC}}_{1\cdot} - \widehat{\mathrm{AUC}}_{2\cdot}\right)$ for rating data simulated from the unconstrained unequal-variance RM model rating data. In Appendix D (available in the online Supporting Materials) I describe how unbiased error variance and covariance estimates can be computed. Use of the OR method with unbiased error and covariance methods has not been previously investigated, and thus is an area for future research.

Of course, it could be that the OR model using unbiased error variance and covariance estimates does not perform as well, e.g., in terms of type I error, as it does using another method, such as DeLong's method or the one-sample jackknife. If this is true, one may want to consider using two error covariance estimation methods: one for making inferences and the other for estimating parameters that are needed for sizing future studies.

5.4. Empirical validation of the Table 3 formulas

To empirically validate the Table 3 formulas, I compared unbiased Monte Carlo estimates of the variance of the empirical reader-averaged AUC test difference with the predicted variance computed using the Table 3 formula, given by

$$\mathrm{var}\left(\widehat{\mathrm{AUC}}_{1}\cdot-\widehat{\mathrm{AUC}}_{2}\cdot\right)=\frac{2}{r}\left[\overline{E\left(\sigma_{\varepsilon;\,\mathrm{OR}}^{2}\right)}+\sigma_{TR:\mathrm{OR}}^{2}-E(\mathrm{Cov}_{1})+(r-1)\left(\overline{E(\mathrm{Cov}_{2})}-E(\mathrm{Cov}_{3})\right)\right]$$

For each of the 108 different simulation settings, I computed the corresponding Monte Carlo estimate of $\mathrm{var}\left(\widehat{\mathrm{AUC}}_{1}\cdot-\widehat{\mathrm{AUC}}_{2}\cdot\right)$, which is the sample variance of the empirical $\left(\widehat{\mathrm{AUC}}_{1}\cdot-\widehat{\mathrm{AUC}}_{2}\cdot\right)$. differences computed across the 30,000 simulated samples. Because the sample variance is unbiased, the expected value of the difference between the Monte Carlo and predicted values should be zero if the formulas are correct. It follows that, for a sufficiently large number of simulated samples, the mean difference between the Monte Carlo and predicted values should be approximately zero, and a fitted linear regression line fitted to a plot of Monte Carlo vs. predicted values should have slope and intercept approximately equal to one and zero, respectively. Both of these results were true for the data simulated from the nonull constrained unequal-variance RM model: the mean percentage difference (difference/predicted×100) was 0.0014%, and the fitted regression line (see Figure 5) had slope and intercept approximately equal to one and zero. Results were similar for data simulated using the original RM model and the null constrained equal-variance RM model. In addition to validating the AUC difference variance formula, these results implicitly validate the formulas for $\overline{E\left(\sigma_{\varepsilon;\,\mathrm{OR}}^{2}\right)}$, $\sigma_{TR:\mathrm{OR}}^{2}$, $E(\mathrm{Cov}_{1})$, $\overline{E(\mathrm{Cov}_{2})}$, and (Cov_{3}) that comprise its definition.

6. Example: Validation of an upper one-sided confidence bound for the OR test×reader variance component

Hillis et al [11] describe how to compute power for comparing the performance of diagnostic tests when using an OR analysis. Power computation requires values for $\sigma_{TR:\mathrm{OR}}^{2}$, $\sigma_{\varepsilon:\mathrm{OR}}^{2}$, r_1, r_2, and r_3. However, an estimate of $\sigma_{TR:\mathrm{OR}}^{2}$ acquired from pilot data may have low precision if there are only a few readers in the pilot study. Because power is inversely related to $\sigma_{TR:\mathrm{OR}}^{2}$ [11], if the value of $\sigma_{TR:\mathrm{OR}}^{2}$ is greatly underestimated actual power can be considerably lower than the nominal power. One solution to this problem is to provide a conservative estimate of $\sigma_{TR:\mathrm{OR}}^{2}$. In this example I propose a methodfor computing an upper one-sided confidence bound for $\sigma_{TR:\mathrm{OR}}^{2}$ and illustrate how the method can be validated using the RM-model relationship formulas.

Let $\zeta_{1-\alpha}$ denote a $(1-\alpha)$ 100% upper one-sided confidence bound for $\sigma^2_{TR:\mathrm{OR}}$ computed from pilot data; i.e., $\Pr(\zeta_{1-\alpha} > \sigma^2_{TR:\mathrm{OR}}) = 1-\alpha$. Note that $\zeta_{1-\alpha}$ is a random variable. Under the assumptions of the OR model, it has been shown [9, 30] that the distribution of the OR test-by-reader mean square is given by

$$\mathrm{MS(T*R)} \sim \chi^2_{(t-1)(r-1)} \frac{E[\mathrm{MS(T*R)}]}{(t-1)(r-1)} \quad (24)$$

where

$$E[\mathrm{MS(T*R)}] = \sigma^2_{TR:\mathrm{OR}} + \sigma^2_{\varepsilon:\mathrm{OR}}(1 - r_1 - r_2 + r_3) \quad (25)$$

and t and r are the number of tests and readers, respectively. It follows from (24) and (25) that

$$\Pr\left(\sigma^2_{TR:\mathrm{OR}} < \frac{\mathrm{MS(T*R)}(t-1)(r-1)}{\chi^2_{\alpha;(t-1)(r-1)}} - \sigma^2_{\varepsilon:\mathrm{OR}}(1 - r_1 - r_2 + r_3)\right) = 1-\alpha$$

where $\chi^2_{\alpha;(t-1)(r-1)}$ is the α (100)th percentile of a $\chi^2_{(t-1)(r-1)}$ distribution. Thus a $(1-\alpha)$ 100% upper one-sided confidence bound for $\sigma^2_{TR:\mathrm{OR}}$ is given by

$$\zeta_{1-\alpha} = \frac{\mathrm{MS(T*R)}(t-1)(r-1)}{\chi^2_{\alpha;(t-1)(r-1)}} - \sigma^2_{\varepsilon:\mathrm{OR}}(1 - r_1 - r_2 + r_3)$$

I propose using

$$\hat{\zeta}_{1-\alpha} = \frac{\mathrm{MS(T*R)}(t-1)(r-1)}{\chi^2_{\alpha;(t-1)(r-1)}} - \hat{\sigma}^2_{\varepsilon:\mathrm{OR}}(1 - \hat{r}_1 - \hat{r}_2 + \hat{r}_3) \quad (26)$$

or equivalently,

$$\hat{\zeta}_{1-\alpha} = \frac{\mathrm{MS(T*R)}(t-1)(r-1)}{\chi^2_{\alpha;(t-1)(r-1)}} - \left(\hat{\sigma}^2_{\varepsilon:\mathrm{OR}} - \widehat{\mathrm{Cov}}_1 - \widehat{\mathrm{Cov}}_2 + \widehat{\mathrm{Cov}}_3\right)$$

as an estimate of $\zeta_{1-\alpha}$, where MS(T*R), $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}$, $\hat{r}_1$, $\hat{r}_2$, and $\hat{r}_3$ (or $\widehat{\mathrm{Cov}}_1$, $\widehat{\mathrm{Cov}}_2$, and $\widehat{\mathrm{Cov}}_3$) are computed from pilot data.

To validate the proposed estimator, I compute $\hat{\zeta}_{1-\alpha}$ for each of the samples simulated from the original and constrained unequal-variance RM-model simulation studies described in

Section 5. For each input combination I compute the empirical coverage: the proportion of the 30,000 samples such that $\hat{\zeta}_{1-\alpha}$ exceeds the true value of $\sigma^2_{TR:\mathrm{OR}}$, computed using the formula in Table 3. For each of the upper bounds $\hat{\zeta}_{.70}$, $\hat{\zeta}_{.80}$, and $\hat{\zeta}_{.90}$, Fig. 4 presents results for the nonnull constrained unequal-variance model for 108 simulation combinations; these are the same combinations included in Figs. 1 and 2, except that now all 3 reader levels (3, 5, and 10 readers) are included. Empirical coverage is plotted for each combination.

From Fig. 4 we see that empirical coverage tends to be less than the nominal confidence level. For the A$_z$ pairs (0.632, 0.765) and (0.812, 0.892), all except one of the empirical coverages for all 3 upper bounds is between the nominal level and nominal − 0.05. For example, for $\hat{\zeta}_{0.70}$, 71 of 72 empirical coverages are between 0.65 and 0.70, with the one exception being 0.64; for $\hat{\zeta}_{0.80}$, all 72 coverages are between 0.75 and 0.80, and for $\hat{\zeta}_{0.90}$, all 72 coverages are between 0.85 and 0.90. For the A$_z$ pair (0.948, 0.972), results are similar for the two low reader-variance structures HL and LL: all but one of the empirical coverages are between the nominal level and nominal − 0.05. However, for the two high reader-variance structures, HH and LH, the empirical coverages are spread between the nominal level and nominal − 0.15.

That the empirical-versus-nominal discrepancy is much greater for the combination of high A$_z$ and high reader variability is not surprising, since the OR assumption of normality of the reader AUC estimates will not be as well approximated for large A$_z$. These results suggest that except for the high A$_z$ and high reader variability situation, in practice this approach is acceptable, with nominal confidence levels slightly overstating the true level. Conclusions were similar for the original and null unequal-variance RM-model simulations (not shown).

6.1. Real-data illustration

To illustrate use of the proposed method, I revisit an example of sample size estimation given by Hillis et al [11]. They perform an OR analysis of data from a study [51] that compares the relative performance of single spin-echo magnetic resonance imaging (MRI) to cinematic presentation of MRI for the detection of thoracic aortic dissection. Five radiologists independently interpreted 114 images using a 5-point ordinal scale. With the binormal likelihood ratio (also known as “bi-chi-squared” or “proper binormal”) [52, 53, 54] AUC as the outcome, the OR analysis results in MS(T*R) = 0.000623, $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}} = 0.00139$, $\hat{r}_1 = 0.252$, $\hat{r}_2 = 0.249$, $\hat{r}_3 = 0.159$ and $\hat{\sigma}^2_{TR:\mathrm{OR}} = -0.00294$. Averaged across readers, the two modality AUCs are 0.910 and 0.950, with $p = .092$ for testing the null hypothesis of equivalence.

They perform sample size computations treating these data as pilot data. Because $\hat{\sigma}^2_{TR:\mathrm{OR}}$ is negative, they use $\sigma^2_{TR:\mathrm{OR}} = 0$ for the computations. They note, however, that it seems reasonable that there should be some test-by-reader interaction. A solution that avoids this problem is to set $\sigma^2_{TR:\mathrm{OR}}$ equal to the estimated upper one-sided confidence bound obtained from the pilot data. For example, suppose it is desired to use an upper 75% one-sided confidence bound. Using (26) with $t = 2$, $r = 5$, $\alpha = 0.25$ yields $\hat{\zeta}_{0.75} = 0.000379$. Although the average AUC estimate (0.93) is high, the negative value of $\sigma^2_{TR:\mathrm{OR}}$ suggests that reader variability is not high – thus I have not adjusted the nominal level. For comparison, to obtain 80% power to detect an AUC difference of .05 requires 7 readers and 365 cases if $\sigma^2_{TR:\mathrm{OR}} = 0.000379$, whereas 7 readers and 200 cases are required if $\sigma^2_{TR:\mathrm{OR}} = 0.0$. (These sample-size results are based on the methodology in Reference [11] and were obtained using sample-size software created by Hillis and Schartz [38], which is freely available at http://perception/radiology.uiowa.edu/).

6.2. Further comments

Although use of the upper one-sided confidence bound was useful in the preceding example for avoiding a negative $\sigma^2_{TR:\mathrm{OR}}$ estimate, I recommend using it also with positive estimates to provide more confidence that the actual power is at least as great as the nominal power. A limitation is that the simulation study only specifically provided validation for the empirical AUC estimate for continuous DV data. Although it seems reasonable that the method should perform similarly with discrete rating data and other types of AUC estimators, it would be useful to establish this empirically in future research.

This example illustrates the importance of the Table 3 formulas: without the formula for the true value of σ^2_{TR} we would not have been able to evaluate the performance of the upper one-sided confidence bound approach. Note that the true value of σ^2_{TR} cannot be estimated empirically by the empirical mean of $\hat{\sigma}^2_{TR}$ unless it is known that $\hat{\sigma}^2_{TR}$ is an unbiased estimator of σ^2_{TR}. The Table 3 formulas made it possible in the previous section to determine that $\hat{\sigma}^2_{TR}$ is unbiased only if the error covariance and covariance estimates are unbiased.

7. Discussion

The RM simulation model has been a valuable tool for empirically validating various MRMC analysis methods for diagnostic radiologic studies. However, because its parameters are defined in terms of a case-level continuous DV, it has never been clear how closely it emulates real-data studies where reader performance outcomes, such as the AUC, are analyzed. In particular, it has not been possible to know the exact relationship between the RM-model parameters and those of the Obuchowski-Rockette and related Dorfman-Berbaum-Metz analysis methods, which are the most frequently used methods for analysis of diagnostic reader performance outcomes that generalize to both the case and reader populations.

The primary contribution of this paper is Table 3, which provides analytic formulas that relate the RM-model parameters to corresponding OR model parameters when the reader performance outcome is the empirical AUC. Thus any specification of RM-model parameter values can be interpreted in terms of corresponding OR parameter values. Because the OR model parameters have a one-to-one relationship with the DBM model parameters and have more intuitive interpretations, only results for the OR model were presented. The focus of this paper was the empirical AUC for two reasons: (1) it is a frequently used reader-performance measure and (2) it results in analytic relationship formulas. This second reason is why the empirical AUC was chosen instead of the frequently used AUC estimators resulting from maximum likelihood estimation under the assumption of a latent binormal model [55, 56] or a latent binormal likelihood ratio model [52, 53, 54].

Although the RM model generates DV data according to a binormal model for each test-reader combination, the Table 3 relationships hold for any strictly increasing transformation of the data, which can include distribution pairs that do not remotely resemble normal distributions and can be quite skewed. This is because the empirical ROC curve, and hence the empirical AUC, is a function only of the ranks of the data. For example, Hanley [57] illustrates how many pairs of distributions having shapes that appear to be quite nonnormal can be made close to binormal by a suitable increasing transformation.

The approach taken for deriving the Table 3 relationships was to express each OR parameter in terms of the distribution of either the AUC estimates or the reader-specific expected AUCs, and then apply these definitions to the RM-model empirical AUC estimates and reader-specific expected empirical AUCs. This approach differs from previous work by Gallas and Hillis [15] that focused only on deriving the means, variances, and covariances of the empirical AUC or AUC difference in terms of the RM-model parameters. Although the OR model implies that the fixed-reader variance and covariances are the same for each reader, they actually vary across readers for RM-model simulated data; thus the expected values of these parameters are considered to be the true values. Alternatively, I could have taken the approach of defining the true values for Cov_2 and Cov_3 using the relationships $\underset{j \ne j'}{\mathrm{cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right) = \mathrm{Cov}_2$ and $\underset{i \ne i', j \ne j}{\mathrm{cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}\right) = \mathrm{Cov}_3$ (note that here readers are treated as random), which are also implied by the OR model. Not surprisingly, these two approaches give equivalent formulas in terms of the RM model, as shown by (C4) and (C5).

An advantage of Table 3 is the savings in time that results from not having to obtain approximate true parameter values through simulations. But more importantly, Table 3 makes it *possible* to obtain true values for some of the OR parameters for which approximate true values cannot be obtained from covariance analysis of the simulated reader performance outcomes. For example, the OR model implies $\mathrm{var}\left(\hat{\theta}_{ij}\right) = \sigma^2_{R:\mathrm{OR}} + \sigma^2_{TR:\mathrm{OR}} + \sigma^2_{\varepsilon:\mathrm{OR}}$, $\underset{j \ne j'}{\mathrm{cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right) = \mathrm{Cov}_2$, $\underset{i \ne i'}{\mathrm{cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j}\right) = \sigma^2_{R:\mathrm{OR}} + \mathrm{Cov}_1$, and $\underset{i \ne i', j \ne j}{\mathrm{cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}\right) = \mathrm{Cov}_3$. Hence although a multivariate covariance matrix computed from test × reader empirical AUC vectors can provide unbiased estimates for Cov_2 and Cov_3, it cannot provide unbiased estimates for $\sigma^2_{R:\mathrm{OR}}$, $\sigma^2_{TR:\mathrm{OR}}$, $\sigma^2_{\varepsilon:\mathrm{OR}}$, or Cov_1.

Table 3 offers many other advantages. It makes it possible to easily check for proper implementation of the RM simulation model by checking if estimates of the OR parameters approximately agree with the true values. It makes it possible to assess the realism of the RM model by comparing the corresponding OR parameter values with those from real-data sets; for example, we saw that the OR error covariances and correlations were not ordered as usually observed for real-data sets for half of the RM input combinations in Table 6, showing a lack of realism. We also saw how Table 3 makes it possible to not only assess the bias of OR parameter estimates, but also to explain the source of any bias.

Table 3 made it possible to assess the performance of the proposed upper one-sided confidence bound for the OR test × reader variance component. Many other performance assessments can similarly be made which previously were not possible. For instance, although in this paper I only used the DeLong et al [27] method for computing the error variance and covariances, I could have used other methods, such as the one-sample jackknife, the bootstrap, and the unbiased method discussed in Sect. 5.3. Because the true values of the error variance and covariances can be computed using the Table 3 formulas, these four methods can easily be compared with respect to bias and mean squared error.

Biases of the OR estimate of the variance for the reader-averaged test difference of the 36 empirical AUC estimates in Fig. 2 (for 10 readers) were relatively small, ranging between −0.30% and 1.43%. The range is similar when estimates for 3 and 5 readers are included: −0.87% to 1.43%. To illustrate that this level bias will have little effect on inferences, consider the situation where an unbiased variance estimate results in a *p*-value of 0.05 using a large sample *z* test. An increase in the variance estimate of 1.43% would result in *p* =0.0516; when degrees of freedom are accounted for using a *t* test the change will be even less. This is important because it demonstrates how the OR method can perform well even when some of its assumptions, such as constant fixed-reader variance and covariances, do not strictly apply to the outcomes. Furthermore, we saw that this variance bias could be eliminated by using unbiased error variance and covariance estimates, at least for rating data simulated from the unconstrained equal-variance RM model, although this is an approach needs to be further investigated.

Biases of the $\hat{\sigma}^2_{TR:\mathrm{OR}}$ estimates in Fig. 2 ranged between −61.6% and 0.4%; when estimates for 3 and 5 readers are included, the range is similar: −61.6% to 1.4%. This range of bias is of concern for sample sizing purposes: because power is inversely related to $\hat{\sigma}^2_{TR:\mathrm{OR}}$, negative bias will result in overly optimistic sample size estimates. The bias is attributable to biases in the $\hat{\sigma}^2_{\varepsilon:\mathrm{OR}}, \widehat{\mathrm{Cov}}_1, \widehat{\mathrm{Cov}}_2$, and $\widehat{\mathrm{Cov}}_3$ estimates and is less for larger case samples, with biases for 100 + /100 ranging between −4.3% and 1.4% across the 36 estimates which include readers = 3, 5, and 10. Thus I make the following recommendations for pilot studies. First, having 100 + /100− cases is acceptable for estimating $\hat{\sigma}^2_{TR:\mathrm{OR}}$. Second, for the typical situation when the number of cases is less than 100 + /100, using an unbiased method for estimating $\sigma^2_{\varepsilon:\mathrm{OR}}$, Cov_1, Cov_2, and Cov_3, as discussed in Sect. 5.3 for the empirical AUC, will result in unbiased estimates of $\sigma^2_{TR:\mathrm{OR}}$. Another option is to use the proposed upper one-sided confidence bound in place of $\hat{\sigma}^2_{TR:\mathrm{OR}}$; however, the confidence level will diverge more from the nominal level as $\hat{\sigma}^2_{TR:\mathrm{OR}}$ becomes more biased. These two options (unbiased error covariances and the one-sided confidence bound) could be used together – this is an area for future research.

Gallas et al [58, 59] proposed an alternative approach that is often used for computing unbiased estimates of the variances of the reader-averaged test and test-difference estimates. In this approach, variances are expressed as linear combinations of product moments of functions of the ratings. The moments are then replaced by corresponding unbiased *U*-statistic estimates, resulting in unbiased variance estimates. A limitation of this approach is that the reader-performance outcome must be a *U*-statistic, such as the empirical AUC. In contrast, the OR method is applicable to all reader performance outcomes and is essentially a marginal ANOVA model, as discussed by Hillis [30].

Table 3 makes it possible to calibrate the RM model so that its corresponding OR parameters are similar to those resulting from OR analyses of real data sets. This will be done in future research. In addition to having OR parameter values similar to those for real data sets, to be realistic the RM model should also be based on underlying ROC curves similar in shape to those for real data sets, which can be accomplished by using a similar mean-to-sigma ratio. Although such calibration can be done by “trial and error,” ideally the calibration process could be accomplished through use of an OR-to-RM mapping algorithm which computes RM-model parameter values that correspond to specified OR model parameter values; this is an area for future research.

It would be extremely helpful if researchers, when reporting results from OR analyses, would include the estimated OR parameter values for $\sigma^2_{R:\mathrm{OR}}$, $\sigma^2_{TR:\mathrm{OR}}$, $\sigma^2_{\varepsilon:\mathrm{OR}}$, plus either r_1, r_2, r_3 or Cov_1, Cov_2, Cov_3; and similarly report estimates of all of the DBM parameters if a DBM analysis is performed. Estimates of these parameters are printed out in some MRMC software packages, such as the stand-alone multireader ROC software *OR-DBMMRMC 2.5* [50], as well as the program *OR/DBM MRMC 3.0 for SAS* [60], both freely available at http://perception.radiology.uiowa.edu/. Reporting these estimates would provide additional information for calibrating the RM model, as well as for conjecturing parameter values needed for power and sample size estimation.

Limitations of the Table 3 relationships are the empirical AUC outcome and continuous RM-model DV data requirements. Knowledge of relationship formulas for discrete-rating DV data would allow us to determine how the OR parameter values change for a given set of RM-model input parameters when the continuous DV is transformed to a discrete ordinal DV. This would be useful for comparing the performance of the OR method and sample size requirements for continuous and various rating categorizations. Extension of Table 3 relationships to include other types of estimators (such as the binormal AUC) would allow for comparison of different types of estimators. These are both areas for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under Award Numbers R01EB013667 and R01EB019967. I thank Carolyn Van Dyke, M.D. for sharing her data set.

References

1. 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. DOI: 10.1016/S1076-6332(97)80032-3 [PubMed: 9110028]
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. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects, receiver operating characteristic analysis. Academic Radiology. 2000; 7(5):341–9. [PubMed: 10803614]
4. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Academic Radiology. 2001; 8(7):616–22. [PubMed: 11450962]
5. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. Academic Radiology. 2001; 8(4):328–34. [PubMed: 11293781]
6. Song X, Zhou XH. A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. Biostat. 2005; 6(2):303–12.
7. 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]
8. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Academic Radiology. 2005; 12:1534–1541. DOI: 10.1016/j.acra.2005.07.012 [PubMed: 16321742]
9. 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]
10. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Academic Radiology. 2008; 15:647–661. DOI: 10.1016/j.acra.2007.12.015 [PubMed: 18423323]
11. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Academic Radiology. 2011 Feb; 18(2):129–42. [PubMed: 21232681]
12. Hillis SL. Simulation of unequal-variance binormal multireader ROC decision data: An extension of the Roe and Metz simulation model. Academic Radiology. 2012; 19:1518–1528. DOI: 10.1016/j.acra.2012.09.011 [PubMed: 23122571]
13. Chakraborty DP. Prediction Accuracy of a Sample-size Estimation Method for ROC Studies. Academic Radiology. 2010; 17:628–638. DOI: 10.1016/j.acra.2010.01.007 [PubMed: 20380980]
14. Abbey CK, Samuelson FW, Gallas BD. Statistical Power Considerations for a Utility Endpoint in Observer Performance Studies. Academic Radiology. 2013; 20(7):798–806. DOI: 10.1016/j.acra.2013.02.008 [PubMed: 23611439]
15. Gallas BD, Hillis SL. Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances. Journal of Medical Imaging. 2014; 1(3):031006–031006. DOI: 10.1117/1.JMI.1.3.031006 [PubMed: 26158048]
16. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis for binary data. Journal of the Optical Society of America A. 2007; 24(12):B70–B80. DOI: 10.1364/JOSAA.24.000B70

17. Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC Studies with Split-plot Designs: A Comparison of Statistical Methods. Academic Radiology. 2012; 19(12):1508–1517. DOI: 10.1016/j.acra.2012.09.012 [PubMed: 23122570]

18. Chen W, Wunderlich A, Petrick N, Gallas BD. Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing. Journal of Medical Imaging. 2014; 1(3):031011–031011. DOI: 10.1117/1.JMI.1.3.031011 [PubMed: 26158051]

19. Metz, C., Wang, P-L., Kronman, H. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck, F., editor. Information Processing in Medical Imaging. Springer; Netherlands: 1984. p. 432-445.

20. 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.

21. 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]

22. Bamber D. Area above ordinal dominance graph and area below receiver operating characteristic graph. Journal of Mathematical Psychology. 1975; 12(4):387–415.

23. 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]

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

25. Shao, J., Dongshen, T. The Jackknife and Bootstrap. New York: Springer-Verlag; 1995.

26. Efron, B., Tibshirani, RJ. An introduction to the bootstrap. New York: Chapman and Hall; 1993.

27. 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(3):837–845. [PubMed: 3203132]

28. Obuchowski NA. Multireader receiver operating characteristic studies: A comparison of study designs. Academic Radiology. 1995; 2(8):709–716. DOI: 10.1016/S1076-6332(05)80441-6 [PubMed: 9419629]

29. Obuchowski NA. Reducing the Number of Reader Interpretations in MRMC Studies. Academic Radiology. 2009; 16(2):209–217. [PubMed: 19124107]

30. Hillis SL. A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. Statistics in Medicine. 2014; 33:330–360. DOI: 10.1002/sim.5926 [PubMed: 24038071]

31. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. Academic Radiology. 2000; 7(7):516–525. [PubMed: 10902960]

32. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images 1. Radiology. 1975; 116(3):533–538. [PubMed: 1153755]

33. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Medical Physics. 1996; 23(10):1709–1725. DOI: 10.1118/1.597758 [PubMed: 8946368]

34. Popescu LM. Nonparametric ROC and LROC analysis. Medical Physics. 2007; 34(5):1556–1564. DOI: 10.1118/1.2717407 [PubMed: 17555237]

35. Chakraborty DP, Winter LHL. Free-response methodology - alternate analysis and a new observer-performance experiment. Radiology. 1990; 174(3):873–881. [PubMed: 2305073]

36. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis, and validation. Medical Physics. 2004; 31(8):2313–2330. [PubMed: 15377098]

37. Obuchowski NA, Hillis SL. Sample Size Tables for Computer-Aided Detection Studies. American Journal of Roentgenology. 2011; 197(5):W821–W828. DOI: 10.2214/AJR.11.6764 [PubMed: 22021528]

38. Hillis SL, Schartz KM. Demonstration of multi- and single-reader sample size program for diagnostic studies software. Proc SPIE 9416, Medical Imaging 2015: Image Perception, Observer

Performance, and Technology Assessment. 2015; 94160E. March 17, 2015. doi: 10.1117/12.2083150

39. Green, DM., Swets, JA. Signal detection theory and psychophysics. Peninsula Publishing; Los Altos: 1988. Original work: Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley, 1966.
40. Hillis SL, Berbaum KS. Using the mean-to-sigma ratio as a measure of the improperness of binormal ROC curves. Academic Radiology. 2011; 18:143–154. [PubMed: 21232682]
41. Yousef WA, Wagner RF, Loew MH. Assessing Classifiers from Two Independent Data Sets Using ROC Analysis: A Nonparametric Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006; 28(11):1809–1817. DOI: 10.1109/TPAMI.2006.218 [PubMed: 17063685]
42. Chen W, Gallas BD, Yousef WA. Classifier variability: Accounting for training and testing. Pattern Recognition. 2012; 45(7):2661–2671. DOI: 10.1016/j.patcog.2011.12.024
43. Tihansky DP. Properties of the bivariate normal cumulative distribution. Journal of the American Statistical Association. 1972; 67(340):903–905.
44. Rockette HE, Campbell WL, Britton CA, Holbert JM, King JL, Gur D. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. Academic Radiology. 1999; 6(12):723–729. [PubMed: 10887893]
45. SAS for Windows, Version 9.3, copyright (c) 2002-2010 by SAS Institute Inc., Cary, NC, USA.
46. Sen PK. On Some Convergence Properties of UStatistics. Calcutta Statistical Association Bulletin. 1960; 10(1-2):1–18.
47. Efron B. The jackknife, the bootstrap and other resampling plans: SIAM. 1982
48. Bandos AI, Rockette HE, Gur D. Resampling methods for the area under the ROC curve. ROC Analysis in Machine Learning. 2006:1–8.
49. Arvesen JN. Jackknifing U-Statistics. Annals of Mathematical Statistics. 1969; 40(6):2076–2100.
50. Schartz, KM., Hillis, SL., Pesce, LL., Berbaum, KS. OR-DBM MRMC (Version 2.5) [Computer software]. Available for download from http://perception.radiology.uiowa.edu. Accessed July 22, 2015
51. 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;
52. 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]
53. Metz CE, Pan XC. “Proper” binormal ROC curves: theory and maximum-likelihood estimation. Journal of Mathematical Psychology. 1999; 43:1–33. [PubMed: 10069933]
54. Hillis SL. Equivalence of binormal likelihood-ratio and bi-chi-squared ROC curve models. Statistics in Medicine. 2016; 35(12):2031–2057. DOI: 10.1002/sim.6816 [PubMed: 26608405]
55. 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.
56. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in Medicine. 1998; 17(9):1033–1053. [PubMed: 9612889]
57. Hanley JA. The use of the ‘binormal’ model for parametric ROC analysis of quantitative diagnostic tests. Statistics in Medicine. 1996; 15(14):1575–1585. [PubMed: 8855483]
58. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. Academic Radiology. 2006; 13(3):353–362. DOI: 10.1016/j.acra.2005.11.030 [PubMed: 16488848]
59. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators. Communications in Statistics - Theory and Methods. 2009; 38(15):2586–2603. DOI: 10.1080/03610920802610084
60. Hillis, SL., Schartz, KM., Berbaum, KS. OR/DBM MRMC for SAS (Version 3.0) [Computer software]. Available for download from http://perception.radiology.uiowa.edu. Accessed July 22, 2015.

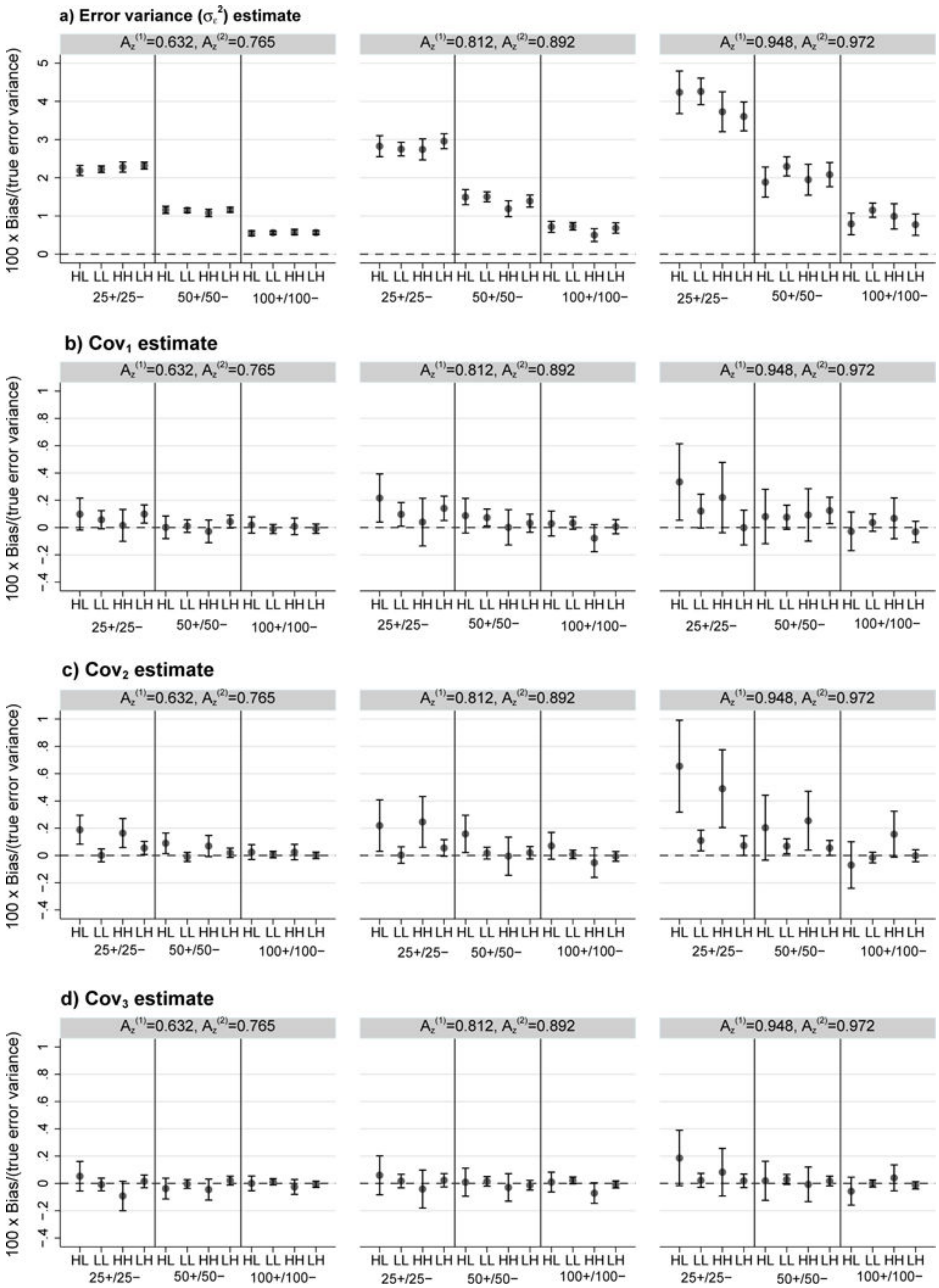

Figure 1.

Bias of Obuchowski-Rockette error variance and covariance estimates, computed using the Delong et al [27] method, expressed as a percent of the true error variance $\sigma^2_{\varepsilon:\mathrm{OR}}$. Outcome is the empirical AUC. Results are computed from data simulated from the constrained nonnull unequal-variance RM model, with 30,000 samples simulated for each of $N = 36$ input combinations: 3 sample sizes (25+/25−, 50+/50−, 100+,100−) × 4 structures(HL. LL, HH, LH) × 3 nonnull curve pairs (with $A_z^{(1)}$ and $A_z^{(2)}$ denoting the test 1 and test 2 values.)

Ten readers are used for each combination. The parameter values are the same as given in Table 2, except that to produce nonnull simulations the μ_+ values are replaced by $\mu_+ -0.3$ and $\mu_+ +0.3$ for tests 1 and 2 and A_z values replaced by corresponding $A_z^{(1)} = \Phi\left[\left(\mu_+ - 0.3\right)/\sqrt{1 + b^{-2}}\right]$ and $A_z^{(2)} = \Phi\left[\left(\mu_+ + 0.3\right)/\sqrt{1 + b^{-2}}\right]$ values, respectively. Bias is computed with respect to the corresponding parameters in Table 3. Error bars indicate corresponding large sample 95% confidence intervals.

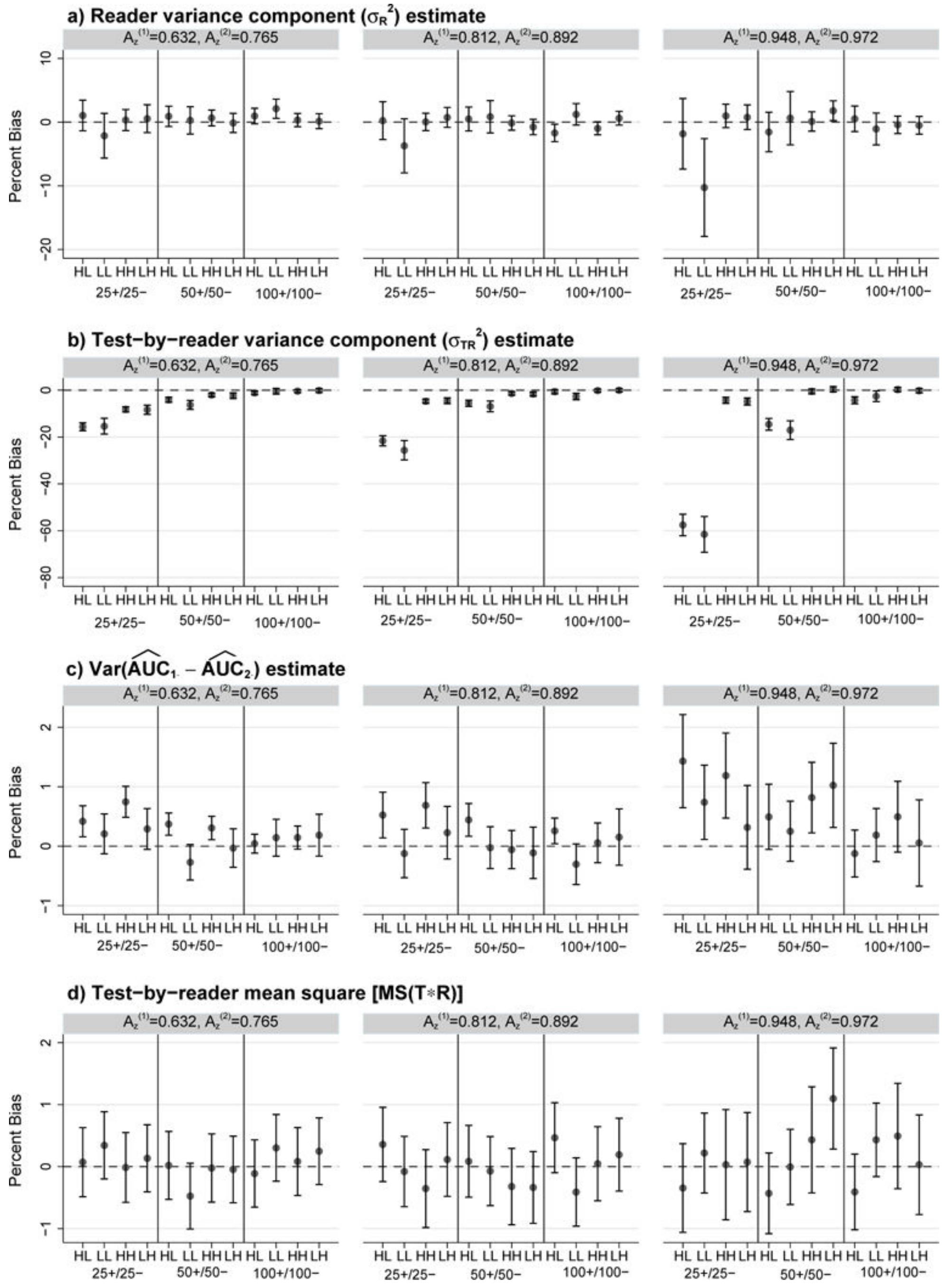

Figure 2.

Percent bias of Obuchowski-Rockette estimates for the reader variance component $\sigma^2_{R:\text{OR}}$, test×reader variance component $\sigma^2_{TR:\text{OR}}$, $\text{var}\left(\widehat{\text{AUC}}_{1\,\bullet} - \widehat{\text{AUC}}_{2\,\bullet}\right)$ and $E\,[\text{MS (T*R)}]$ based on data simulated from the constrained nonnull unequal-variance RM model as described in Figure 1. Ten readers are used for each combination. The OR estimates are defined by $\hat{\sigma}^2_{R:\text{OR}} = \frac{1}{2}[\text{MS(R)} - \text{MS(T} * \text{R)}] - \widehat{\text{Cov}}_1 + \widehat{\text{Cov}}_3$; $\hat{\sigma}^2_{TR:\text{OR}} = \text{MS(T}*\text{R)} - \hat{\sigma}^2_{\varepsilon} + \widehat{\text{Cov}}_1 + \widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3$;

and $\widehat{\text{var}}\left(\widehat{\text{AUC}}_{1} \cdot - \widehat{\text{AUC}}_{2} \cdot\right) = \frac{2}{r}\left[\text{MS}(\text{T}*\text{R}) + r\left(\widehat{\text{Cov}}_2 - \widehat{\text{Cov}}_3\right)\right]$, where MS (R) and MS (T*R) are the OR reader and test×reader mean squares, respectively. Bias is computed with respect to the corresponding parameters in Table 3. Percent bias = 100×bias/ (estimand value).

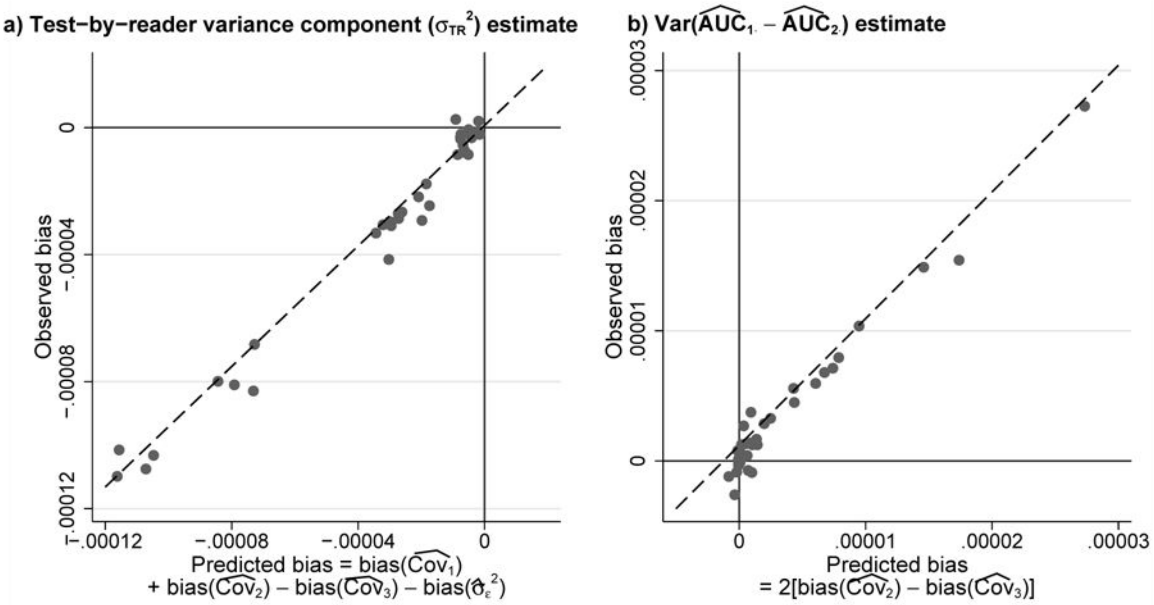

Figure 3.

Observed versus predicted bias for the Obuchowski-Rockette test×reader variance component estimate, $\hat{\sigma}_{\varepsilon}^{2}$, and the variance of the difference of the empirical reader-averaged AUC estimate, $\widehat{\mathrm{var}}\left(\widehat{\mathrm{AUC}}_{1}\cdot-\widehat{\mathrm{AUC}}_{2}\cdot\right)$, based on data simulated from the constrained nonnull unequal-variance RM model as described in Figure 1. Observed bias = (mean estimate – estimand), where the estimand is computed from the Table 3 formulas. The predicted bias is a function of the error variance and covariances, as given by (22–23), but with simulation-study estimates of the error-variance and covariance biases used in place of the true bias values.

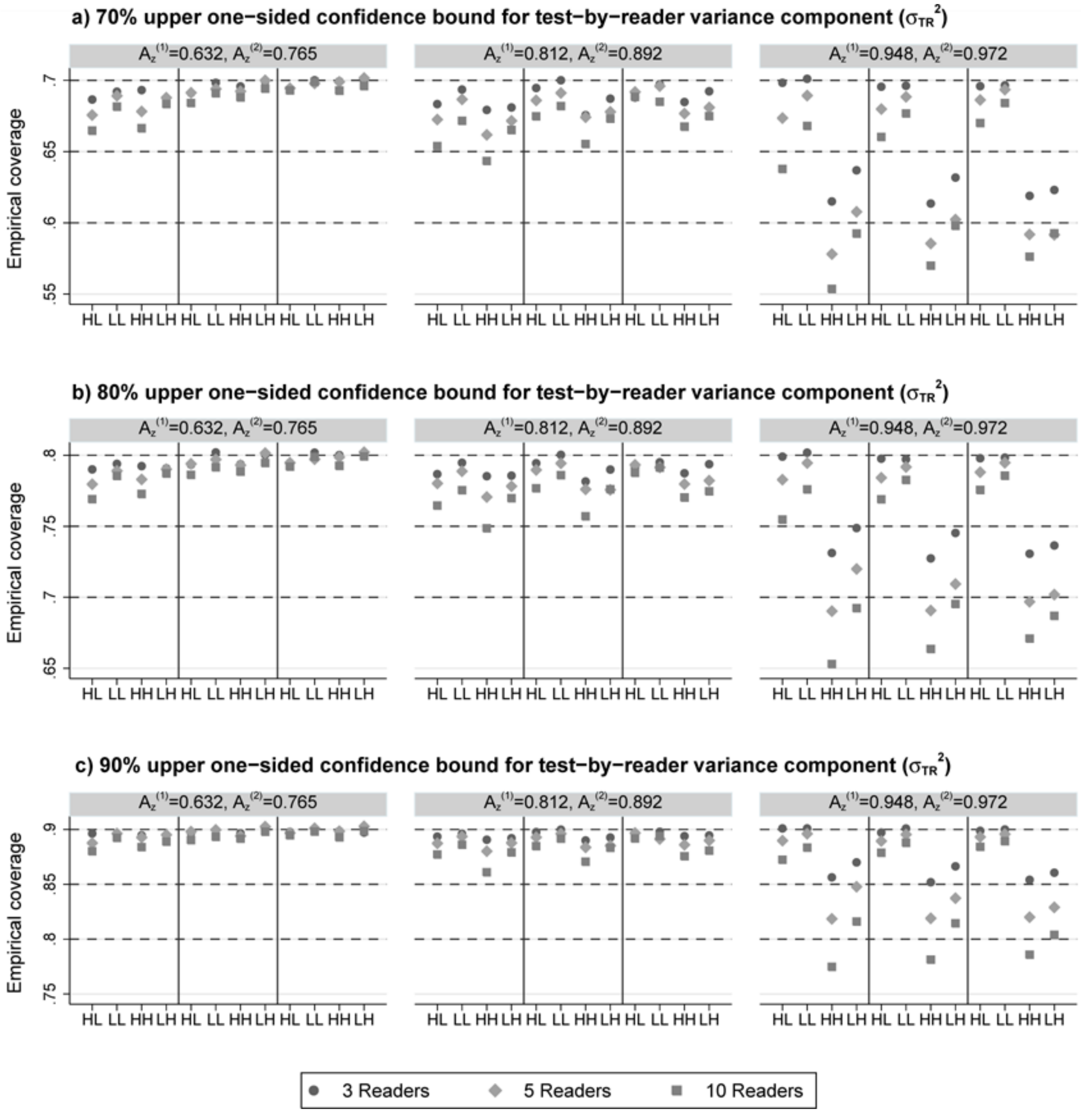

Figure 4.

Empirical coverage of the proposed upper one-sided confidence bound for the Obuchowski-Rockette test-by-reader variance component $\sigma^2_{TR:\mathrm{OR}}$, based on data simulated from the constrained nonnull unequal-variance RM model as described in Figure 1, except that there are 108 simulation settings because all three reader levels (3, 5, 10 readers) are included. Empirical coverage is the proportion of samples for a particular input combination such that $\hat{\zeta}_{1-\alpha} \geq \sigma^2_{TR:\mathrm{OR}}$, where $\hat{\zeta}_{1-\alpha}$ is the estimated $(1-\alpha)$ 100% upper one-sided confidence bound and $\sigma^2_{TR:\mathrm{OR}}$ is the true value computed from the Table 3 formulas.

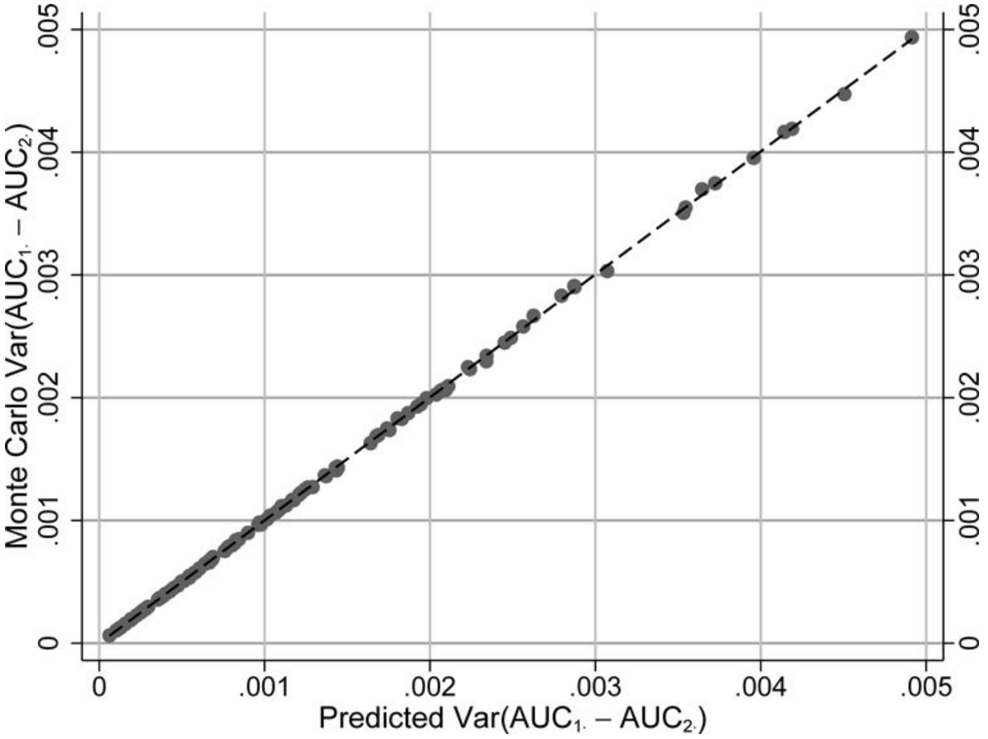

Figure 5.

Monte Carlo versus predicted variance of the empirical reader-averaged AUC test difference. The Monte Carlo values are based on data simulated from the constrained nonnull unequal-variance RM model as described in Figure 1, except that there are 108 simulation settings because all three reader levels (3, 5, 10 readers) are included. The predicted variance is computed using the Table 3 formula. The dashed line is the fitted linear regression line.

Table 1

Corrected and revised Roe and Metz [1] table of null simulation parameter values.

Structure	μ_+	A_z	Correlations				Variance Components					
			ρ_{WR} (incorrect)	ρ_{WR} (correct)	ρ_{BR1}	ρ_{BR2}	σ^2_C	$\sigma^2_{\tau C}$	σ^2_{RC}	σ^2_ε	σ^2_R	$\sigma^2_{\tau R}$
HL	0.75	0.702	(0.8)	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.0055	0.0055
HL	1.50	0.856	(0.8)	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.0055	0.0055
HL	2.50	0.961	(0.8)	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.0055	0.0055
LL	0.75	0.702	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.0055	0.0055
LL	1.50	0.856	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.0055	0.0055
LL	2.50	0.961	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.0055	0.0055
HH	0.75	0.702	(0.8)	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.011	0.011
HH	1.50	0.856	(0.8)	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.030	0.030
HH	2.50	0.961	(0.8)	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.056	0.056
LH	0.75	0.702	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.011	0.011
LH	1.50	0.856	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.030	0.030
LH	2.50	0.961	(0.4)	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.056	0.056

This table is reprinted, adapted and revised with permission from Roe and Metz [1, Table 1]. Notation is the same as in Reference [1]. For these null simulations $\mu_- = \tau_{i-} = \tau_{i+} = 0$, $i = 1, 2$; thus μ_+ is the median and mean separation of the normal and abnormal decision variable distributions across the reader population and $A_z = \Phi\left(\mu_+/\sqrt{2}\right)$ is the median reader-specific true AUC. This table contains slight corrections (e.g., A_Z =0.961 instead of 0.962) from Reference [1] so that A_Z corresponds to μ_+. Structure: HH = high data correlation, high reader variance; LL = low data correlation, low reader variance; LH = low data correlation, high reader variance; HL = high data correlation, low reader variance. $\rho_{WR} = \left(\sigma^2_C + \sigma^2_{RC}\right)/\left(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_\varepsilon\right)$ is the correct formula for the within-reader correlation between DV values for one reader and both tests, whereas the incorrect formula used by Roe and Metz is $\rho_{WR} = \left(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC}\right)/\left(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_\varepsilon\right)$;values for the incorrect formula are shown in parentheses. $\rho_{BR1} = \left(\sigma^2_C + \sigma^2_{\tau C}\right)/\left(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_\varepsilon\right)$ is the correlation between decision variable outcomes for different fixed readers using the same test and $\rho_{BR2} = \sigma^2_C/\left(\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_\varepsilon\right)$ is the correlation between decision variable outcomes for different fixed readers using different tests. In the original table ρ_{BR1} was labeled as ρ_{BR} and ρ_{BR2} was not included.

Table 2

Constrained unequal-variance RM-model [12] table of null simulation parameter values with the median mean-to-sigma ratio equal to 4.5.

							Variance Components									
				Correlations			**Normal cases**				**Abnormal cases**					
Structure	μ_+	A_z	b	ρ_{WR}	ρ_{BR1}	ρ_{BR2}	$\sigma^2_{C(-)}$	$\sigma^2_{\tau C(-)}$	$\sigma^2_{RC(-)}$	$\sigma^2_{\varepsilon(-)}$	$\sigma^2_{C(+)}$	$\sigma^2_{\tau C(+)}$	$\sigma^2_{RC(+)}$	$\sigma^2_{\varepsilon(+)}$	σ^2_R	$\sigma^2_{\tau R}$
HL	0.821	0.702	0.846	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.42	0.42	0.28	0.28	0.0066	0.0066
HL	1.831	0.856	0.711	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.59	0.59	0.40	0.40	0.0082	0.0082
HL	3.661	0.961	0.551	0.4	0.6	0.3	0.3	0.3	0.2	0.2	0.99	0.99	0.66	0.66	0.0118	0.0118
LL	0.821	0.702	0.846	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.14	0.14	0.28	0.84	0.0066	0.0066
LL	1.831	0.856	0.711	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.20	0.20	0.40	1.19	0.0082	0.0082
LL	3.661	0.961	0.551	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.33	0.33	0.66	1.97	0.0118	0.0118
HH	0.821	0.702	0.846	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.42	0.42	0.28	0.28	0.0132	0.0132
HH	1.831	0.856	0.711	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.59	0.59	0.40	0.40	0.0447	0.0447
HH	3.661	0.961	0.551	0.5	0.6	0.3	0.3	0.3	0.2	0.2	0.99	0.99	0.66	0.66	0.1201	0.1201
LH	0.821	0.702	0.846	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.14	0.14	0.28	0.84	0.0132	0.0132
LH	1.831	0.856	0.711	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.20	0.20	0.40	1.19	0.0447	0.0447
LH	3.661	0.961	0.551	0.3	0.2	0.1	0.1	0.1	0.2	0.6	0.33	0.33	0.66	1.97	0.1201	0.1201

This table is reprinted, adapted and revised with permission from Hillis [12, Table 2]. For these null simulations $\mu_- = \tau_{i-} = \tau_{i+} = 0$, for all i; thus μ_+ is the median and mean separation of the normal and abnormal decision variable distributions across the reader population and $A_z = \Phi\left(\mu_+/\sqrt{1+b^{-2}}\right)$ is the median reader-specific true AUC. Structure and correlations ρ_{WR}, ρ_{BR1} and ρ_{BR2} are defined in Table 1. Variance components involving case for the abnormal cases were computed by multiplying the corresponding variance components for the normal cases by b^{-2}; more precise values of b are 0.84566, 0.71082, and 0.55140. The median mean-to-sigma ratio across readers for each test is given by $\bar{r} = \mu_+/\left(\sigma_+ - \sigma_-\right) = 4.5$, where $\sigma_- = \sqrt{\sigma^2_{C(-)} + \sigma^2_{\tau C(-)} + \sigma^2_{RC(-)} + \sigma^2_{\varepsilon(-)}} = 1$ and $\sigma_+ = \sigma_-/b = 1/b$.

Table 3

Obuchowski-Rockette parameters, mean test-difference variance and expected test×reader mean square expressed in terms of RM-model parameters for the empirical AUC, assuming the unconstrained unequal-variance RM model.

$$E(\mathrm{Cov}_1) = \sum_{m=1}^{4} c_m F_{\mathrm{BVN}}\left(\frac{\delta_1}{\sqrt{V}}, \frac{\delta_2}{\sqrt{V}}; \frac{\rho_m\left(\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}\right) + 2\sigma^2_R}{V}\right)$$

$$\text{where } \rho_1 = \frac{\sigma^2_{RC(-)} + \sigma^2_{C(-)} + \sigma^2_{RC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_2 = \frac{\sigma^2_{RC(-)} + \sigma^2_{C(-)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_3 = \frac{\sigma^2_{RC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_4 = 0$$

$$\overline{E(\mathrm{Cov}_2)} = \frac{1}{2}\sum_{i=1}^{2} \mathrm{E}(\mathrm{Cov}_2 \mid \mathrm{test} = i) =$$

$$\frac{1}{2}\sum_{i=1}^{2}\sum_{m=1}^{4} c_m F_{\mathrm{BVN}}\left(\frac{\delta_i}{\sqrt{V}}, \frac{\delta_i}{\sqrt{V}}; \frac{\rho_m\left(\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}\right)}{V}\right)$$

$$\text{where } \rho_1 = \frac{\sigma^2_{TC(-)} + \sigma^2_{C(-)} + \sigma^2_{TC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_2 = \frac{\sigma^2_{TC(-)} + \sigma^2_{C(-)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_3 = \frac{\sigma^2_{TC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_4 = 0$$

$$E(\mathrm{Cov}_3) = \sum_{m=1}^{4} c_m F_{\mathrm{BVN}}\left(\frac{\delta_1}{\sqrt{V}}, \frac{\delta_2}{\sqrt{V}}; \frac{\rho_m\left(\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}\right)}{V}\right)$$

$$\text{where } \rho_1 = \frac{\sigma^2_{C(-)} + \sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_2 = \frac{\sigma^2_{C(-)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_3 = \frac{\sigma^2_{C(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_4 = 0$$

$$\overline{E(\sigma^2_{\varepsilon;\mathrm{OR}})} = \frac{1}{2}\sum_{i=1}^{2} E(\sigma^2_{\varepsilon;\mathrm{OR}} \mid \mathrm{test} = i) = \frac{1}{2}\sum_{i=1}^{2}\sum_{m=1}^{4} c_m F_{\mathrm{BVN}}\left(\frac{\delta_i}{\sqrt{V}}, \frac{\delta_i}{\sqrt{V}}; \frac{\rho_m\left(\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}\right) + 2\left(\sigma^2_R + \sigma^2_{TR}\right)}{V}\right)$$

$$\text{where } \rho_1 = 1, \rho_2 = \frac{\sigma^2_{TC(-)} + \sigma^2_{RC(-)} + \sigma^2_{C(-)} + \sigma^2_{\varepsilon(-)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_3 = \frac{\sigma^2_{TC(+)} + \sigma^2_{RC(+)} + \sigma^2_{C(+)} + \sigma^2_{\varepsilon(+)}}{\sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)}}, \rho_4 = 0$$

$$\mu_{\mathrm{OR}} + \tau_{i:\mathrm{OR}} = \Phi\left(\frac{\delta_i}{\sqrt{V}}\right)$$

$$\sigma^2_{R:\mathrm{OR}} = F_{\mathrm{BVN}}\left(\frac{\delta_1}{\sqrt{V}}, \frac{\delta_2}{\sqrt{V}}; \frac{2\sigma^2_R}{V}\right) - \left[\Phi\left(\frac{\delta_1}{\sqrt{V}}\right)\Phi\left(\frac{\delta_2}{\sqrt{V}}\right)\right]$$

$$\sigma^2_{TR:\mathrm{OR}} = .5\sum_{i=1}^{2}\left\{F_{\mathrm{BVN}}\left(\frac{\delta_i}{\sqrt{V}}, \frac{\delta_i}{\sqrt{V}}; \frac{2\left(\sigma^2_R + \sigma^2_{TR}\right)}{V}\right) - \left[\Phi\left(\frac{\delta_i}{\sqrt{V}}\right)\right]^2\right\} - \sigma^2_{R:\mathrm{OR}}$$

$$\mathrm{var}\left(\widehat{\mathrm{AUC}}_{1.} - \widehat{\mathrm{AUC}}_{2.}\right) = \frac{2}{r}\left[\overline{E\left(\sigma^2_{\varepsilon;\,\mathrm{OR}}\right)} + \sigma^2_{TR:\mathrm{OR}} - E(\mathrm{Cov}_1) + (r-1)\left(\overline{E(\mathrm{Cov}_2)} - E(\mathrm{Cov}_3)\right)\right]$$

$$E[\mathrm{MS(T * R)}] = \overline{E\left(\sigma^2_{\varepsilon;\,\mathrm{OR}}\right)} + \sigma^2_{TR:\mathrm{OR}} - E(\mathrm{Cov}_1) - \overline{E(\mathrm{Cov}_2)} + E(\mathrm{Cov}_3)$$

$$E[\mathrm{MS(R)}] = \overline{E\left(\sigma^2_{\varepsilon;\,\mathrm{OR}}\right)} + \sigma^2_{R:\mathrm{OR}} + \sigma^2_{TR:\mathrm{OR}} + E(\mathrm{Cov}_1) - \overline{E(\mathrm{Cov}_2)} - E(\mathrm{Cov}_3)$$

Because $\sigma^2_{\varepsilon;\,\mathrm{OR}}$, Cov$_1$, Cov$_2$, and Cov$_3$ can differ by reader, the formulas give their expectations across the reader population. Because $\sigma^2_{\varepsilon;\,\mathrm{OR}}$ and Cov$_2$ can also differ by test, their expectations are also averaged across tests, with averages indicated by $\overline{E\left(\sigma^2_{\varepsilon;\,\mathrm{OR}}\right)} \equiv \frac{1}{2}\sum_{i=1}^{2} E\left(\sigma^2_{\varepsilon;\,\mathrm{OR}} \mid \mathrm{test} = i\right)$ and $\overline{E\left(\mathrm{Cov}_2\right)} \equiv \frac{1}{2}\sum_{i=1}^{2} E\left(\mathrm{Cov}_2 \mid \mathrm{test} = i\right)$ F_{BVN} (., .; ρ) is the standardized bivariate normal distribution function with correlation ρ; $\delta_i = \mu_+ + \tau_{i+}$; $V = \sigma^2_{\mathrm{fixed}(-)} + \sigma^2_{\mathrm{fixed}(+)} + 2\left(\sigma^2_R + \sigma^2_{TR}\right)$, where $\sigma^2_{\mathrm{fixed}(-)} = \sigma^2_{C(-)} + \sigma^2_{\tau C(-)} + \sigma^2_{RC(-)} + \sigma^2_{\varepsilon(-)}$ and $\sigma^2_{\mathrm{fixed}(+)} = \sigma^2_{C(+)} + \sigma^2_{\tau C(+)} + \sigma^2_{RC(+)} + \sigma^2_{\varepsilon(+)}$; $c_1 = 1/(n_0 n_1)$; $c_2 = (n_1-1)/(n_0 n_1)$; $c_3 = (n_0-1)/(n_0 n_1)$; $c_4 = (1-n_0-n_1)/(n_0 n_1)$; $\mathrm{MS(T * R)} = \sum_{i=1}^{2}\sum_{j=1}^{r}\left(\widehat{\mathrm{AUC}}_{ij} - \widehat{\mathrm{AUC}}_{.j} - \widehat{\mathrm{AUC}}_{i.} + \widehat{\mathrm{AUC}}_{..}\right)^2/(r-1)$ the test×reader interaction mean square, where r is the number of readers; $\mathrm{MS(R)} = \sum_{j=1}^{r}\left(\widehat{\mathrm{AUC}}_{.j} - \widehat{\mathrm{AUC}}_{..}\right)^2/(r-1)$ is the reader mean square.

Table 4

Obuchowski-Rockette parameters $\mu_{\mathrm{OR}} + \tau_{1:\mathrm{OR}}$, $\mu_{\mathrm{OR}} + \tau_{2:\mathrm{OR}}$, $\sigma^2_{R:\mathrm{OR}}$, $\sigma^2_{TR:\mathrm{OR}}$ corresponding to RM-model parameters.

a) Original RM model [1]

	Original RM-model parameters				OR parameters			
Structure	μ_+	A_z	σ^2_R	$\sigma^2_{\tau R}$	$\mu_{\mathrm{OR}} + \tau_{1:\mathrm{OR}}$	$\mu_{\mathrm{OR}} + \tau_{2:\mathrm{OR}}$	$\sigma^2_{R:\mathrm{OR}}$	$\sigma^2_{TR:\mathrm{OR}}$
HL	0.75	0.702	0.0055	0.0055	0.7011	0.7011	0.000656	0.000657
HL	1.50	0.856	0.0055	0.0055	0.8543	0.8543	0.000285	0.000287
HL	2.50	0.961	0.0055	0.0055	0.9606	0.9606	0.000040	0.000040
LL	0.75	0.702	0.0055	0.0055	0.7011	0.7011	0.000656	0.000657
LL	1.50	0.856	0.0055	0.0055	0.8543	0.8543	0.000285	0.000287
LL	2.50	0.961	0.0055	0.0055	0.9606	0.9606	0.000040	0.000040
HH	0.75	0.702	0.011	0.011	0.7001	0.7001	0.001303	0.001307
HH	1.50	0.856	0.030	0.030	0.8485	0.8485	0.001582	0.001629
HH	2.50	0.961	0.056	0.056	0.9532	0.9532	0.000517	0.000590
LH	0.75	0.702	0.011	0.011	0.7001	0.7001	0.001303	0.001307
LH	1.50	0.856	0.030	0.030	0.8485	0.8485	0.001582	0.001629
LH	2.50	0.961	0.056	0.056	0.9532	0.9532	0.000517	0.000590

b) Constrained unequal-variance RM model [12]

	Constrained RM-model parameters				OR parameters			
Structure	μ_+	A_z	σ^2_R	$\sigma^2_{\tau R}$	$\mu_{\mathrm{OR}} + \tau_{1:\mathrm{OR}}$	$\mu_{\mathrm{OR}} + \tau_{2:\mathrm{OR}}$	$\sigma^2_{R:\mathrm{OR}}$	$\sigma^2_{TR:\mathrm{OR}}$
HL	0.821	0.702	0.0066	0.0066	0.7010	0.7010	0.000657	0.000658
HL	1.831	0.856	0.0082	0.0082	0.8543	0.8543	0.000286	0.000287
HL	3.661	0.961	0.0118	0.0118	0.9606	0.9606	0.000040	0.000040
LL	0.821	0.702	0.0066	0.0066	0.7010	0.7010	0.000657	0.000658
LL	1.831	0.856	0.0082	0.0082	0.8543	0.8543	0.000286	0.000287
LL	3.661	0.961	0.0118	0.0118	0.9606	0.9606	0.000040	0.000040
HH	0.821	0.702	0.0132	0.0132	0.7000	0.7000	0.001304	0.001308
HH	1.831	0.856	0.0447	0.0447	0.8486	0.8486	0.001582	0.001629

b) Constrained unequal-variance RM model [12]

Structure	Constrained RM-model parameters				OR parameters			
	μ_+	A$_z$	σ_R^2	$\sigma_{\tau R}^2$	$\mu_{\mathrm{OR}} + \tau_{1:\mathrm{OR}}$	$\mu_{\mathrm{OR}} + \tau_{2:\mathrm{OR}}$	$\sigma_{R:\mathrm{OR}}^2$	$\sigma_{TR:\mathrm{OR}}^2$
HH	3.661	0.961	0.1201	0.1201	0.9532	0.9532	0.000517	0.000590
LH	0.821	0.702	0.0132	0.0132	0.7000	0.7000	0.001304	0.001308
LH	1.831	0.856	0.0447	0.0447	0.8486	0.8486	0.001582	0.001629
LH	3.661	0.961	0.1201	0.1201	0.9532	0.9532	0.000517	0.000590

RM parameter values for $\sigma_C^2, \sigma_{\tau C}^2, \sigma_{RC}^2$, and σ_ε^2, not shown, are the same as in Table 1.

RM parameter values for $\sigma_C^2, \sigma_{\tau C}^2, \sigma_{RC}^2$, and σ_ε^2, not shown, are the same as in Table 2.

Table 5

Illustration of computations for Tables 4 and 6.

a) Computations for first row in Table 4a

$$\delta_1 = \mu_+ + \tau_{1+} = 0.75 + 0 = 0.75; \delta_2 = \mu_+ + \tau_{2+} = 0.75 + 0 = 0.75$$

$$V = \sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)} + 2\left(\sigma^2_R + \sigma^2_{TR}\right) = 1 + 1 + 2(0.0055 + 0.0055) = 2.022$$

$$\mu_{\text{OR}} + \tau_{i:\text{OR}} = \Phi\left(\delta_i/\sqrt{V}\right) = \Phi(0.75/\sqrt{2.022}) = 0.7011, i = 1, 2$$

$$\sigma^2_{R:\text{OR}} = F_{\text{BVN}}\left(\frac{\delta_1}{\sqrt{V}}, \frac{\delta_2}{\sqrt{V}}, \frac{2\sigma^2_R}{\sqrt{V}}\right) - \left[\Phi\left(\frac{\delta_1}{\sqrt{V}}\right), \Phi\left(\frac{\delta_2}{\sqrt{V}}\right)\right] = F_{\text{BVN}}\left(\frac{0.75}{\sqrt{2.0220}}, \frac{0.75}{\sqrt{2.0220}}, \frac{2(0.0055)}{2.022}\right) - \left[\Phi\left(\frac{0.75}{\sqrt{2.022}}\right)\right]^2 = 0.000656$$

$$\sigma^2_{TR:\text{OR}} = .5\sum_{i=1}^{2}\left\{F_{\text{BVN}}\left(\frac{\delta_i}{\sqrt{V}}, \frac{\delta_i}{\sqrt{V}}, \frac{2\left(\sigma^2_R + \sigma^2_{TR}\right)}{V}\right) - \left[\Phi\left(\frac{\delta_i}{\sqrt{V}}\right)\right]^2\right\} - \sigma^2_{R:\text{OR}} = F_{\text{BVN}}\left(\frac{0.75}{\sqrt{2.0220}}, \frac{0.75}{\sqrt{2.0220}}, \frac{2(0.0055 + 0.0055)}{2.022}\right) - \left[\Phi\left(\frac{0.75}{\sqrt{2.022}}\right)\right]^2 - 0.000656 = 0$$

b) Computations for Cov$_1$ in first row in Table 6

$$\text{Cov}_1 = \sum_{m=1}^{4} c_m F_{\text{BVN}}\left(\frac{\delta_1}{\sqrt{V}}, \frac{\delta_2}{\sqrt{V}}; \frac{\rho_m(\sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)}) + 2\sigma^2_R}{V}\right) = .00346 \text{ where}$$

$$\delta_1 = \delta_2 = 0.75,\ V = 2.022,\ \sigma^2_R = 0.0055,\ \sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)} = 2,\ \sigma^2_{C(-)} = \sigma^2_{C(+)} = 0.3,\ \sigma^2_{\tau C(-)} = \sigma^2_{\tau C(+)} = 0.3,$$

$$\sigma^2_{RC(-)} = \sigma^2_{RC(+)} = 0.2,\ \sigma^2_{\varepsilon(-)} = \sigma^2_{\varepsilon(+)} = 0.2,\ n_0 = 90,\ n_1 = 10,\ c_1 = \frac{1}{n_0 n_1} = \frac{1}{10(90)},$$

$$c_3 = \frac{n_0 - 1}{n_0 n_1} = \frac{89}{10(90}, c_2 = \frac{n_1 - 1}{n_0 n_1} = \frac{9}{10(90)}, c_4 = \frac{1 - n_0 - n_1}{n_0 n_1} = \frac{-99}{10(90)}, \rho_1 = \frac{\sigma^2_{RC(-)} + \sigma^2_{C(-)} + \sigma^2_{RC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)}} = \frac{1}{2},$$

$$\rho_2 = \frac{\sigma^2_{RC(-)} + \sigma^2_{C(-)}}{\sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)}} = \frac{0.5}{2}, \rho_3 = \frac{\sigma^2_{RC(+)} + \sigma^2_{C(+)}}{\sigma^2_{\text{fixed}(-)} + \sigma^2_{\text{fixed}(+)}} = \frac{0.5}{2}, \rho_4 = 0.$$

Table 6

Obuchowski-Rockette parameters Cov$_1$, Cov$_2$, Cov$_3$, $\sigma^2_{\varepsilon:\mathrm{OR}}$, r_1, r_2, and r_3 corresponding to original Roe and Metz [1] model parameter values.

Roe and Metz parameters										Obuchowski-Rockette parameters						
Structure	μ_+	**A$_z$**	ρ_{WR}	ρ_{BR1}	ρ_{BR2}	σ^2_C	$\sigma^2_{\tau C}$	σ^2_{RC}	σ^2_ε	**Cov$_1$×10^5**	**Cov$_2$×10^5**	**Cov$_3$×10^5**	$\sigma^2_{\varepsilon:\mathrm{OR}} \times 10^5$	r_1	r_2	r_3
a) Sample size = 10+/90−																
HL	0.75	0.702	0.4	0.6	0.3	0.3	0.3	0.2	0.2	346	418	203	743	0.47	0.56	0.27
HL	1.50	0.856	0.4	0.6	0.3	0.3	0.3	0.2	0.2	165	203	94	384	0.43	0.53	0.24
HL	2.50	0.961	0.4	0.6	0.3	0.3	0.3	0.2	0.2	29	37	15	83	0.35	0.44	0.18
LL	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	204	134	67	743	0.27	0.18	0.09
LL	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	94	61	30	384	0.24	0.16	0.08
LL	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	15	9	4	83	0.18	0.11	0.05
HH	0.75	0.702	0.5	0.6	0.3	0.3	0.3	0.2	0.2	343	415	202	740	0.46	0.56	0.27
HH	1.50	0.856	0.5	0.6	0.3	0.3	0.3	0.2	0.2	168	201	93	395	0.43	0.51	0.24
HH	2.50	0.961	0.5	0.6	0.3	0.3	0.3	0.2	0.2	36	41	17	106	0.34	0.39	0.16
LH	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	202	133	66	740	0.27	0.18	0.09
LH	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	96	61	30	395	0.24	0.15	0.08
LH	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	20	11	5	106	0.18	0.10	0.05
b) Sample size = 25+/25−																
HL	0.75	0.702	0.4	0.6	0.3	0.3	0.3	0.2	0.2	249	302	146	541	0.46	0.56	0.27
HL	1.50	0.856	0.4	0.6	0.3	0.3	0.3	0.2	0.2	119	146	67	281	0.42	0.52	0.24
HL	2.50	0.961	0.4	0.6	0.3	0.3	0.3	0.2	0.2	21	27	11	61	0.34	0.43	0.18
LL	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	147	97	48	541	0.27	0.18	0.09
LL	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	68	44	21	281	0.24	0.16	0.08
LL	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	11	7	3	61	0.18	0.11	0.05
HH	0.75	0.702	0.5	0.6	0.3	0.3	0.3	0.2	0.2	248	299	145	539	0.46	0.56	0.27
HH	1.50	0.856	0.5	0.6	0.3	0.3	0.3	0.2	0.2	121	145	67	288	0.42	0.50	0.23
HH	2.50	0.961	0.5	0.6	0.3	0.3	0.3	0.2	0.2	26	30	12	78	0.34	0.38	0.16
LH	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	146	96	48	539	0.27	0.18	0.09
LH	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	69	44	21	288	0.24	0.15	0.07
LH	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	14	8	4	78	0.18	0.10	0.05

	Roe and Metz parameters									Obuchowski-Rockette parameters						
Structure	μ_+	A_z	ρ_{WR}	ρ_{BR1}	ρ_{BR2}	σ^2_C	$\sigma^2_{\tau C}$	σ^2_{RC}	σ^2_ε	$Cov_1 \times 10^5$	$Cov_2 \times 10^5$	$Cov_3 \times 10^5$	$\sigma^2_{\varepsilon:OR} \times 10^5$	r_1	r_2	r_3
c) Sample size = 100+/100−																
HL	0.75	0.702	0.4	0.6	0.3	0.3	0.3	0.2	0.2	62	75	37	133	0.47	0.57	0.28
HL	1.50	0.856	0.4	0.6	0.3	0.3	0.3	0.2	0.2	30	36	17	69	0.43	0.53	0.25
HL	2.50	0.961	0.4	0.6	0.3	0.3	0.3	0.2	0.2	5	7	3	15	0.35	0.45	0.18
LL	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	37	24	12	133	0.28	0.18	0.09
LL	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	17	11	5	69	0.25	0.16	0.08
LL	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	3	2	1	15	0.19	0.11	0.05
HH	0.75	0.702	0.5	0.6	0.3	0.3	0.3	0.2	0.2	62	75	36	132	0.47	0.56	0.27
HH	1.50	0.856	0.5	0.6	0.3	0.3	0.3	0.2	0.2	30	36	17	70	0.43	0.51	0.24
HH	2.50	0.961	0.5	0.6	0.3	0.3	0.3	0.2	0.2	7	7	3	19	0.35	0.39	0.16
LH	0.75	0.702	0.3	0.2	0.1	0.1	0.1	0.2	0.6	36	24	12	132	0.28	0.18	0.09
LH	1.50	0.856	0.3	0.2	0.1	0.1	0.1	0.2	0.6	17	11	5	70	0.24	0.15	0.08
LH	2.50	0.961	0.3	0.2	0.1	0.1	0.1	0.2	0.6	4	2	1	19	0.19	0.10	0.05

Notes: Structure and correlations ρ_{WR}, ρ_{BR1} and ρ_{BR1} are defined in Table 1; SS = sample size: “10+/90−” indicates 10 diseased and 90 nondiseased images, etc.; Roe and Metz parameter values for σ^2_R and σ^2_{TR}, not shown, are the same as in Table 1; Cov_1, Cov_2, Cov_3 and $\sigma^2_{\varepsilon:OR}$ are the expected values, $E(Cov_1)$, $\overline{E(Cov_2)}$, $E(Cov_3)$ and $\overline{E(\sigma^2_{\varepsilon:OR})}$ computed using the formulas in Table 3; $r_{i:OR} = Cov_i / \sigma^2_{\varepsilon:OR}$, $i = 1,2,3$.

Table 7

Obuchowski-Rockette parameters Cov$_1$, Cov$_2$, Cov$_3$, $\sigma^2_{\varepsilon:\mathrm{OR}}$, r_1, r_2, and r_3 corresponding to Hillis [12] constrained unequal-variance Roe and Metz model values.

Roe and Metz parameters							Obuchowski-Rockette parameters						
Structure	μ_+	**A$_z$**	b	ρ_{WR}	ρ_{BR1}	ρ_{BR2}	**Cov$_1$×10^5**	**Cov$_2$×10^5**	**Cov$_3$×10^5**	$\sigma^2_{\varepsilon:\mathrm{OR}} \times 10^5$	r_1	r_2	r_3
a) Sample size = 10+/90−													
HL	0.821	0.702	0.846	0.4	0.6	0.3	395	478	231	860	0.46	0.56	0.27
HL	1.831	0.856	0.711	0.4	0.6	0.3	217	267	121	529	0.41	0.51	0.23
HL	3.661	0.961	0.551	0.4	0.6	0.3	48	63	24	161	0.30	0.39	0.15
LL	0.821	0.702	0.846	0.3	0.2	0.1	232	153	76	860	0.27	0.18	0.09
LL	1.831	0.856	0.711	0.3	0.2	0.1	121	78	38	529	0.23	0.15	0.07
LL	3.661	0.961	0.551	0.3	0.2	0.1	24	14	6	161	0.15	0.09	0.04
HH	0.821	0.702	0.846	0.5	0.6	0.3	392	474	229	856	0.46	0.55	0.27
HH	1.831	0.856	0.711	0.5	0.6	0.3	219	264	120	540	0.41	0.49	0.22
HH	3.661	0.961	0.551	0.5	0.6	0.3	59	69	27	197	0.30	0.35	0.14
LH	0.821	0.702	0.846	0.3	0.2	0.1	230	151	75	856	0.27	0.18	0.09
LH	1.831	0.856	0.711	0.3	0.2	0.1	124	78	38	540	0.23	0.14	0.07
LH	3.661	0.961	0.551	0.3	0.2	0.1	30	16	7	197	0.15	0.08	0.04
b) Sample size = 25+/25−													
HL	0.821	0.702	0.846	0.4	0.6	0.3	250	302	147	544	0.46	0.56	0.27
HL	1.831	0.856	0.711	0.4	0.6	0.3	121	149	68	292	0.41	0.51	0.23
HL	3.661	0.961	0.551	0.4	0.6	0.3	23	30	11	75	0.31	0.40	0.15
LL	0.821	0.702	0.846	0.3	0.2	0.1	147	97	48	544	0.27	0.18	0.09
LL	1.831	0.856	0.711	0.3	0.2	0.1	68	44	21	292	0.23	0.15	0.07
LL	3.661	0.961	0.551	0.3	0.2	0.1	12	7	3	75	0.15	0.09	0.04
HH	0.821	0.702	0.846	0.5	0.6	0.3	248	300	145	542	0.46	0.55	0.27
HH	1.831	0.856	0.711	0.5	0.6	0.3	123	147	68	299	0.41	0.49	0.23
HH	3.661	0.961	0.551	0.5	0.6	0.3	28	33	13	93	0.30	0.35	0.14
LH	0.821	0.702	0.846	0.3	0.2	0.1	146	96	48	542	0.27	0.18	0.09
LH	1.831	0.856	0.711	0.3	0.2	0.1	70	44	21	299	0.23	0.15	0.07

	Roe and Metz parameters						Obuchowski-Rockette parameters						
Structure	μ_+	A_z	b	ρ_{WR}	ρ_{BR1}	ρ_{BR2}	$\mathrm{Cov}_1\times10^5$	$\mathrm{Cov}_2\times10^5$	$\mathrm{Cov}_3\times10^5$	$\sigma^2_{\varepsilon:\mathrm{OR}}\times10^5$	r_1	r_2	r_3
LH	3.661	0.961	0.551	0.3	0.2	0.1	15	8	4	93	0.16	0.09	0.04
c) Sample size = 100+/100−													
HL	0.821	0.702	0.846	0.4	0.6	0.3	62	75	37	134	0.47	0.56	0.27
HL	1.831	0.856	0.711	0.4	0.6	0.3	30	37	17	71	0.42	0.52	0.24
HL	3.661	0.961	0.551	0.4	0.6	0.3	6	7	3	18	0.31	0.41	0.16
LL	0.821	0.702	0.846	0.3	0.2	0.1	37	24	12	134	0.27	0.18	0.09
LL	1.831	0.856	0.711	0.3	0.2	0.1	17	11	5	71	0.24	0.15	0.07
LL	3.661	0.961	0.551	0.3	0.2	0.1	3	2	1	18	0.16	0.09	0.04
HH	0.821	0.702	0.846	0.5	0.6	0.3	62	75	36	133	0.46	0.56	0.27
HH	1.831	0.856	0.711	0.5	0.6	0.3	31	37	17	73	0.42	0.50	0.23
HH	3.661	0.961	0.551	0.5	0.6	0.3	7	8	3	23	0.31	0.36	0.14
LH	0.821	0.702	0.846	0.3	0.2	0.1	36	24	12	133	0.27	0.18	0.09
LH	1.831	0.856	0.711	0.3	0.2	0.1	17	11	5	73	0.24	0.15	0.07
LH	3.661	0.961	0.551	0.3	0.2	0.1	4	2	1	23	0.16	0.09	0.04

Notes: Structure and correlations ρ_{WR}, ρ_{BR1} and ρ_{BR1} are defined in Table 1; SS = sample size: "10+/90−" indicates 10 diseased and 90 nondiseased images, etc.; parameter values for $\sigma^2_R, \sigma^2_{TR}, \sigma^2_{C(-)}, \sigma^2_{\tau C(-)}, \sigma^2_{RC(-)}, \sigma^2_{\varepsilon(-)}, \sigma^2_{C(+)}, \sigma^2_{\tau C(+)}, \sigma^2_{RC(+)}, \sigma^2_{\varepsilon(+)}$, not shown, are the same as in Table 2; Cov_1, Cov_2, Cov_3 and $\sigma^2_{\varepsilon:\mathrm{OR}}$ are the expected values, $E(\mathrm{Cov}_1)$, $\overline{E\left(\mathrm{Cov}_2\right)}$, $E(\mathrm{Cov}_3)$ and $\overline{E\left(\sigma^2_{\varepsilon:\mathrm{OR}}\right)}$ computed using the formulas in Table 3; $r_{i:\mathrm{OR}} = \mathrm{Cov}_i/\sigma^2_{\varepsilon:\mathrm{OR}}$, $i = 1,2,3$.