Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances

Brandon D. Gallas
Stephen L. Hillis

SPIE.

Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances

Brandon D. Gallas[a,*] **and Stephen L. Hillis**[b,c,*]
[a]CDRH/FDA, Division of Imaging and Applied Mathematics, Bldg. 62, Rm 3124, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993-0002, United States
[b]University of Iowa, Departments of Radiology and Biostatistics, 3170 Medical Laboratories, 200 Hawkins Drive, Iowa City, Iowa 52242-1077, United States
[c]Comprehensive Access and Delivery Research and Evaluation Center, VA Health Care System, Iowa City, Iowa 52242-1077, United States

Abstract. Modeling and simulation are often used to understand and investigate random quantities and estimators. In 1997, Roe and Metz introduced a simulation model to validate analysis methods for the popular endpoint in reader studies to evaluate medical imaging devices, the reader-averaged area under the receiver operating characteristic (ROC) curve. Here, we generalize the notation of the model to allow more flexibility in recognition that variances of ROC ratings depend on modality and truth state. We also derive and validate equations for computing population variances and covariances for reader-averaged empirical AUC estimates under the generalized model. The equations are one-dimensional integrals that can be calculated using standard numerical integration techniques. This work provides the theoretical foundation and validation for a Java application called iRoeMetz that can simulate multireader multicase ROC studies and numerically calculate the corresponding variances and covariances of the empirical AUC. The iRoeMetz application and source code can be found at the "iMRMC" project on the google code project hosting site. These results and the application can be used by investigators to investigate ROC endpoints, validate analysis methods, and plan future studies.
 [DOI: 10.1117/1.JMI.1.3.031006]

Keywords: medical imaging; simulations; radiology; receiver operating characteristic curves; multireader multicase; reader studies.

Paper 14047SSR received Apr. 16, 2014; revised manuscript received Jun. 26, 2014; accepted for publication Aug. 22, 2014; published online Sep. 25, 2014.

1 Introduction

The area under the receiver operating characteristic (ROC) curve, denoted AUC, is a common endpoint in reader studies to evaluate medical imaging devices.[1] ROC data from reader studies are confidence-of-disease ratings from clinicians (readers) evaluating images (cases). Therefore, the endpoint in a multireader multicase (MRMC) ROC study is affected by two important sources of variability—the readers and the cases—and in studies past the exploratory stage, we often want to account for both sources in our analyses.

Modeling and simulation are tools that help us to understand and investigate the distribution and statistical behavior of random quantities and estimators. Roe and Metz[2] (R&M) proposed a simulation model that launched the study of different analysis methods and endpoints related to MRMC ROC studies. Their model was developed to validate the Dorfman, Berbaum, and Metz (DBM) method that compares AUCs from two modalities.[3] The R&M model simulates ROC ratings according to a binormal model for each reader and generates data from a "fully crossed" study design where each patient is imaged by two or more modalities, with the resulting cases evaluated once by each reader.

With few modifications, the R&M model has been used to validate and characterize many other MRMC ROC analysis methods for almost two decades.[4–10] The R&M model has also been used to investigate power and sizing methods for ROC studies,[11,12] and adapted to yield discrete ROC ratings[13,14] and explore alternative study designs.[15] Finally, while AUC has been the primary reader performance measure analyzed with the R&M model to date, the R&M model has also been used to analyze binary performance measures[16,17] and utility.[18]

The R&M model assumes a four-factor modality × reader × case × truth mixed-effects analysis of variance model for the ROC ratings. In the original work, reader and case effects and their interactions were assumed to be random and the variance components were chosen to be the same across truth states and across modalities. As we show below, the assumptions applied to the R&M model in the original work can be relaxed. In the current work, we generalize the notation to clarify that the variance components can depend on truth state and modality.

As discussed by Hillis,[19] when the variance components are assumed to be the same across truth states and modalities, the R&M model has the following interpretations: (1) ROC ratings for each reader are generated from an equal-variance binormal model (i.e, a binormal model such that variances of the nondiseased and diseased ROC ratings are equal); and (2) the expected differences (or separations) between the nondiseased and diseased ROC ratings vary across readers, with the separations having the same variance for each modality. This last result

*Address all correspondence to: Brandon D. Gallas, E-mail: brandon.gallas@fda.hhs.gov; Stephen L. Hillis, E-mail: steve-hillis@uiowa.edu

implies that for a simulation study that assumes equal AUCs across modalities (i.e., a null-hypothesis study), the resulting AUC estimates will have the same variance for each modality. It is natural to question this assumption, especially when comparing an imaging modality with and without a computer aid. Beiden et al.[20] found that the reader variability of readers' AUCs was much smaller when using a computer aid in classifying microcalcifications in mammograms compared to those without the aid.

When fitting an ROC curve with a binormal model, it is generally recognized that for real data the distributions of the latent decision variables of the diseased and nondiseased ROC ratings will often have different variances, with the diseased distribution typically wider. This causes the ROC curve to be unsymmetric about the negative diagonal. Such unsymmetric ROC curves have been seen as far back as the early psychophysical experiments of the 1960s[21,22] and in recent studies evaluating medical imaging modalities.[19,23–26] Unsymmetric ROC curves have motivated other models of ROC ratings[27–29] and are sometimes characterized by a mean-to-sigma ratio defined as the difference of the binormal means divided by the difference of the binormal standard deviations across truth states in an unequal-variance binormal model.[19,21,26] For this reason, Hillis[19] introduced an unequal-variance binormal model by allowing some of the variance components to depend on truth but with some additional constraints.

In this paper, our main purpose is to present the exact nature of the relationship between the R&M model inputs (these include the model parameters and numbers of readers and cases) and the means, variances and covariances of the resulting reader-averaged empirical AUC estimates. R&M note that they were not able to determine such a relationship. Knowledge of this relationship builds on earlier work by Gallas et al.[16] and allows an investigator to, among other things, (1) verify that the simulation model has been correctly programmed by comparing parameter estimates based on the simulations to the true values of the parameters; and (2) quantify the bias of AUC variance estimates by similarly comparing simulation results to the true values. We consider this paper to be an important first step toward our ultimate aim of being able to calibrate a simulation model that will produce data that matches a real data set with respect to the estimated parameters from an analysis method, such as that proposed by Obuchowski and Rockette.[30]

In this paper, we begin by generalizing the notation of the R&M model by allowing all of its variance components to depend on both modality and truth state. Then, we present and validate equations for computing the population variances and covariances for empirical AUC estimates computed from data simulated from the generalized R&M model. The generalized model includes the original R&M model (with the equal variance-components assumption) and the unequal-variance model proposed by Hillis[19] as special cases. Although these two special cases are sufficient for many situations, we anticipate that researchers may want to use other special cases of the generalized R&M model, or the generalized model itself for simulating data. Presenting equations for the generalized model eliminates the need to derive equations for each special case in the future.

2 Methods

2.1 *Generalized Roe and Metz Model*

The R&M simulation model is for simulating rating data that emulate an ROC reader-performance study that has N_0 nondiseased cases and N_1 diseased cases that are interpreted and rated by N_R readers. Here, we focus on a fully crossed study design to compare two modalities, denoted by A and B, where "fully crossed" refers to the data collection: all the readers rate all the cases in both modalities with respect to confidence of disease. The result of such a study is a dataset with $2 \times (N_0 + N_1) \times N_R$ ROC ratings $[(\text{modalities}) \times (\text{cases}) \times (\text{readers})]$. R&M denote the ROC ratings by X_{ijkt}, where i denotes modality ($i =$ "A" or "B"), j denotes reader, k denotes case, and t denotes truth $(t = 0,1 \equiv \text{nondiseased, diseased})$.

Using the notation of R&M, the model is given by

$$X_{ijkt} = \mu_t + \tau_{it} + R_{jt} + C_{kt} + [RC]_{jkt} + [\tau R]_{ijt} + [\tau C]_{ikt} + [\tau RC]_{ijkt} + E_{ijkt}, \quad (1)$$

where X_{ijkt} denotes the value of the ROC rating for modality i, reader j, case k, and truth state t. Modality and truth are fixed factors and reader and case are random factors, i.e., effects involving reader or case are random effects and all other effects are fixed. Consequently, the Greek terms μ_t, τ_{it} are fixed effects and the remaining seven terms are random.

The random terms in the R&M model are all independent zero-mean Gaussian random variables: R is the reader effect, C is the case effect, $[RC]$ is the reader × case effect, $[\tau R]$ is the modality × reader effect, $[\tau C]$ is the modality × case effect, $[\tau RC]$ is the modality × reader × case effect, and E is an independent random error term. The corresponding variance components are denoted $\sigma_R^2, \sigma_C^2, \sigma_{RC}^2, \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{\tau RC}^2$, and σ_E^2. The independent error term E_{ijkt} can be attributed to a reader's inability to exactly reproduce their ROC rating for a case; E is sometimes referred to as internal noise or reader jitter. R&M pointed out that the pure error and the three-way interaction variance components cannot be separately estimated "without multiple readings of each case in each modality by all readers." As such, they defined an aggregate term $\varepsilon_{ijkt} = [\tau RC]_{ijkt} + E_{ijkt}$.

2.1.1 *Updated notation*

Here, we generalize the R&M model, clarifying that the model is, in fact, a four-factor model. In addition to modality, reader, and case, truth is a factor. In particular, truth has a hierarchical relationship with cases:[31] each case can have only one truth state and cases are "nested" within truth. Mathematically, we will use α to express the truth factor and rewrite the R&M model as

$$X_{ijkt} = \alpha_t + [\tau\alpha]_{it} + [R\alpha]_{jt} + [C\alpha]_{kt} + [RC\alpha]_{jkt} + [\tau R\alpha]_{ijt} + [\tau C\alpha]_{ikt} + [\tau RC\alpha]_{ijkt}, \quad (2)$$

where we have omitted the pure error term since it cannot be distinguished from the four-way interaction term without replications. Following statistical conventions, we could surround α and its subscript t with parentheses (for terms that include case) to indicate the nesting of cases within truth. However, we will not follow this convention for simplicity.

Now, the four-factor structure of the model is clearly shown in Eq. (2). We point out that this model includes only interactions with truth or effects nested within truth; in particular, there are no effects for modality alone or reader alone. The rationale for omitting the modality and reader effects is that these terms would have no effect on the ROC curve for a given reader and test, since the ROC curve is invariant to location

shifts of the decision variable. We note that the fixed interaction $[\tau\alpha]_{it}$ allows different modalities to have different ROC curves.

2.1.2 *Allow variances to depend on modality and truth*

Given the new notation for the R&M model, we now generalize it to allow the variance components to depend on modality and truth.

There are three random-effect terms that do not include modality: $[R\alpha]_{jt}, [C\alpha]_{kt}, [RC\alpha]_{jkt}$. They correspond to six variance components: $\sigma^2_{R0}, \sigma^2_{C0}, \sigma^2_{RC0}$ are for nondiseased cases ($t = 0$) and $\sigma^2_{R1}, \sigma^2_{C1}, \sigma^2_{RC1}$ are for diseased cases ($t = 1$). The sum of these six variance components that do not depend on modality is

$$\sigma^2_\Omega = \sigma^2_{R0} + \sigma^2_{C0} + \sigma^2_{RC0} + \sigma^2_{R1} + \sigma^2_{C1} + \sigma^2_{RC1}. \tag{3}$$

There are three random-effect terms that include modality: $[\tau R\alpha]_{ijt}$, $[\tau C\alpha]_{ikt}$, and $[\tau RC\alpha]_{ijkt}$. They correspond to 12 variance components: $\sigma^2_{\mathrm{A}R0}, \sigma^2_{\mathrm{A}C0}, \sigma^2_{\mathrm{A}RC0}$ are for modality A, nondiseased cases ($i = \mathrm{A}; t = 0$), $\sigma^2_{\mathrm{A}R1}, \sigma^2_{\mathrm{A}C1}, \sigma^2_{\mathrm{A}RC1}$ are for modality A, diseased cases ($i = \mathrm{A}; t = 1$), $\sigma^2_{BR0}, \sigma^2_{BC0}, \sigma^2_{BRC0}$ are for modality B, nondiseased cases ($i = \mathrm{B}; t = 0$), and $\sigma^2_{BR1}, \sigma^2_{BC1}$, σ^2_{BRC1} are for modality B, diseased cases ($i = \mathrm{B}; t = 1$). The sums of the variance components that are specific to modality A and B are

$$\sigma^2_{\mathrm{A}} = \sigma^2_{\mathrm{A}R0} + \sigma^2_{\mathrm{A}C0} + \sigma^2_{\mathrm{A}RC0} + \sigma^2_{\mathrm{A}R1} + \sigma^2_{\mathrm{A}C1} + \sigma^2_{\mathrm{A}RC1}, \tag{4}$$

$$\sigma^2_{\mathrm{B}} = \sigma^2_{\mathrm{B}R0} + \sigma^2_{\mathrm{B}C0} + \sigma^2_{\mathrm{B}RC0} + \sigma^2_{\mathrm{B}R1} + \sigma^2_{\mathrm{B}C1} + \sigma^2_{\mathrm{B}RC1}. \tag{5}$$

The original R&M model is a special case of the generalized model. All we need to do is assume, as R&M did, that the variance components of the ROC ratings do not depend on modality or truth. We can replicate the original R&M model by setting generalized R&M model variance components equal to the original ones as given in Table 1. Other simplifications can be similarly handled.

2.2 *Expected AUCs*

Here, we examine the expected value of the empirical estimate of AUC, also known as the trapezoidal estimate,[32] given the R&M model. Without loss of generality, we focus on modality A. A similar discussion can be derived for modality B.

The estimated reader-averaged AUC for modality A is

$$\widehat{\mathrm{AUC}}_{\mathrm{A}} = \sum_{j=1}^{N_R} \sum_{k=1}^{N_0} \sum_{k'=1}^{N_1} s(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0})/N_0 N_1 N_R, \tag{6}$$

where we shall refer to $s(x)$ as the "success function;" $s(x)$ equals 1.0 when reader j successfully rates diseased case k' higher than nondiseased case k, $s(x)$ equals 0.0 if the ratings are in the wrong order, and $s(x)$ equals 0.5 if the ratings are tied.

As we consider the expected reader-averaged AUC for modality A, we are averaging over readers and cases as we average over the ROC ratings. We express the expectation of $\widehat{\mathrm{AUC}}_{\mathrm{A}}$ [Eq. (6)] as

Table 1 Equivalences needed to replicate original R&M model with the generalized R&M model.

Original R&M	Generalized R&M
$\sigma^2_R =$	$\sigma^2_{R0} = \sigma^2_{R1}$
$\sigma^2_C =$	$\sigma^2_{C0} = \sigma^2_{C1}$
$\sigma^2_{RC} =$	$\sigma^2_{RC0} = \sigma^2_{RC1}$
$\sigma^2_{\tau R} =$	$\sigma^2_{\mathrm{A}R0} = \sigma^2_{\mathrm{B}R0} = \sigma^2_{\mathrm{A}R1} = \sigma^2_{\mathrm{B}R1}$
$\sigma^2_{\tau C} =$	$\sigma^2_{\mathrm{A}C0} = \sigma^2_{\mathrm{B}C0} = \sigma^2_{\mathrm{A}C1} = \sigma^2_{\mathrm{B}C1}$
$\sigma^2_\varepsilon =$	$\sigma^2_{\mathrm{A}RC0} = \sigma^2_{\mathrm{B}RC0} = \sigma^2_{\mathrm{A}RC1} = \sigma^2_{\mathrm{B}RC1}$

$$\begin{aligned}\mathrm{AUC}_{\mathrm{A}} &= E\left[\sum_{j=1}^{N_R} \sum_{k=1}^{N_0} \sum_{k'=1}^{N_1} s(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0})/N_0 N_1 N_R\right] \\ &= E[s(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0})] = E[s_{\mathrm{A}jkk'}],\end{aligned} \tag{7}$$

where we have introduced the random variable $s_{\mathrm{A}jkk'} = s(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0})$ that we refer to as a "success observation." The equivalences above are true because we can pull the summations out of the expected value and all the summands yield the same result. To be clear, $E[s_{\mathrm{A}jkk'}]$ is the expected value of $s_{\mathrm{A}jkk'}$ for a randomly selected reader reading randomly selected diseased and nondiseased cases. In a sense, the expected value averages over the subscripts, and is not dependent on them.

In Eq. (6), $X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0}$ is nothing more than the difference of a few fixed effects and many zero-mean random effects. Specifically, it is a normal random variable with a mean and variance given by

$$\begin{aligned}\Delta_{\mathrm{A}} &= (\alpha_1 + [\tau\alpha]_{\mathrm{A}1}) - (\alpha_0 + [\tau\alpha]_{\mathrm{A}0}), \\ \mathrm{var}(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0}) &= \sigma^2_\Omega + \sigma^2_{\mathrm{A}},\end{aligned} \tag{8}$$

where σ^2_Ω is the sum of the six variance components that do not depend on modality [Eq. (3)] and σ^2_{A} is the sum of the six variance components that are specific to modality A [Eq. (4)]. Therefore, we can express the expected value of $\widehat{\mathrm{AUC}}_A$ analytically as

$$\begin{aligned}\mathrm{AUC}_{\mathrm{A}} &= E(s_{\mathrm{A}jkk'}) = \Pr(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0} > 0) \\ &= \Phi(\Delta_{\mathrm{A}}/\sqrt{\sigma^2_\Omega + \sigma^2_{\mathrm{A}}}),\end{aligned} \tag{9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

2.3 *Variances and Covariances*

Here, we turn our attention to variances and covariances. Both appear in the variance of the difference of estimated reader-averaged AUCs:

$$\begin{aligned}V &= \mathrm{var}(\widehat{\mathrm{AUC}}_{\mathrm{A}} - \widehat{\mathrm{AUC}}_{\mathrm{B}}) \\ &= \mathrm{var}(\widehat{\mathrm{AUC}}_{\mathrm{A}}) + \mathrm{var}(\widehat{\mathrm{AUC}}_{\mathrm{B}}) - 2\mathrm{cov}(\widehat{\mathrm{AUC}}_{\mathrm{A}}, \widehat{\mathrm{AUC}}_{\mathrm{B}}).\end{aligned} \tag{10}$$

Table 2 Coefficients of the moments that are found in the variance, Eq. (11) and in the covariance (Eq. 13). For the fully-crossed study design considered in this paper, $\underline{c}_A = \underline{c}_B = \underline{c}_{AB}$.

$[\underline{c}_A]_1 = \frac{1}{N_0 N_1 N_R}$	$[\underline{c}_A]_5 = \frac{(N_R-1)}{N_0 N_1 N_R}$
$[\underline{c}_A]_2 = \frac{(N_0-1)}{N_0 N_1 N_R}$	$[\underline{c}_A]_6 = \frac{(N_0-1)(N_R-1)}{N_0 N_1 N_R}$
$[\underline{c}_A]_3 = \frac{(N_1-1)}{N_0 N_1 N_R}$	$[\underline{c}_A]_7 = \frac{(N_1-1)(N_R-1)}{N_0 N_1 N_R}$
$[\underline{c}_A]_4 = \frac{(N_0-1)(N_1-1)}{N_0 N_1 N_R}$	$[\underline{c}_A]_8 = \frac{(N_0-1)(N_1-1)(N_R-1)}{N_0 N_1 N_R} - 1$

Without loss of generality, we first discuss the variance of $\widehat{\mathrm{AUC}}_A$. A similar discussion can be given for the variance $\widehat{\mathrm{AUC}}_B$. We will then discuss the covariance of $\widehat{\mathrm{AUC}}_A$ and $\widehat{\mathrm{AUC}}_B$.

2.3.1 *Variance*

The variance of a single modality can be decomposed into different representations. We shall use the success-moment representation that can be derived using U-statistics;[33] i.e.,

$$\mathrm{var}(\widehat{\mathrm{AUC}}_A) = \underline{c}_A^t \underline{M}_A, \tag{11}$$

where $\underline{c}_A$ is a vector of coefficients (see Table 2 for the coefficients of the fully-crossed study design considered in this paper) and $\underline{M}_A$ is a vector of eight product moments (see Table 3, every other row of column 1). Each element, $[\underline{M}_A]_l$ $(l = 1,2,\ldots,8)$, is the expected value of the product of two success observations from that modality. There are eight moments because the two success observations may come from the same or different readers (when reader subscripts match or not), the same or different nondiseased cases (when the nondiseased case subscripts match or not), and the same or different diseased cases (when the diseased case subscripts match or not). To see this clearly or to understand any of the discussion below, it may be useful to write a success observation in terms of the success function acting on a difference in ratings; recall $s_{Ajkk'} = s(X_{Ajk'1} - X_{Ajk0})$. Furthermore, it may even be necessary to write out the difference in ratings in terms of the constituent random effects.

The particular moment will be clear from the subscripts as, from this point forward, we will not allow different subscripts to take on the same value. The interpretation of each moment is driven by the unique subscripts that appear. For example, in the expression $[\underline{M}_A]_1 = E(s_{Ajkk'}s_{Ajkk'})$ we see that the subscripts are identical on both success observations; therefore, the expression is the expected value for a randomly selected reader reading a randomly selected diseased and nondiseased case. In contrast, in the expression $[\underline{M}_A]_8 = E(s_{Ajkk'}s_{Aj'k''k'''})$ we see that every subscript on the first success observation is different from its

Table 3 Each constituent moment of $\mathrm{var}(\mathrm{AUC}_A)$ and $\mathrm{cov}(\mathrm{AUC}_A, \mathrm{AUC}_B)$ calculated using the generalized R&M variance components. Each row is an equation for the variances $\sigma^2_{A(\cdot)}$ and $\sigma^2_{\Omega(\cdot)}$ that appear in Eq. (12) and (15). For example, in the row corresponding to $[\underline{M}_{AB}]_5$, $\sigma^2_{\Omega(5)} = \sigma^2_{R0} + \sigma^2_{RC0} + \sigma^2_{R1} + \sigma^2_{RC1}$. Note that the generalized R&M variance components are organized in columns, leaving spaces when variance components are not included.

Moments	Variance components						
$[\underline{M}_A]_1 = E(s_{Ajkk'}s_{Ajkk'})$:	Special case, refer to the text						
$[\underline{M}_{AB}]_1 = E(s_{Ajkk'}s_{Bjkk'})$:	$\sigma^2_{\Omega(1)} =$	0					
$[\underline{M}_A]_2 = E(s_{Ajkk'}s_{Ajk''k'})$:	$\sigma^2_{A(2)} =$		σ^2_{AC0}	$+\sigma^2_{ARC0}$			
$[\underline{M}_{AB}]_2 = E(s_{Ajkk'}s_{Bjk''k'})$:	$\sigma^2_{\Omega(2)} =$		σ^2_{C0}	$+\sigma^2_{RC0}$			
$[\underline{M}_A]_3 = E(s_{Ajkk'}s_{Ajkk''})$:	$\sigma^2_{A(3)} =$					σ^2_{AC1}	$+\sigma^2_{ARC1}$
$[\underline{M}_{AB}]_3 = E(s_{Ajkk'}s_{Bjkk''})$:	$\sigma^2_{\Omega(3)} =$					σ^2_{C1}	$+\sigma^2_{RC1}$
$[\underline{M}_A]_4 = E(s_{Ajkk'}s_{Ajk''k'''})$:	$\sigma^2_{A(4)} =$		σ^2_{AC0}	$+\sigma^2_{ARC0}$		$+\sigma^2_{AC1}$	$+\sigma^2_{ARC1}$
$[\underline{M}_{AB}]_4 = E(s_{Ajkk'}s_{Bjk''k'''})$:	$\sigma^2_{\Omega(4)} =$		σ^2_{C0}	$+\sigma^2_{RC0}$		$+\sigma^2_{C1}$	$+\sigma^2_{RC1}$
$[\underline{M}_A]_5 = E(s_{Ajkk'}s_{Aj'kk'})$:	$\sigma^2_{A(5)} =$	σ^2_{AR0}		$+\sigma^2_{ARC0}$	$+\sigma^2_{AR1}$		$+\sigma^2_{ARC1}$
$[\underline{M}_{AB}]_5 = E(s_{Ajkk'}s_{Bj'kk'})$:	$\sigma^2_{\Omega(5)} =$	σ^2_{R0}		$+\sigma^2_{RC0}$	$+\sigma^2_{R1}$		$+\sigma^2_{RC1}$
$[\underline{M}_A]_6 = E(s_{Ajkk'}s_{Aj'k''k'})$:	$\sigma^2_{A(6)} =$	σ^2_{AR0}	$+\sigma^2_{AC0}$	$+\sigma^2_{ARC0}$	$+\sigma^2_{AR1}$		$+\sigma^2_{ARC1}$
$[\underline{M}_{AB}]_6 = E(s_{Ajkk'}s_{Bj'k''k'})$:	$\sigma^2_{\Omega(6)} =$	σ^2_{R0}	$+\sigma^2_{C0}$	$+\sigma^2_{RC0}$	$+\sigma^2_{R1}$		$+\sigma^2_{RC1}$
$[\underline{M}_A]_7 = E(s_{Ajkk'}s_{Aj'kk''})$:	$\sigma^2_{A(7)} =$	σ^2_{AR0}		$+\sigma^2_{ARC0}$	$+\sigma^2_{AR1}$	$+\sigma^2_{AC1}$	$+\sigma^2_{ARC1}$
$[\underline{M}_{AB}]_7 = E(s_{Ajkk'}s_{Bj'kk''})$:	$\sigma^2_{\Omega(7)} =$	σ^2_{R0}		$+\sigma^2_{RC0}$	$+\sigma^2_{R1}$	$+\sigma^2_{C1}$	$+\sigma^2_{RC1}$
$[\underline{M}_A]_8 = E(s_{Ajkk'}s_{Aj'k''k'''})$:	Special case, refer to the text						
$[\underline{M}_{AB}]_8 = E(s_{Ajkk'}s_{Bj'k''k'''})$:	Special case, refer to the text						

counterpart on the second success observation; therefore, the expression is an expected value over a pair of randomly selected readers (that are unique), reading randomly selected diseased and nondiseased cases (that are all unique).

To derive the eight moments, we first treat two special cases, then we discuss the rest. The two special cases are the examples above: $[\underline{M}_A]_1$ and $[\underline{M}_A]_8$. In the first special case, the success observations are identical and both are equal to one or zero (ROC ratings from the generalized R&M model are continuous); consequently, their product equals one or zero. Therefore, $[\underline{M}_A]_1 = E(s_{Ajkk'}) = \mathrm{AUC}_A$. In the second special case, the success observations are independent because they come from different readers reading different nondiseased and diseased cases. Therefore, the expected value can be factored and $[\underline{M}_A]_8 = E(s_{Ajkk'})E(s_{Aj'k''k'''}) = \mathrm{AUC}_A^2$.

The remaining moments of $\underline{M}_A$ (moments 2 to 7) can be written as

$$[\underline{M}_A]_l = \int_{-\infty}^{\infty} \Phi\left(\frac{\Delta_A + x\sqrt{\sigma_\Omega^2 + \sigma_A^2 - \sigma_{\Omega(l)}^2 - \sigma_{A(l)}^2}}{\sqrt{\sigma_{\Omega(l)}^2 + \sigma_{A(l)}^2}}\right)^2 \phi(x)\mathrm{d}x, \tag{12}$$

where expressions for the variances $\sigma_{\Omega(l)}^2$ and $\sigma_{A(l)}^2$ ($l = 2,\ldots,7$) are listed in alternate rows of Table 3, and $\phi(\cdot)$ is the probability density function of the standard normal distribution. We point out that $\sigma_\Omega^2 + \sigma_A^2 - \sigma_{\Omega(l)}^2 - \sigma_{A(l)}^2$ is the sum of the variances of the random effects that are common to the two success observations within each moment.

Given a fully-specified R&M model, we can compute the moments above (and the moments corresponding to the variance of $\widehat{\mathrm{AUC}}_B$) using basic numerical integration methods. The derivation of the expression above follows a common outline. We exemplify the derivation for one of the moments in Appendix A. Here is the outline:

1. Identify the random effects that are common to the two success observations. These will be the effects corresponding to shared subscripts.
2. Write the moment as the expected value of the product of two conditional expected values, one for each success observation. Each conditional expected value assumes that the random effects found in Step 1 are fixed.
3. Recognize that each conditional expected value equals a probability that can be expressed in the form $\Phi(Z)$, where Z is a normal random variable.
4. Integrate over the random variables that were initially fixed in Step 1. This is a one-dimensional (1-D) integral of the product of two normal cumulative distribution functions and one normal probability distribution function.

2.3.2 *Covariance*

The covariance of $\widehat{\mathrm{AUC}}_A$ and $\widehat{\mathrm{AUC}}_B$ can be decomposed into a success-moment representation analogous to the variance; namely,

$$\mathrm{cov}(\widehat{\mathrm{AUC}}_A, \widehat{\mathrm{AUC}}_B) = \underline{c}_{AB}^t \underline{M}_{AB}, \tag{13}$$

where $\underline{c}_{AB} = \underline{c}_A$ for the fully-crossed study design considered in this paper (see Table 2) and $\underline{M}_{AB}$ is a vector of eight moments (see Table 3, every other row of column 1). Here, each moment is the expected value of the product of one success observation from modality A and another from B. For example,

$$[\underline{M}_{AB}]_5 = E(s_{Ajkk'}s_{Bj'kk'}), \tag{14}$$

where the two success observations come from different random readers reading the same cases in different modalities.

There is only one special case this time: $[\underline{M}_{AB}]_8$. The success observations for this moment are again independent because they again come from different readers reading different cases. Therefore, the expected value can be factored and $[\underline{M}_{AB}]_8 = E(s_{Ajkk'})E(s_{Bj'k''k'''}) = \mathrm{AUC}_A\mathrm{AUC}_B$.

The remaining moments of $\underline{M}_{AB}$ (moments 1 to 7) can be written as

$$[\underline{M}_{AB}]_l = \int_{-\infty}^{\infty} \Phi\left(\frac{\Delta_A + x\sqrt{\sigma_\Omega^2 - \sigma_{\Omega(l)}^2}}{\sqrt{\sigma_A^2 + \sigma_{\Omega(l)}^2}}\right) \times \Phi\left(\frac{\Delta_B + x\sqrt{\sigma_\Omega^2 - \sigma_{\Omega(l)}^2}}{\sqrt{\sigma_B^2 + \sigma_{\Omega(l)}^2}}\right)\phi(x)\mathrm{d}x, \tag{15}$$

where the variances $\sigma_{\Omega(l)}^2$ are listed in every other row of Table 3. We point out that $\sigma_\Omega^2 - \sigma_{\Omega(l)}^2$ is the sum of the variances of the random effects that are common to the two success observations within each moment.

Given a fully specified R&M model, we can compute the moments above using basic numerical integration methods. The derivation of the expression above follows the same common outline as above. We illustrate the derivation for $[\underline{M}_{AB}]_5$ in Appendix A.

2.4 *Design of Simulation Studies*

In this paper, we mimic the Monte Carlo (MC) simulation experiments run by R&M; however, we modify their experiments by perturbing the input variance components so that the variance components of the ROC ratings depend on modality and truth state. The purpose of the experiments here is to validate the numerical calculations given above against simulated results. Like R&M, we simulate data sets that are fully crossed: all readers read all images in both modalities with no re-reading. In what follows, we describe the original R&M simulations and detail how the simulations in this paper are built on the original.

R&M set to zero the effect for nondiseased cases and all four modality-truth interaction effects:

$$\alpha_0 = [\tau\alpha]_{A0} = [\tau\alpha]_{B0} = [\tau\alpha]_{A1} = [\tau\alpha]_{B1} = 0. \tag{16}$$

Consequently, the expected differences in ROC ratings for both modalities are equal to the effect for diseased cases ($\Delta_A = \Delta_B = \alpha_1$). R&M considered three levels of this effect; namely, $\alpha_1 = 0.75$, 1.50, and 2.50. These three levels ultimately control the expected AUCs via Eqs. (8) and (9). Now, since the variances of the random effects did not depend on modality, they effectively simulated a null-hypothesis experiment: $\mathrm{AUC}_A = \mathrm{AUC}_B$. The experiments in this paper will use the same values for the

fixed effects. However, because we perturb some variance components in order to generate ROC ratings that depend on modality and truth state (described below), AUC_A will not equal AUC_B and they will both be different from the AUCs of the original R&M simulation.

R&M investigated experiments with three or five readers and case-sets that ranged from 50 to 400. They also explored ratios of case mixes (nondiseased to diseased) of 1:1 and 9:1. We shall only investigate experiments with five readers, 50 nondiseased cases, and 50 diseased cases. We feel that limiting the investigations related to the size of the experiment is appropriate since the purpose here is more modest than the original purpose (validating a hypothesis test).

R&M based their original simulation experiments on correlation estimates found in actual ROC analyses using the CORROC algorithm,[34] which assumes an underlying bivariate binormal distribution. They investigated high versus low data correlation between and within readers (due to reading the same cases) at both high and low values of total reader variability. We explore a similar set of variance structures, but we perturb them to get variance structures that are different across modality and truth state. Specifically, we halve the variances of effects that involve modality A and truth = diseased, and we double the variances of effects that involve modality B and truth = diseased:

$$\begin{aligned} &\sigma^2_{AR1} = \sigma^2_{\tau R} \times 0.5, \quad \sigma^2_{AC1} = \sigma^2_{\tau C} \times 0.5, \quad \sigma^2_{ARC1} = \sigma^2_{\tau C} \times 0.5, \\ &\sigma^2_{BR1} = \sigma^2_{\tau R} \times 2.0, \quad \sigma^2_{BC1} = \sigma^2_{\tau C} \times 2.0, \quad \sigma^2_{BRC1} = \sigma^2_{\tau C} \times 2.0. \end{aligned} \tag{17}$$

This perturbation causes σ^2_A to be different from σ^2_B, which then causes AUC_A to be different from AUC_B [Eq. (9)].

In total, we explore two levels of between- and within-reader data correlation, four levels of reader variability, and three levels of performance in a factorial fashion for a total of 24 simulation configurations. Details are provided in Appendix B that will allow the interested reader to replicate our experiments.

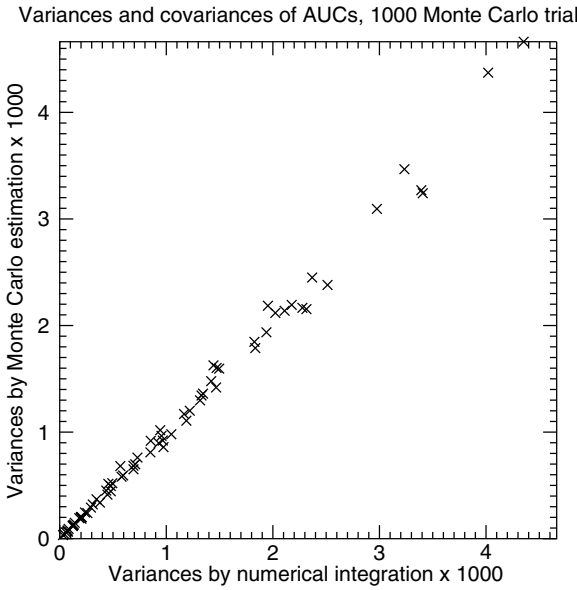

Fig. 1 The variances and covariances (estimates versus numerical results) of the 24 generalized R&M simulation configurations described in the text. The estimates are described in the text. The results here are for 1000 Monte Carlo (MC) trials. Differences converge as the number of MC trials increases.

3 Results

Figure 1 shows the variances and covariances of $\widehat{AUC}_A$ and $\widehat{AUC}_B$ (numerical results versus estimates) of the 24 simulation experiments described above: 8 configurations × 3 performance levels. The x-axis shows the variances by numerical integration. The y-axis shows the MC estimates of variance given 1000 MC trials. There is some variability along the line of equality due to the relatively small number of MC trials. The absolute value of the relative differences between the numerical results and the estimates averaged over all 24 simulation configurations is 5.2%. For 100,000 MC trials, the averaged absolute value of the relative differences is 0.5% and the variability along the line of equality is not visible (not shown).

4 Conclusions

In this paper, we generalized the R&M model by allowing all of its variance components to depend on both modality and truth state. This will allow investigators to model and simulate MRMC ROC studies that better fit their data. Additionally, we presented and validated equations for computing the population variances and covariances for empirical AUC estimates computed from data simulated from the generalized R&M model. These equations show the core relationships between the ROC data and the reader-averaged AUCs. These equations and relationships should help investigators to validate new MRMC variance and covariance estimation methods, explore novel study designs, size future trials, and model MRMC ROC data.

For the interested investigator, the first author has made available a stand-alone Java application called iRoeMetz. iRoeMetz and its source code can be found in Ref. 35. iRoeMetz can simulate ROC studies according to the generalized R&M model and numerically calculate the expected moments and variances found in this paper. Additionally, all the original R&M simulation configurations and those used in this paper can be downloaded and tested.

Future work is needed in this area, investigating real data sets and calibrating the different possible R&M models of decision scores to be consistent with real data estimates of AUC variance components. To cover the full range, datasets that study different anatomical locations and diseases, and different imaging modalities (with and without computer aids as appropriate) are needed. The results of such future work would provide investigators with more information and examples to help them calibrate the R&M model (or any other model) to their situation so that they can appropriately size future reader studies. We are collecting datasets for this work for public sharing toward this goal; we welcome any and all contributions.

There are other interesting future efforts that could support the need mentioned above. One direction of study could link the success moments at the core of the work here to the parameters of the Obuchowski and Rockette[30] model for AUC observations. Another direction of study could be to uncover the inverse mapping: given the parameters of the Obuchowski and Rockette model for AUC observations, solve for the variance components of the ROC ratings that are consistent with them. It is not currently known whether the solution is unique or if it can be treated efficiently and effectively.

Appendix A

In this appendix, we derive one of the product moments, $[\underline{M}_{AB}]_5$, that enters into the computation of the covariance between $\widehat{AUC}_A$ and $\widehat{AUC}_B$ via Eq. (13). Adapting the results to the

other moments that are required for computing the variance of $\widehat{\mathrm{AUC}}_{\mathrm{A}}$ and the variance of $\widehat{\mathrm{AUC}}_{\mathrm{B}}$ is straightforward.

Although the generalized R&M model has many parameters and the variances and the covariance of the reader-averaged AUCs are fairly complex, all of the success moments are derived the same way and the results have a common form. The derivations follow a common outline. The results are straightforward 1-D integrals involving the normal distribution that can be calculated using basic numerical integration methods. We illustrate the derivation for $[\underline{M}_{\mathrm{AB}}]_5$.

Here is the derivation outline:

1. Identify the random effects that are common to the two success observations. These will be the effects corresponding to shared subscripts.
2. Write the moment as the expected value of the product of two conditional expected values, one for each success observation. Each conditional expected value assumes that the random effects found in Step 1 are fixed.
3. Recognize that each conditional expected value equals a probability that can be expressed in the form $\Phi(Z)$, where Z is a normal random variable equal to a difference in ROC ratings with the random effects found in Step 1 temporarily fixed.
4. Integrate over the random variables that were initially fixed in Step 1. This is a 1-D integral of the product of two normal cumulative distribution functions and one normal probability distribution function.

Step 1: The fifth product moment is

$$[\underline{M}_{\mathrm{AB}}]_5 = E(s_{\mathrm{A}jkk'}s_{\mathrm{B}j'kk'}) = E[s(X_{\mathrm{A}jk'1} - X_{\mathrm{A}jk0})s(X_{\mathrm{B}j'k'1} - X_{\mathrm{B}j'k0})]. \quad (18)$$

The random effects that are common to both success functions are the ones that do not depend on reader or modality: $[C\alpha]_{k'1}, [C\alpha]_{k0}$. Note that these random effects are the effects corresponding to shared subscripts. This can be seen if we write out the differences in ROC ratings in terms of the constituent random effects.

Step 2: Here, we first write the moment as the expected value of a conditional expected value, where the random effects found in Step 1 are fixed; namely,

$$[M_{\mathrm{AB}}]_5 = E[E(s_{\mathrm{A}jkk'}s_{\mathrm{B}j'kk'}|[C\alpha]_{k'1}, [C\alpha]_{k0})] \quad (19)$$

With the common random effects fixed, $s_{Ajkk'}$ is independent of $s_{\mathrm{B}j'kk'}$. Therefore, we can factor the conditional expected value to obtain

$$[M_{\mathrm{AB}}]_5 = E[E(s_{\mathrm{A}jkk'}|[C\alpha]_{k'1}, [C\alpha]_{k0}) \times E(s_{\mathrm{B}j'kk'}| \times [C\alpha]_{k'1}, [C\alpha]_{k0})]. \quad (20)$$

Step 3: In the expected value concerning modality A, the argument of the success function is a normal random variable Z with mean and conditional variance given by

$$\mu_Z = \Delta_{\mathrm{A}} + [C\alpha]_{k'1} - [C\alpha]_{k0}, \quad (21)$$

$$\sigma_Z^2 = \sigma_{R0}^2 + \sigma_{RC0}^2 + \sigma_{R1}^2 + \sigma_{RC1}^2 + \sigma_{AR0}^2 + \sigma_{AC0}^2 + \sigma_{\mathrm{A}RC0}^2 + \sigma_{\mathrm{A}R1}^2 + \sigma_{\mathrm{A}C1}^2 + \sigma_{\mathrm{A}RC1}^2. \quad (22)$$

Using the definitions given in Eq. (4) and Table 3, we see that $\sigma_Z^2 = \sigma_{\mathrm{A}}^2 + \sigma_{\Omega(5)}^2$. We have a similar result for the argument of the success function in the expected value concerning modality B. Since the argument to each conditional expected value is a normal random variable, we can write

$$[M_{\mathrm{AB}}]_5 = E\left[\Phi\left(\frac{\Delta_{\mathrm{A}} + [C\alpha]_{k'1} - [C\alpha]_{k0}}{\sqrt{\sigma_{\mathrm{A}}^2 + \sigma_{\Omega(5)}^2}}\right) \times \Phi\left(\frac{\Delta_{\mathrm{B}} + [C\alpha]_{k'1} - [C\alpha]_{k0}}{\sqrt{\sigma_{\mathrm{B}}^2 + \sigma_{\Omega(5)}^2}}\right)\right]. \quad (23)$$

Step 4: Last, we see that the only randomness that remains above is the difference $[C\alpha]_{k'1} - [C\alpha]_{k0}$. This difference is a normal random variable $Y = [C\alpha]_{k'1} - [C\alpha]_{k0}$ with mean zero and variance $\sigma_{\Omega}^2 - \sigma_{\Omega(5)}^2 = \sigma_{C0}^2 + \sigma_{C1}^2$. After a change of variables, $X = Y/\sqrt{\sigma_{\Omega}^2 - \sigma_{\Omega(5)}^2}$ we can effectively integrate over the random variables that were initially fixed, $[C\alpha]_{k'1}, [C\alpha]_{k0}$, with the following:

$$[\underline{M}_{\mathrm{AB}}]_5 = \int_{-\infty}^{\infty} \Phi\left(\frac{\Delta_{\mathrm{A}} + x\sqrt{\sigma_{\Omega}^2 - \sigma_{\Omega(5)}^2}}{\sqrt{\sigma_{\mathrm{A}}^2 + \sigma_{\Omega(5)}^2}}\right) \times \Phi\left(\frac{\Delta_{\mathrm{B}} + x\sqrt{\sigma_{\Omega}^2 - \sigma_{\Omega(5)}^2}}{\sqrt{\sigma_{\mathrm{B}}^2 + \sigma_{\Omega(5)}^2}}\right)\phi(x)\mathrm{d}x. \quad (24)$$

This last expression can be calculated using basic numerical integration methods. In our software, we simply sample the 1-D integral at 256 points on the interval $(-10, 10)$ and use the midpoint rule (rectangle rule).

Appendix B

Here, we document the simulation configurations used in this paper. We start by defining original R&M variance components. We then map them to generalized R&M variance components (Table 1). Finally, we perturb them to get variance structures that are different across modality and truth state.

The simulation configurations in this paper explore the following settings in a factorial fashion for 5 readers, 50 nondiseased cases, and 50 diseased cases:

- Two levels of between- and within-reader data correlation:

  $\mathrm{High}\,\rho_{\mathrm{BR}}$: $\sigma_C^2 = 0.3$ $\sigma_{\tau C}^2 = 0.3$ $\sigma_{RC}^2 = 0.2$ $\sigma_{\tau RC}^2 = 0.2$
  $\mathrm{Low}\,\rho_{\mathrm{BR}}$: $\sigma_C^2 = 0.1$ $\sigma_{\tau C}^2 = 0.1$ $\sigma_{RC}^2 = 0.2$ $\sigma_{\tau RC}^2 = 0.6$

- Four levels of reader variability: $\sigma_R^2 = \sigma_{\tau R}^2 = 0.0055$, 0.011, 0.030, and 0.056.
- Three levels of performance: $\Delta_{\mathrm{A}} = \Delta_{\mathrm{B}} = \alpha_1 = 0.75$, 1.50, and 2.5.

Let us consider the simulation with the high between- and within-reader data correlation and the lowest reader variability. After mapping and perturbing, the generalized R&M variance components are

$$\begin{array}{lll} \sigma^2_{R0} = 0.0055, & \sigma^2_{C0} = 0.3, & \sigma^2_{RC0} = 0.2, \\ \sigma^2_{R1} = 0.00550, & \sigma^2_{C1} = 0.30, & \sigma^2_{RC1} = 0.2, \\ \sigma^2_{AR0} = 0.0055, & \sigma^2_{AC0} = 0.3, & \sigma^2_{ARC0} = 0.2, \\ \sigma^2_{AR1} = 0.00275, & \sigma^2_{AC1} = 0.15, & \sigma^2_{ARC1} = 0.1, \\ \sigma^2_{BR0} = 0.0055, & \sigma^2_{BC0} = 0.3, & \sigma^2_{BRC0} = 0.2, \\ \sigma^2_{BR1} = 0.01100, & \sigma^2_{BC1} = 0.60, & \sigma^2_{BRC1} = 0.4. \end{array}$$

This variance structure used with the lowest level of performance ($\Delta_A = \Delta_B = \alpha_1 = 0.75$) leads to the following expectations of performance and uncertainty for 5 readers, 50 nondiseased cases, and 50 diseased cases:

$$\begin{array}{ll} \mathrm{AUC_A} = 0.714, & \sqrt{\mathrm{var}(\widehat{\mathrm{AUC}}_A)} = 0.044, \\ \mathrm{AUC_B} = 0.681, & \sqrt{\mathrm{var}(\widehat{\mathrm{AUC}}_B)} = 0.035, \\ \mathrm{AUC_A} - \mathrm{AUC_B} = 0.032, & \sqrt{\mathrm{var}(\widehat{\mathrm{AUC}}_A - \widehat{\mathrm{AUC}}_B)} = 0.047. \end{array}$$

Acknowledgments

This research was partially supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB013667. The content is solely the responsibility of the authors and does not necessarily represent the offical views of the National Institutes of Health, the Department of Veterans Affairs, or the United States government.

References

1. B. D. Gallas et al., "Evaluating imaging and computer-aided detection and diagnosis devices at the FDA," *Acad. Radiol.* **19**, 463–477 (2012).
2. C. A. Roe and C. E. Metz, "Dorfman-Berbaum-Metz method for statistical analysis ofmultireader, multimodality receiver operating characteristic (ROC) data: validation with computer simulation," *Acad. Radiol.* **4**(4), 298–303 (1997).
3. D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method," *Invest. Radiol.* **27**(9), 723–731 (1992).
4. S. V. Beiden, R. F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis," *Acad. Radiol.* **7**(5), 341–349 (2000).
5. S. V. Beiden et al., "Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis," *Acad. Radiol.* **8**(7), 616–622 (2001).
6. X. Song and X.-H. Zhou, "A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data," *Biostatistics* **6**(2), 303–312 (2005).
7. S. L. Hillis and K. S. Berbaum, "Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification," *Acad. Radiol.* **12**(12), 1534–1541 (2005).
8. S. L. Hillis, "A comparison of denominator degrees of freedom methods for multiple observer ROC analysis," *Stat. Med.* **26**(3), 596–619 (2007).
9. S. L. Hillis, K. S. Berbaum, and C. E. Metz, "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Acad. Radiol.* **15**(5), 647–661 (2008).
10. A. Skaron, K. Li, and X.-H. Zhou, "Statistical methods for MRMC ROC studies," *Acad. Radiol.* **19**(12), 1499–1507 (2012).
11. D. P. Chakraborty, "Prediction accuracy of a sample-size estimation method for ROC studies," *Acad. Radiol.* **17**(5), 628–638 (2010).
12. S. L. Hillis, N. A. Obuchowski, and K. S. Berbaum, "Power estimation for multireader ROC methods an updated and unified approach," *Acad. Radiol.* **18**(2), 129–142 (2011).
13. D. D. Dorfman et al., "Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design," *Acad. Radiol.* **5**(9), 591–602 (1998).
14. R. F. Wagner, S. V. Beiden, and C. E. Metz, "Continuous versus categorical data for ROC analysis: some quantitative considerations," *Acad. Radiol.* **8**(4), 328–334 (2001).
15. N. Obuchowski, B. D. Gallas, and S. L. Hillis, "Multi-reader ROC studies with split-plot designs: a comparison of statistical methods," *Acad. Radiol.* **19**(12), 1508–1517 (2012), Invited paper for Special Metz Memorial Issue I.
16. B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am. A, Spec. Issue on Image Qual.* **24**(12), B70–B80 (2007).
17. W. Chen et al., "A general framework for MRMC reader studies with binary assessments: simulation for validation and sizing," *JMI*, Submitted (2014).
18. C. K. Abbey, F. W. Samuelson, and B. D. Gallas, "Statistical power considerations for a utility endpoint in observer performance studies," *Acad. Radiol.* **20**(7), 798–806 (2013), Invited paper for Special Metz Memorial Issue II.
19. S. L. Hillis, "Simulation of unequal-variance binormal multireader ROC decision data: an extension of the Roe and Metz simulation model," *Acad. Radiol.* **19**(12), 1518–1528 (2012).
20. S. V. Beiden et al., "Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities," *Acad. Radiol.* **8**(7), 605–615 (2001).
21. J. Swets, W. P. Tanner, and T. G. Birdsall, "Decision processes in perception," *Psychol. Rev.* **68**(5), 301–340 (1961).
22. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, Wiley, New York (1966), [Reprint Krieger, New York (1974)].
23. H. P. Chan et al., "Digital mammography: observer performance study of the effects of pixel size on the characterization of malignant and benign microcalcifications," *Acad. Radiol.* **8**(6), 454–466 (2001).
24. B. Sahiner et al., "Multi-modality CADx: ROC study of the effect on radiologists' accuracy in characterizing breast masses on mammograms and 3D ultrasound images," *Acad. Radiol.* **16**(7), 810–818 (2009).
25. E. A. Rafferty et al., "Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial," *Radiology* **266**(1), 104–113 (2013).
26. F. W. Samuelson and X. He, "Comparison of semiparametric receiver operating characteristic models on observer data," *J. Med. Imag.***1**(3), 031004 (2014).
27. D. D. Dorfman and K. S. Berbaum, "A contaminated binormal model for ROC data: Part II. A formal model," *Acad. Radiol.* **7**(6), 427–437 (2000).
28. C. E. Metz and X. Pan, "'Proper' binormal ROC curves: theory and maximum-likelihood estimation," *J. Math. Psychol.* **43**(1), 1–33 (1999).
29. L. L. Pesce and C. E. Metz, "Reliable and computationally efficient maximum-likelihood estimation of "proper" binormal ROC curves," *Acad. Radiol.* **14**(7), 814–829 (2007).
30. N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations," *Commun. Stat. B-Simul.* **24**(2), 285–308 (1995).
31. R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences*, 3rd ed., Wadsworth Publishing, Belmont, CA (1994).
32. E. R. DeLong, D. M. Delong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
33. B. D. Gallas et al., "A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators," *Commun. Stat. A-Theory* **38**(15), 2586–2603 (2009).
34. C. E. Metz, P. L. Wang, and K. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from

correlated data," in *Information Processing in Medical Imaging VIII*, F. Deconick, Ed., pp. 432–445, Springer, Netherlands (1984).
35. B. D. Gallas, "iROEMETZ v1.2: Application for Simulating MRMC Reader Studies," code.google.com/p/imrmc/wiki/iRoeMetzGuide (4 August 2014).

Brandon D. Gallas provides mathematical, statistical, and modeling expertise to the evaluation of medical imaging devices at the FDA. His main areas of contribution are in the design and statistical analysis of reader studies, image quality, computer-aided diagnosis, and imaging physics. Before working at the FDA, he was in Dr. Harrison Barrett's research group at the University of Arizona, earning his PhD from the Graduate Interdisciplinary Program in Applied Mathematics.

Stephen L. Hillis is a research professor in the Departments of Radiology and Biostatistics at the University of Iowa, and Senior Statistician for the Iowa City VA Healthcare System. He earned his PhD in statistics from the University of Iowa in 1987 and is the author of 85 peer-reviewed articles. Since 1998 his research has focused on methodology for analyzing diagnostic radiological imaging studies.