MRMC reader studies with binary agreement data: simulation, analysis, validation, and sizing

Weijie Chen, Adam Wunderlich, Nicholas Petrick, Brandon D. Gallas

Food and Drug Administration, Center for Devices and Radiological Health,

Office of Science and Engineering Laboratories, Division of Imaging and Applied Mathematics

10903 New Hampshire Avenue, Silver Spring, MD, 20993

Address all correspondence to:

Weijie Chen, PhD

FDA/CDRH/OSEL

10903 New Hampshire Avenue

Silver Spring, MD, 20993

Tel: +1 301-796-2663

Fax: +1 301-796-9925

E-mail: weijie.chen@fda.hhs.gov

Abstract. In this paper we treat multi-reader multi-case (MRMC) reader studies for which a reader's diagnostic assessment is converted to binary agreement (1: agree with the truth state, 0: disagree with the truth state). We present a mathematical model for simulating binary MRMC data with a desired correlation structure across readers, cases, and two modalities, assuming the expected probability of agreement is equal for the two modalities ($P_1 = P_2$). This model can be used to validate the coverage probabilities of 95% confidence intervals (of P_1, P_2, or $P_1 - P_2$ when $P_1 - P_2 = 0$), validate the type I error of a superiority hypothesis test, and size a non-inferiority hypothesis test (which assumes $P_1 = P_2$). To illustrate the utility of our simulation model, we adapt the Obuchowski-Rockette-Hillis (ORH) method for the analysis of MRMC binary agreement data. Moreover, we use our simulation model to validate the ORH method for binary data and to illustrate sizing in a non-inferiority setting. Our software package is publicly available on the Google code project hosting site (`http://code.google.com/p/imrmc/wiki/iMRMC_Binary`) for use in simulation, analysis, validation, and sizing of MRMC reader studies with binary agreement data.

Keywords: reader study, multi-reader multi-case, Monte Carlo simulation, sizing, binary data.

1 Introduction

The effectiveness of a new imaging modality is typically established by showing that the performance of physicians using the new imaging modality is either non-inferior[1] or superior[2] to a conventional imaging modality. When the assessment is based on a sample of representative physicians, or "readers", who read images from the new and conventional imaging modalities for a sample of representative patients, or "cases," it is considered a multi-reader, multi-case (MRMC) study.[3,4] By statistically treating both the readers and cases as random samples, the conclusion(s) regarding the effectiveness of a new device can be generalized to both the population of readers and the population of cases, i.e., the findings are not limited to a particular set of study readers or cases.[5]

The research literature on methodologies for study designs, sizing, and data analysis of MRMC reader studies has a history of over 30 years going back to the seminal work of Swets and Pickett.[6] This research, summarized in the next paragraph, has focused on the area under the receiver operating characteristic curve (AUC) as a summary figure of merit.[7] This long-term focus could lead to the erroneous conclusion that MRMC assessment is only for AUC. Instead, MRMC analysis should be understood to treat readers and cases as random regardless of the endpoint. However, the research on MRMC analysis for AUC has laid a solid foundation for MRMC analysis for other endpoints such as the binary endpoint that we will investigate in this paper.

The first MRMC data analysis method that gained wide practical use was introduced by Dorfman, Berbaum, and Metz (DBM).[8] Roe and Metz subsequently developed a simulation model[9] to validate the DBM method and a variance-component approach[10] to model different sources of variations in the MRMC problem; generalizations of the Roe and Metz model have also been proposed.[11,12] Several alternative analysis methods have been developed and validated over the years, for example, methods by Obuchowski and Rockette (OR),[13] Beiden, Wagner, and Campbell,[14] Song and Zhou,[15] and Gallas,[16,17] to name a few. More recently, Hillis et al.[18–21] refined the degrees of freedom estimate in the OR method and showed that the

OR method is equivalent to the DBM method under certain circumstances. We call the Hillis-refined OR method the ORH method. Although the ORH method was designed for superiority studies, it has been adopted to non-inferiority settings.[22] Wagner, Metz, and Campbell[4] provided a nice review and tutorial on medical imaging system assessment and MRMC study designs in reviewing the most common analysis methods and summarizing many practical study design issues. There is a general consensus in the medical imaging community supporting the use of MRMC methods and AUC to evaluate medical imaging systems, as summarized in a paper by Gallas et al.[23]

The use of ROC metrics requires the reader assessment to consist of scores on an ordinal scale and the reference standard ("truth") be binary. However, these requirements are not always met in clinical MRMC studies. Examples include some studies for the evaluation of whole slide imaging (WSI) digital pathology devices.[24–28] In these studies, pathologists operate under real-world sign-out conditions; they complete a full diagnostic report, which may include a complicated checklist or orders for additional tests or samples (i.e., their assessments can hardly be ordinal). In addition, the reference standard may be the original clinical report or may be the diagnostic report of an expert pathologist or panel, which is rarely a clear-cut binary truth-state. As a result, these studies often use agreement with the reference as the endpoint, where the agreement is determined by a comparison between the study pathologist's diagnostic report and a reference diagnostic report: "1" means that the diagnostic report by the study reader agrees with the reference, and "0" means that it does not agree.

As such, the binary agreement data is binary *outcome* data (agrees with reference, disagrees with reference). It should not be confused with binary *response* data, which is a binary assessment by the reader (e.g., cancer, non-cancer). Of course a binary response can be converted to a binary outcome by comparing with a reference standard, but other types of response data (e.g., categorical data) can be converted to binary outcome data as well.

A natural summary performance metric for the binary outcome data described above is the probability of agreement with the reference standard, which we denote as P. In practice, it is sometimes called percent correct or percent agreement. The performance of a reader is estimated by averaging his/her binary agreement data over the cases. The mean performance for a modality is estimated by averaging the readers' probability of agreement or, equivalently, by averaging all the binary agreement data over the cases and over the readers. We note that there are some implicit assumptions in using the probability of agreement as a performance metric, for example, it weights different kinds of errors equally as well as mixing cases of varying degrees of disease.

In this paper, we do not aim to compare different types of data collection and performance metrics (e.g., ROC vs. binary) and hence a detailed discussion of those topics is beyond the scope of this paper. The purpose of this paper is to investigate the application of MRMC methods to studies that yield binary agreement data such as the WSI studies mentioned above. The variability of performance estimates in MRMC studies arises from both the random readers and the random cases. Both sources of variability should be accounted for appropriately if the performances of imaging modalities are to be generalized to both the reader and case populations. To our knowledge, methodologies for sizing a MRMC study with binary outcomes and analyzing MRMC binary outcome data accounting for both sources of variability have not been fully established. Therefore, our goals in this paper include (i) developing a model for simulating MRMC binary agreement data and using this simulation model to (ii) validate a potential analysis method and (iii) size/power a new study.

The existing MRMC methodology literature provides a strong foundation to tackle the MRMC problem more generally, and conventional MRMC analysis methods can be adapted to studies that only include binary data. For example, Gallas et al. adapted the one-shot, U-statistics method[16] to estimate the variance of the binary performance metrics.[29] Likewise, as we show in this paper, the ORH method[13,18,21] can be adapted to binary data because it applies to any performance measure in a MRMC setting, whether it is AUC or probability of agreement.

In this paper, we develop a Monte Carlo simulation model for MRMC reader studies with binary agreement data. The MRMC data can be generated with a desired correlation structure across readers, cases, and two modalities, assuming the expected probability of agreement is equal for the two modalities ($P_1 = P_2$). This model can be used to validate the coverage probabilities of 95% confidence intervals (of P_1, P_2, or $P_1 - P_2$ when $P_1 - P_2 = 0$), validate the type I error of a superiority hypothesis test, and size a non-inferiority hypothesis test (which assumes $P_1 = P_2$).

The rest of the paper is organized as follows. In section 2, we present our simulation model for MRMC binary data and illustrate how to use it for (i) validating the coverage probability of confidence intervals for $P_1 - P_2$ when $P_1 = P_2$, and (ii) sizing a non-inferiority hypothesis test with desired statistical power (assuming $P_1 = P_2$). We demonstrate these procedures by applying the adapted ORH analysis method to data simulated with hypothetical parameters. In section 3, we present results of these simulations.. We conclude in Section 4 with discussions of future work.

2 Methods

2.1 Simulation model

Our simulation model is a threshold model, i.e., we generate latent continuous data and then apply a threshold to obtain binary data. The user of the model specifies parameters that characterize the binary data. These parameters are mapped to parameters that control the distribution of the latent continuous data. A diagram overview of our simulation model is given in Figure 1 and details are given below.

2.1.1 Notation and parameters characterizing binary MRMC data

We assume a two-modality fully-crossed MRMC design, in which all readers read all cases on both modalities. In this setting, we denote the number of readers as N_r, the number of cases as N_c, and the binary outcome for the i^{th} modality, the j^{th} reader, and the k^{th} case as Y_{ijk} ($Y_{ijk} = 1$ if the reader's assessment agrees with the reference and $Y_{ijk} = 0$ if it disagrees with the reference), where $i = 1, 2$, $j = 1, 2, \ldots, N_r$,

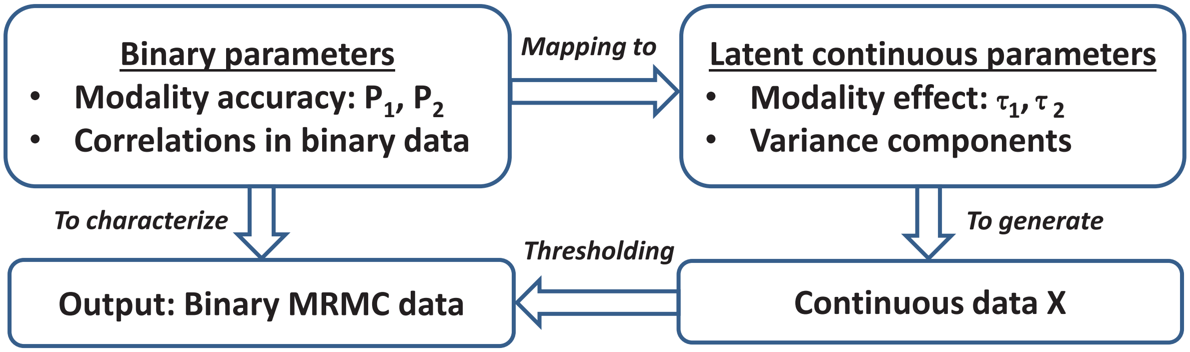
Binary parameters
• Modality accuracy: P_1, P_2
• Correlations in binary data
Mapping to
Latent continuous parameters
• Modality effect: τ_1, τ_2
• Variance components
To characterize
To generate
Thresholding
Output: Binary MRMC data
Continuous data X

Figure 1: An overview of the procedure for simulating binary MRMC data (additional input parameters include the number of readers N_r and the number of cases N_c).

and $k = 1, 2, \ldots, N_c$. The probability of agreement for modality i is the expectation of Y_{ijk} over the population of readers and the population of cases, i.e., $P_i = \mathrm{E}[Y_{ijk}]$. We assume that the mean performance is the same for both modalities, i.e., $P_1 = P_2$, which is restrictive but tractable and useful as we shall show. Extension to a more general model to overcome this restriction is planned future work as discussed in Section 4.

For a fully-crossed design, correlations are induced in the binary scores in three ways: between modalities (because each modality is evaluated by the same readers reading the same cases), between readers (because each reader reads the same cases), and between cases (because each case is read by the same readers). We assume that (i) these correlations are independent of any particular pair of readers or any particular pair of cases and (ii) the binary scores are independent when both the readers and cases are different. These are standard MRMC modeling assumptions that are made as a trade-off between capturing major characteristics in real data and controlling the complexity (the number of parameters) of the model. Consequently, there are a total of eight distinct correlations given the three factors (modality, reader, and case) and our assumptions, as summarized in Table 1. Each correlation is denoted with some combination of the subscripts τ, r and c to indicate that the correlation is for different modalities, readers, or cases, respectively. Of the eight correlations, five of them are non-trivial and are used in our model to characterize the binary data Y_{ijk} and three of them are trivial constants under our model assumptions. Range conditions for the correlations listed in Table 1 are consistent with our simulation model described below, and can be understood intuitively

Table 1: Binary data correlation parameters.

definition	description	range conditions
ρ_c	same modality, same reader, different case	$\rho_c \geq \rho_{\tau c}$
ρ_r	same modality, different reader, same case	$\rho_r \geq \rho_{\tau r}$
ρ_τ	different modality, same reader, same case	$\rho_\tau \geq \rho_{\tau c}, \rho_{\tau r}$
$\rho_{\tau c}$	different modality, same reader, different case	$\rho_{\tau c} \geq 0$
$\rho_{\tau r}$	different modality, different reader, same case	$\rho_{\tau r} \geq 0$
ρ_{rc}	same modality, different reader, different case	$\rho_{rc} \equiv 0$
$\rho_{\tau rc}$	different modality, different reader, different case	$\rho_{\tau rc} \equiv 0$
ρ_0	same modality, same reader, same case	$\rho_0 \equiv 1$

on the grounds that the correlation between readings will be stronger when more reading conditions are the same. For example, ρ_c is the correlation when the modality and reader are the same, whereas $\rho_{\tau c}$ is the correlation when only the reader is the same, and therefore, $\rho_c \geq \rho_{\tau c}$. Summarizing, for a fixed number of readers, N_r, and a fixed number of cases, N_c, binary MRMC data with $P_1 = P_2 = P$ are described by six parameters: one probability of agreement P, and five correlations.

2.1.2 Data generation

To simulate binary MRMC data with the parameters defined above, we first need to map the parameters in the binary score domain to parameters in the latent continuous domain (Top half of Figure 1). However, we defer the description of this mapping mechanism to section 2.1.3 when parameters in both domains are defined. The binary data generation is a two-step process (Bottom half of Figure 1). In the first step, we generate continuous-valued MRMC data using a model similar to the well-known Roe-Metz[9] method for simulating continuous MRMC ROC data, but with the dependence on truth state removed. In this way, all the correlations in Table 1 are built into the data. In the second step, the continuous valued data are dichotomized to obtain binary MRMC data. Further details are given below.

We start by generating realizations of a latent, continuous random variable, X_{ijk}, which represents the propensity to agree, with the mixed-effect model

$$X_{ijk} = \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + e_{ijk}, \tag{1}$$

where τ_i is the fixed effect of modality i, R_j is the random effect for reader j, C_k is the random effect for case k, and $(\tau R)_{ij}$, $(\tau C)_{ik}$, and $(RC)_{jk}$ are the corresponding two-way interactions. The term e_{ijk} effectively includes two terms: a three-way interaction term $(\tau RC)_{ijk}$ and a pure random error term that represents the reader's inability to exactly reproduce his/her assessments when reading the same case multiple times. The six random terms are assumed to be independent, normal random variables with zero means and variances σ_R^2, σ_C^2, $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$, σ_{RC}^2, and σ_e^2, respectively, so that the total variance is $\sigma_X^2 = \sigma_R^2 + \sigma_C^2 + \sigma_{\tau R}^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_e^2$. Note that each of the six variance components arises from one or both sources of randomness, namely the randomness of the readers (σ_R^2, $\sigma_{\tau R}^2$, σ_{RC}^2, σ_e^2) and the randomness of cases (σ_C^2, $\sigma_{\tau C}^2$, σ_{RC}^2, σ_e^2). The decomposition of the variance into independent components allows each component to be generated independently. Once the latent continuous data, X_{ijk}, has been generated, binary MRMC data, Y_{ijk}, is obtained by thresholding with zero, i.e.,

$$Y_{ijk} = \begin{cases} 0 & \text{if } X_{ijk} \leq 0 \\ \\ 1 & \text{if } X_{ijk} > 0 \end{cases}. \tag{2}$$

2.1.3 Parameter mapping

Note that there is a fundamental difference between our binary data simulation and conventional continuous MRMC data simulation, as exemplified by the well-known Roe-Metz model.[9] In the Roe-Metz model, the model parameters that are needed to generate the data are the same as the measurable parameters to characterize the data. By contrast, in binary data simulation, the measurable parameters for characterizing the data (i.e., the probability of agreement P and the five correlations defined in Table 1) must be mapped back to unobservable parameters in the latent space to generate the data (the fixed modality effect τ_i and the six variance component parameters, i.e., the sigmas) (See top half of Figure 1). Because there is one more free parameter in the latent space, we apply an additional constraint to the latent-space parameters to make the mapping unique. Namely, without loss of generality, we choose to set the total variance of X to be unity (i.e., $\sigma_X = 1$).

With the constraint $\sigma_X = 1$ and the range conditions in Table 1, there is a unique mapping between the parameters in the two domains. First, the expected performance P is uniquely mapped to the fixed modality effect τ_i. From the definition of P_i and Equation 2, we have

$$P_i = \mathrm{E}[Y_{ijk}] = \Pr\left(X_{ijk} > 0\right) = \Phi(\tau_i/\sigma_X) = \Phi(\tau_i),$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. With the assumption $P_1 = P_2 = P$, we obtain $\tau_1 = \tau_2 = \Phi^{-1}(P)$.

Second, the binary correlation parameters are uniquely mapped to variance component parameters. The relations between the binary correlation parameters and the latent variance component parameters, under the assumption $P_1 = P_2 = P$, are shown in Equations 3-7, where $\Psi\left(z_1, z_2; r\right)$ is the CDF of the standard bivariate normal distribution with correlation r (one way to derive these is to use a result in Emrich and Piedmonte[30]).

$$\rho_c = \frac{\Psi\left(\Phi^{-1}(P), \Phi^{-1}(P); \sigma_R^2 + \sigma_{\tau R}^2\right) - P^2}{P(1-P)} \tag{3}$$

$$\rho_r = \frac{\Psi\left(\Phi^{-1}(P), \Phi^{-1}(P); \sigma_C^2 + \sigma_{\tau C}^2\right) - P^2}{P(1-P)} \tag{4}$$

$$\rho_\tau = \frac{\Psi\left(\Phi^{-1}(P), \Phi^{-1}(P); \sigma_R^2 + \sigma_C^2 + \sigma_{RC}^2\right) - P^2}{P(1-P)} \tag{5}$$

$$\rho_{\tau c} = \frac{\Psi\left(\Phi^{-1}(P), \Phi^{-1}(P); \sigma_R^2\right) - P^2}{P(1-P)} \tag{6}$$

$$\rho_{\tau R} = \frac{\Psi\left(\Phi^{-1}(P), \Phi^{-1}(P); \sigma_C^2\right) - P^2}{P(1-P)} \tag{7}$$

Using these equations, we can find the variance component parameters (i.e., the σ parameters on the right-hand side of the equations) given binary correlations (i.e., the ρ parameters on the left-hand side of the equations) through a sequential univariate optimization procedure: first find σ_R^2 and σ_C^2 using Equations 6 and 7 respectively, then find $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$, and σ_{RC}^2 using the solved σ_R^2 and σ_C^2 and Equations 5-7. Because $\Psi(z_1, z_2; r)$ is a monotonic function of r, the procedure is guaranteed to converge to a unique and globally optimal solution.

In summary, as depicted in Figure 1, our simulation model works as follows. The user specifies the parameters in the binary data domain (i.e., the probability of agreement and the five correlations defined in Table 1), which can be measured in a real study. These parameters are mapped to the latent-space parameters (fixed modality effect τ_i and the six variance component parameters, i.e., the sigmas). Then these computed parameters are used to generate continuous data X using Equation 1. Finally, the continuous data are dichotomized into binary data (Equation 2).

2.2 Use of simulation model for analysis method validation

Our simulation model is useful for validating analysis methods for MRMC studies with binary agreement data. In particular, one can validate a confidence interval estimator by following the procedure below:

1. Specify a set of parameters: expected performance for each modality $P_1 = P_2 = P$, the correlation parameters (ρ_c, ρ_r, ρ_τ, $\rho_{\tau c}$, and $\rho_{\tau r}$), the number of cases N_c, and the number of readers N_r.

2. Randomly generate a dataset with the parameters specified in step 1 using our simulation model.

3. Apply the analysis method in need of validation to the dataset. The analysis gives a confidence interval for the performance difference $P_1 - P_2$ (here $P_1 - P_2 = 0$ since $P_1 = P_2$).

4. Repeat steps 2 and 3 multiple times (e.g., 10,000) and estimate the coverage probability, i.e., the proportion of experiments for which the confidence interval for the performance difference covers the true value 0.

To demonstrate the robustness of an analysis method, a broad range of parameters in step 1 can be used. We say a method is conservative (or liberal) if the coverage probability of the confidence interval is greater (or smaller) than its targeted value (e.g., for a 95% confidence interval, we expect the coverage probability be 0.95).

Note that the procedure above for validating the coverage probability of confidence intervals does not require specifying the type of hypothesis testing. The situation $P_1 = P_2$ corresponds to the alternative hypothesis in non-inferiority testing and the procedure is just a validation of the coverage probability of confidence intervals under that situation. On the other hand, the situation $P_1 = P_2$ corresponds to the null hypothesis in superiority testing. Since one minus the coverage probability of performance difference is the type I error rate for superiority tests under the null hypothesis, the validation of the coverage probability in this situation is equivalent to validating the type I error rate in superiority tests (e.g., if an analysis method is conservative, the coverage probability of the 95% CI is greater than 95% and correspondingly, the type I error rate is less than 5%).

To illustrate the above procedure, we use it to validate the ORH method adapted to binary MRMC data analysis. The ORH method[13,18,21] uses a correlated analysis of variance (ANOVA) model for reader performance estimates, i.e., an ANOVA that models the observed reader performance as a linear combination of fixed modality effect, random reader effect, random "reader by modality interaction" effect, and a random error term. The error term is correlated rather than independent because of the correlations between (fixed) readers. As such, the method requires estimates of the variances of each reader's performance in both modalities and the covariances between every pair of readers within and across modalities, i.e., the fixed-reader covariance matrix for the vector of reader performance estimates. For binary outcomes, the fixed-reader covariance matrix can be estimated in a straightforward fashion, as described next.

Assuming two modalities and N_r readers, the binary data for each case can be collected in a $2N_r \times 1$ random vector $\mathbf{Z}$, where the first N_r values are for the first modality and the second N_r values are for the second modality. Conditioned on the reader, this vector has mean $\mathbf{P}_{c|r}$ and covariance matrix $\mathbf{\Sigma}_{c|r}$, where the subscript 'c|r' indicates that cases are random and readers are fixed. Given the binary outcome of a sample of N_c cases, denoted $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_{N_c}$, the vector of fixed-reader probabilities of agreement, $\mathbf{P}_{c|r}$, is estimated with the sample mean $\widehat{\mathbf{P}}_{c|r} = (1/N_c)\sum_{k=1}^{N_c} \mathbf{Z}_k$. Now, the fixed-reader covariance matrix of performance estimates required by the ORH method is the covariance matrix of $\widehat{\mathbf{P}}_{c|r}$, which is $\mathbf{\Sigma}_{c|r}/N_c$. Therefore, estimating $\mathbf{\Sigma}_{c|r}$ with the usual unbiased sample covariance matrix given by

$$\mathbf{S}_{c|r} = \frac{1}{N_c - 1}\sum_{k=1}^{N_c}(\mathbf{Z}_k - \widehat{\mathbf{P}}_{c|r})(\mathbf{Z}_k - \widehat{\mathbf{P}}_{c|r})^T, \tag{8}$$

an estimate of the required covariance matrix is $\mathbf{S}_{c|r}/N_c$. Note that this estimate for the fixed-reader covariance matrix of $\widehat{\mathbf{P}}_{c|r}$ is equivalent to the jackknife estimate, since the jackknife pseudovalues for a sample mean (i.e., $\widehat{\mathbf{P}}_{c|r}$) are the same as the original measurements ($\mathbf{Z}_k$). Once the fixed-reader covariance matrix is estimated, the ORH method can be applied to the analysis of a binary dataset.

To demonstrate the use of our simulation model for validating the ORH method, we consider a type of application where the purpose is to compare two imaging modalities in terms of probability of agreement in a fully-crossed design, i.e., all readers read images of all cases from both modalities once (no re-reading). To set the parameters in step 1, we follow a factorial experiment design, i.e., we consider all combinations of the following factors resulting in a total of $3 \times 3 \times 3 \times 8 = 216$ experimental conditions, as outlined below.

- Expected performance for each modality: $P_1 = P_2 = 0.75, 0.85, 0.95$;

- Number of readers: $N_r = 3, 6, 12$;

- Number of cases: $N_c = 50, 100, 200$;
- Correlation parameters in the binary score domain: we consider 8 sets as described below and shown in Table 2.

To specify the correlation parameters (see Table 1), we use three-tuples of L's and H's to denote low and high relative magnitudes of certain correlations. The first L or H represents the relative magnitude of the between-case correlations, i.e., the within-modality between-case correlation (ρ_c) or the between-modality between-case correlation ($\rho_{\tau c}$). The second L or H represents the relative magnitude of between-reader correlations, i.e., the within-modality between-reader correlation (ρ_r) or the between-modality between-reader correlation ($\rho_{\tau r}$), and the relative magnitude of between-modality correlation ρ_τ. The third L or H represents the relative magnitude of the difference between the within-modality correlation and the between-modality correlation ($\rho_c - \rho_{\tau c}$ or $\rho_r - \rho_{\tau r}$). The correlation difference denoted by the third letter is related to the reader-modality variance component or the case-modality variance component that directly contributes to the variance of $\hat{P}_1 - \hat{P}_2$. In specifying these correlation parameters, we calculated the binary correlations corresponding to the variance component parameters in Roe and Metz[9] and took that as a reference for the relative magnitude of these correlations. The "low" (L) or "high" (H) is only relative within the same correlation. For example, according to empirical experience and the Roe and Metz[9] parameters, the between-case correlation is usually much lower if compared with the between-reader correlation, but we still categorized the between-case correlation into (relatively) high and (relatively) low. With these considerations, we specified 8 sets of parameter values shown in Table 2.

To validate the ORH method, for each experimental condition, we repeat the simulation (steps 2 and 3) 10,000 times. We then estimate the coverage probability by calculating the proportion of experiments for which the 95% confidence interval covers the true value of performance difference, which is 0 since we set $P_1 = P_2$.

Table 2: List of correlation parameter values chosen for validating an analysis method

Structure*	ρ_c	$\rho_{\tau c}$	ρ_r	$\rho_{\tau r}$	ρ_τ
LLL	0.006	0.005	0.240	0.200	0.300
LLH	0.008	0.005	0.320	0.200	0.300
LHL	0.006	0.005	0.500	0.400	0.500
LHH	0.008	0.005	0.600	0.400	0.500
HLL	0.040	0.030	0.240	0.200	0.300
HLH	0.050	0.030	0.320	0.200	0.300
HHL	0.040	0.030	0.500	0.400	0.500
HHH	0.050	0.030	0.600	0.400	0.500

* L for low and H for high representing the relative magnitude of specific correlations. The first letter represents the between-case correlation. The second letter represents the between-reader correlation. The third letter represents the difference between within-modality correlation and the between-modality correlation.

2.3 Use of simulation model for sizing

Our simulation model is also a useful tool to aid in sizing a non-inferiority MRMC study when the two modalities are expected to perform the same. To size a study, the expected performance for each modality and the five correlation parameters are needed. These can be measured in a pilot study or adopted from prior similar studies. In addition, the non-inferiority margin must be specified when sizing a non-inferiority study. The non-inferiority margin (δ) is typically chosen on clinical grounds, and quantifies the maximum acceptable performance deficiency of the new modality with respect to the conventional modality.[22]

Once the parameters are fixed, the procedure below can be followed for sizing a study:

1. Set the null ($\mathcal{H}_0$) and alternative ($\mathcal{H}_1$) hypotheses and the following parameters: expected performance for each modality P , the correlation parameters (ρ_c, ρ_r, ρ_τ, $\rho_{\tau c}$, and $\rho_{\tau r}$), and the non-inferiority margin. For a non-inferiority study, the hypotheses are $\mathcal{H}_0 : P_1 - P_2 = -\delta$; $\mathcal{H}_1 : P_1 - P_2 > -\delta$. Initialize sample sizes N_r and N_c.

2. Repeat the following multiple (e.g., 10,000) times and calculate the proportion of experiments that the null hypothesis is rejected, which is the observed statistical power.

   (a) Randomly generate a dataset using our simulation model with the parameters specified in step 1;

(b) Analyze the dataset with a validated analysis method and determine if the null hypothesis is rejected based on the analysis;

3. Compare the observed statistical power obtained in 2 to a desired power (e.g., 80%) and adjust N_r and/or N_c accordingly.

4. Repeat steps 2 and 3 until the desired power is achieved.

This method is iterative and potentially time-consuming, but it can be applied to any analysis method, which is particularly useful when analytical sizing formulas are not available.

We demonstrate our method with a set of hypothetical parameters that illustrate the factors that influence the sample size.. The parameters were chosen according to our empirical experience with practical applications. Namely, we calculate the number of cases N_c needed to achieve a 80% statistical power in a non-inferiority study for all 16 combinations of the following parameters:

- Expected performance for each modality:$P_1 = P_2 = 0.80, 0.90$;
- Number of readers: $N_r = 6, 12$;
- Non-inferiority margin: $\delta = 0.03, 0.05$;
- Correlation structure: LHH and LHL in Table 2.

3 Results

Figure 2 presents the simulation results for the validation of the ORH method in terms of the coverage probability of 95% confidence intervals under various of experimental conditions. Equivalently, the results can be viewed as an investigation of the type I error rate of a superiority hypothesis test at the 0.05 significance level, where the null hypothesis is that $P_1 = P_2$. The results show that the coverage probability of the 95% confidence intervals is around 95% in most situations, indicating the reliability and robustness of the ORH

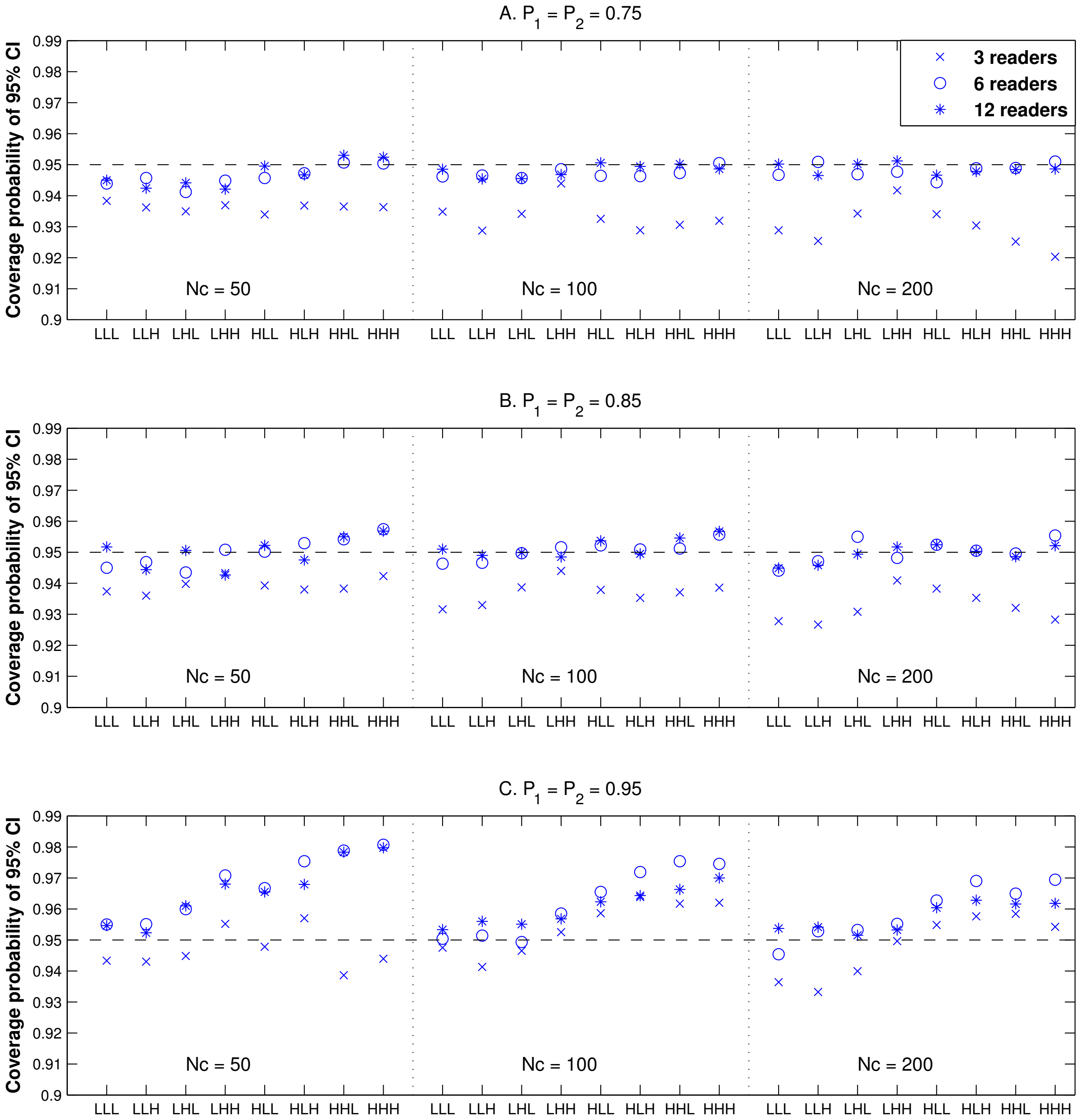

Figure 2: simulation results for the validation of the ORH method in terms of the coverage probability of 95% confidence intervals (CI) under various of experimental conditions. The standard error of the estimated coverage probability is $\sqrt{0.95 \times 0.05/10000} = 0.002$ (error bar not shown for clarity). (L for low and H for high relative magnitude of certain correlations. The first letter represents the between-case correlation. The second letter represents the between-reader correlation. The third letter represents the difference between within-modality correlation and the between-modality correlation.)

Table 3: Sizing results in planning a hypothetical non-inferiority study: the number of cases needed to achieve a statistical power of 0.80.

		$P_1 = P_2 = 0.80$		$P_1 = P_2 = 0.90$	
		$N_r = 6$	$N_r = 12$	$N_r = 6$	$N_r = 12$
$\delta = 0.03$	LHH	>10,000	3,200	>10,000	610
	LHL	1483	493	444	244
$\delta = 0.05$	LHH	889	305	216	150
	LHL	240	150	115	81

δ: non-inferiority margin. LHH, LHL: see Table 2. P_i: expected probability of agreement for modality i. N_r: the number of readers.

method in analyzing binary MRMC data. However, there are two noticeable exceptions. The first is that, when the number of readers is 3, the coverage probability is generally lower than 95%, indicating that the ORH method is liberal when the number of readers is too small. This result is not surprising because we would not expect a reliable estimate of reader variability with so few readers. The second exception is that when $P_1 = P_2 = 0.95$, the coverage probability tends to be higher than expected, i.e., the ORH method is conservative. Conservative coverage for high binomial proportions is similarly found in other methods (not in the MRMC setting).[31] One possible explanation for this observation is that the ORH method does not properly account for the fact that the performance P has an upper limit of 1 (i.e., the distribution of the performance is truncated at 1). It is also interesting to observe a trend that this conservativeness is mitigated when sample sizes increase. Finally, the observations regarding the binary data analysis results reported here appear to be consistent with those observed in the analysis of ROC data.[9,11]

Table 3 presents the sizing results in planning a hypothetical non-inferiority study using our simulation model. The table shows the number of cases needed to achieve a statistical power of 0.80 for various combinations of performance level, non-inferiority margin, correlation structure, and the number of readers. The results show a substantial trade-off between the number of cases and the number of readers needed to achieve the same power. Moreover, the results also demonstrate how the sample sizes are influenced by the performance level, the non-inferiority margin, and the correlation structure.

4 Discussions and Conclusion

Our simulation results indicate that the ORH method adapted to binary agreement data has a reasonably accurate type I error rate (or equivalently, coverage probability of confidence intervals) for six or more readers and 50 or more cases, for agreement probabilities of 0.85 or less. Accuracy extends up to agreement probabilities of 0.95 when the between-case correlation (or reader variability) is relatively low. When the average reader performance is high and the reader variability is relatively high, the ORH method is shown to be conservative. Furthermore, our sizing exercise using hypothetical parameters indicates that the number of cases may vary greatly depending on other experimental factors. Among these factors, the non-inferiority margin should be determined independently on clinical grounds prior to the study. In addition, the correlations (or variance components) and the average performance values should be adopted from prior similar studies or a pilot study. The sizing method presented here allows study designers make a practical and cost-effective trade-off between the number of patient cases and the number of clinical readers.

Our methods for simulation, sizing, and analysis account for two sources of variability: the case sample and the reader sample. We implicitly assume that the truth states of the patient cases are known with certainty. In practice, however, the truth may be imperfect to some extent, especially when the truth is obtained from a panel of readers. This additional source of uncertainty may be ignored only if the truth is highly reliable; otherwise, some accounting of the variability in the truth should be accomplished. Interested readers are referred to Wagner et al.,[4] who suggested using resampling of truthing panel readers to account for truthing uncertainty in the analysis.

There are several directions that would extend the current work. First, while the procedures illustrated in this paper for simulation, validation of analysis, and sizing binary MRMC studies are conceptually general, our simulation is restricted to the situation that the performance of two modalities is equal. The restriction arises from the difficulty in mapping the parameters between the binary data domain and the latent continuous data domain for non-equal performance under our model. We plan to extend our simulation model presented here to a more general model to address this problem in a future publication to allow the performance to differ across modalities. This will allow us to validate a non-inferiority hypothesis test (simulate $P_2 \neq P_1$) and size a superiority hypothesis test (simulate $P_2 > P_1$).

Second, more analysis methods can be investigated and compared utilizing our simulation model. We have investigated the ORH method for the analysis of binary MRMC data in the non-inferiority setting. As mentioned earlier, the U-statistics method of Gallas for MRMC ROC data analysis[16] has been adapted to deal with binary data.[29] In addition, the generalized linear mixed model (GLMM)[32] approach may be suitable for this type of study and may be a natural choice because of its association with binary data analysis. Our simulation results in this study have shown that the ORH method is overly conservative when the agreement rate is close to 1 and the reader variability is relatively high. The GLMM approach may have advantages in these situations. However, implementing this class of model in the MRMC framework is not a trivial extension and thus requires additional statistical modeling and development. It is interesting future work to investigate and compare these methods.

Finally, while most MRMC methods are developed for a fully-crossed study design, the ORH and the Gallas methods for ROC data have been extended to deal with alternative study designs such as the split-plot design.[33] In the split-plot design, a sub-group of readers read their own sub-group of cases, which is useful because sometimes it is inconvenient or impractical to have all the readers read all the cases. We have assumed a fully-crossed design in this paper, but extension to alternative study designs is an important topic for future work.

In conclusion, we have presented a simulation model mimicking MRMC reader studies with binary agreement data and the expected performance of two modalities is equal. We developed a mathematical model that can simulate MRMC data with a correlation structure that incorporates two sources of variation observed in real studies: variability due to random readers and variability due to random cases. We outlined procedures for using our simulation model to validate an analysis method and size a binary MRMC study to achieve a desired statistical power in the non-inferiority setting where the performance of two imaging modalities is expected to be the same. In particular, adapting the ORH method to the analysis of MRMC binary outcome data, we used our simulation model to investigate the confidence interval coverage probability under various experimental conditions and also demonstrated its use for non-inferiority study sizing. The methodological framework we present here and our software package (freely available from the Google code project hosting site[34]) are useful for simulation, analysis validation, and sizing of non-inferiority MRMC reader studies with binary outcomes.

Acknowledgments

The authors thank Dr. Craig Abbey for his critical review of the manuscript and for his constructive suggestions that helped improve the clarity of the paper.

References

1 G. Gennaro, A. Toledano, C. di Maggio, E. Baldan, E. Bezzon, M. L. Grassa, L. Pescarini, I. Polico, A. Proietti, A. Toffoli, and P. C. Muzzio, “Digital breast tomosynthesis versus digital mammography: a clinical performance study,” *European Radiology* **20**(7), 1545–53 (2010).

2 Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, “Improving breast cancer diagnosis with computer-aided diagnosis,” *Academic Radiology* **6**(1), 22–33 (1999).

3 R. F. Wagner, S. V. Beiden, G. Campbell, C. E. Metz, and W. M. Sacks, “Assessment of medical imaging and computer-assist systems: Lessons from recent experience,” *Acad Radiol* **9**(11), 1264–1277 (2002).

4 R. F. Wagner, C. E. Metz, and G. Campbell, “Assessment of medical imaging systems and computer aids: A tutorial review,” *Acad Radiol* **14**(6), 723–748 (2007).

5 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).

6 J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal DetectionTheory*, Academic Press, New York (1982).

7 C. E. Metz, "Basic principles of ROC analysis," *Semin Nucl Med* **8**(4), 283–298 (1978).

8 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," *Investigative Radiology* **27**(9), 723–731 (1992).

9 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**, 298–303 (1997).

10 C. A. Roe and C. E. Metz, "Variance-component modeling in the analysis of receiver operatingcharacteristic (ROC) index estimates," *Acad Radiol* **4**, 587–600 (1997).

11 S. L. Hillis, "Simulation of unequal-variance binormal multireader ROC decision data: An extension of the roe and metz simulation model.," *Acad Radiol* **19**, 1518–1528 (2012).

12 B. D. Gallas and S. L. Hillis, "Simulating ROC experiments: the Roe and Metz model generalized to allow effects to vary across modalities and truth states." submitted to the same issue of Journal of Medical Imaging.

13 A. N. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations," *Communications in Statistics-Simulation and Computation* **24**(2), 285–308 (1995).

14 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).

15 X. Song and X.-H. Zhou, “A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data,” *Biostatistics* **6**, 303–312 (2005).

16 B. D. Gallas, “One-shot estimate of MRMC variance: AUC,” *Acad Radiol* **13**(3), 353–362 (2006).

17 B. D. Gallas, A. Bandos, F. Samuelson, and R. F. Wagner, “A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators,” *Commun Stat A-Theory* **38**(15), 2586–2603 (2009).

18 S. L. Hillis, “A comparison of denominator degrees of freedom methods for multiple observer ROC analysis,” *Stat Med* **26**(3), 596–619 (2007).

19 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**, 647–661 (2008).

20 S. L. Hillis, N. A. Obuchowski, and K. S. Berbaum, “Power estimation for multireader ROC methods: an updated and unified approach,” *Acad Radiol* **18**, 129–142 (2011).

21 S. L. Hillis, “A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data,” *Stat Med* **33**, 330–360 (2014).

22 W. Chen, N. A. Petrick, and B. Sahiner, “Hypothesis testing in noninferiority and equivalence MRMC ROC studies.,” *Acad Radiol* **19**, 1158–1165 (2012).

23 B. D. Gallas, H.-P. Chan, C. J. D’Orsi, L. E. Dodd, M. L. Giger, D. Gur, E. A. Krupinski, C. E. Metz, K. J. Myers, N. A. Obuchowski, B. Sahiner, A. Y. Toledano, and M. L. Zuley, “Evaluating imaging and computer-aided detection and diagnosis devices at the FDA,” *Acad Radiol* **19**, 463–477 (2012).

24 A. A. Renshaw, N. Cartagena, S. R. Granter, and E. W. Gould, “Agreement and error rates using blinded review to evaluate surgical pathology of biopsy material.,” *Am J Clin Pathol* **119**, 797–800 (2003).

25 J. R. Gilbertson, J. Ho, L. Anthony, D. M. Jukic, Y. Yagi, and A. V. Parwani, “Primary histologic

diagnosis using automated whole slide imaging: a validation study," *BMC clinical pathology* **6**(1), 4 (2006).

26 N. Velez, D. Jukic, and J. Ho, "Evaluation of 2 whole-slide imaging applications in dermatopathology," *Human pathology* **39**(9), 1341–1349 (2008).

27 T. W. Bauer, L. Schoenfield, R. J. Slaw, L. Yerian, Z. Sun, and W. H. Henricks, "Validation of whole slide imaging for primary diagnosis in surgical pathology," *Archives of pathology & laboratory medicine* **137**(4), 518–524 (2013).

28 S. Krishnamurthy, K. Mathews, S. McClure, M. Murray, M. Gilcrease, C. Albarracin, J. Spinosa, B. Chang, J. Ho, J. Holt, A. Cohen, M. Dilip Giri, K. Garg, R. L. B. Jr, and K. Liang, "Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections," *Arch Pathol Lab Med.* **137**, 1733–1739 (2013).

29 B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," *J Opt Soc Am A, Special Issue on Image Quality* **24**(12), B70–B80 (2007). Special Issue on Image Quality. Also selected for inclusion in Virtual Journal of Boimedical Optics from all others published in all OSA journals.

30 L. J. Emrich and M. R. Piedmonte, "A method for generating high-dimensional multivariate binary variates," *Statistical Computing* **45**, 302–304 (1991).

31 R. G. Newcombe, "Interval estimation for the difference between independent proportions: comparison of eleven methods," *Statistics in medicine* **17**(8), 873–890 (1998).

32 C. E. McCulloch, J. M. Neuhaus, and S. R. Searle, *Generalized, Linear, and Mixed Models*, John Wiley & Sons (2011).

33 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**, 1508–1517 (2012). Invited paper for Special Metz Memorial Issue I.

34 W. Chen and A. Wunderlich, "iMRMC Software for Binary Agreement Data." `http://code.google.com/p/imrmc/wiki/iMRMC_Binary`. Accessed: 2014-10-24.

List of Figures

List of Tables