One-Shot Estimate of MRMC Variance: AUC[1]

Brandon D. Gallas

Rationale and Objectives. One popular study design for estimating the area under the receiver operating characteristic curve (AUC) is the one in which a set of readers reads a set of cases: a fully crossed design in which every reader reads every case. The variability of the subsequent reader-averaged AUC has two sources: the multiple readers and the multiple cases (MRMC). In this article, we present a nonparametric estimate for the variance of the reader-averaged AUC that is unbiased and does not use resampling tools.

Materials and Methods. The one-shot estimate is based on the MRMC variance derived by the mechanistic approach of Barrett et al. (2005), as well as the nonparametric variance of a single-reader AUC derived in the literature on U statistics. We investigate the bias and variance properties of the one-shot estimate through a set of Monte Carlo simulations with simulated model observers and images. The different simulation configurations vary numbers of readers and cases, amounts of image noise and internal noise, as well as how the readers are constructed. We compare the one-shot estimate to a method that uses the jackknife resampling technique with an analysis of variance model at its foundation (Dorfman et al. 1992). The name one-shot highlights that resampling is not used.

Results. The one-shot and jackknife estimators behave similarly, with the one-shot being marginally more efficient when the number of cases is small.

Conclusions. We have derived a one-shot estimate of the MRMC variance of AUC that is based on a probabilistic foundation with limited assumptions, is unbiased, and compares favorably to an established estimate.

Key Words. MRMC; ROC; AUC; variance; reader variability; case variability; jackknife; ANOVA; bootstrap.

In the field of medical imaging, investigators, manufacturers, regulators, and insurance providers are all faced with assessing imaging system performance. One dimension of this assessment asks how effective are clinicians at diagnosing disease from an image. If a clinician can diagnose a disease better with system A compared with system B, then we may say that system A is better than system B for that task. What has become clear over the years is that human variability—how clinicians interpret images, how good they are at the interpretation, and how reproducible their decisions are—significantly affects the precision of the measures of diagnostic performance. For example, the ability of radiologists to detect cancer in mammograms varies by as much as 20% (1). Other examples of high reader variability are depicted pictorially in Obuchowski et al. (2). The reader variability can hide clinically significant differences in diagnostic performance.

The area under the receiver operating characteristic curve (ROC) curve, denoted AUC, is a commonly used measure of diagnostic performance (3–5). For a single reader, the variance of the nonparametric estimate of AUC has been characterized in great detail under the names the Wilcoxon (6) and the Mann-Whitney U statistics (7–10). Several estimates for the variance of this single-reader AUC estimate exist either assuming distributional forms for the ROC scores (parametric) (11–14) or

Acad Radiol 2006; 13:353–362

[1] From the NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, US FDA/CDRH, Bldg 1, HFZ-140, 12720 Twinbrook Parkway (Rm 158), Rockville MD 20852-1720. Received September 21, 2005; accepted November 7, 2005. **Address correspondence to:** B.D.G. e-mail: brandon.gallas@fda.hhs.gov

doi:10.1016/j.acra.2005.11.030

leaving the distributions undefined (nonparametric) (10,13,15). Additionally, all the single-reader variance estimates can be applied to a comparison of AUCs from two modalities that are correlated by the study design (the same set of patients is used in both modalities) (14–17).

For a single reader, the only random effect in the AUC estimates is from the cases; the reader is considered fixed. However, study designs used in practice involve multiple readers, and investigators started asking what would happen if different readers had participated. The answer was that the results would be different. Thus investigators designed models to include and account for reader variability, and multireader multicase, or MRMC, analysis was born. The first MRMC model was introduced by Swets and Pickett (18), who partitioned AUC into three components: case-sample variation, between-reader variation, and within-reader variation. This work led to an MRMC model by Dorfman, Berbaum, and Metz (19–22) (DBM) that is based on a random effects, linear decomposition of AUC. Namely, for a single modality and no replication, the AUC of the j^{th} given the k^{th} case set is written as

$$A_{jk} = \mu + r_j + c_k + (rc)_{jk} + Z_{jk}, \tag{1}$$

where μ is the AUC averaged over all readers and case sets, r_j is a random reader effect, c_k accounts for a random case effect, $(rc)_{jk}$ quantifies possible reader-case interaction effects, and Z_{jk} models internal noise; all of these are assumed to be independent zero-mean Gaussian random variables. The variances of the random variables, or the components of variance of AUC are estimated by an analysis of variance (ANOVA) model, which at its heart employs the jackknife resampling scheme. Obuchowski and Rockette (2,23–25) employ a similar ANOVA model that gives estimates similar to the jackknife estimates of DBM (26). Other MRMC models decompose AUC on different scales via a link function and solve for the means and variances of the random effects in different ways (27,28).

There are two more MRMC variance estimates of interest in this introduction. The first is from Beiden, Wagner, and Campbell (29) (BWC). Using the DBM model in Eq 1, they build a system of equations that relates certain measurable variances to the components of variance in Eq 1. They estimate the measurable variances with a bootstrap resampling method. The last MRMC estimate that we wish to note here is that by Lee and Rosner (30,31). Unlike the other investigators, their statistic of interest is not the average of several reader AUCs. Instead, it is the AUC you get when you combine all the scores for all the readers to build a single ROC curve. This AUC has been referred to as the "generic reader" AUC (32), and it may be useful when data are extremely scarce. However, the interpretation of the resulting generic reader ROC curve is not obvious and may be severely distorted because readers often use different scales.

Recently, Barrett, Clarkson, and Kupinski (33,34) (BCK) have derived an expression for the variance of the reader-averaged AUC from a probabilistic foundation. This foundation begins with the distributions of the readers and cases and the mechanism for how readers generate scores.

In this article, we present an estimate of the variance expression derived by BCK (33,34). The estimate is completely nonparametric and unbiased. The estimate is given explicitly and does not require any complicated programming, such as iterative procedures or statistical resampling. We call this estimate the one-shot estimate of the MRMC variance of the reader-averaged AUC, and we investigate the bias and variance properties of this estimate through a set of Monte Carlo (MC) simulations.

MATERIALS AND METHODS

One-shot Estimate

Consider an ROC experiment in which R readers interpret and score a set of images G, every reader reading every image. We refer to the set of readers with Γ and the r^{th} reader with γ_r. The set of images G includes N_0 that are signal-absent, designated g_{0i}, and G_1 that are signal-present, designated g_{1j}. The scores from reader γ_r are organized into an $N_0 + N_1$ dimensional array $\mathbf{t}_{Gr}$ with elements t_{0ir} denoting the score for the i^{th} signal-absent image, and t_{1jr} the score for the j^{th} signal-present image. Arranging the scores from all the readers into a matrix $\mathbf{T}_{G\Gamma}$, the reader-averaged AUC is estimated by

$$A(\mathbf{T}_{G\Gamma}) = \frac{1}{R}\sum_{r=1}^{R} a(\mathbf{t}_{Gr}), \tag{2}$$

where $a(\mathbf{t}_{Gr})$ is the r^{th} reader's AUC given by

$$a(\mathbf{t}_{Gr}) = \frac{1}{N_0 N_1}\sum_{i=0}^{N_0}\sum_{j=0}^{N_1} s(t_{1j} - t_{0i}). \tag{3}$$

The step function $s(\cdot)$ in the Eq 3 is 1.0 when the argument is positive, 0.0 when the argument is negative, and 0.5 when the argument is zero (allowing ties).

To proceed, we make some basic assumptions and definitions. Note that "$\sim$" is shorthand for "is distributed as."

1. Images: all the images are independent and identically distributed (IID) from the nondiseased or diseased distribution: $g_{0i} \sim p_0(\mathbf{g}_0)$, $g_{1j} \sim p_1(\mathbf{g}_1)$, and

$$G \sim p_G(G) = \prod_{i=1}^{N_0} p_0(\mathbf{g}_{0i}) \prod_{j=1}^{N_1} p_1(\mathbf{g}_{1j}). \quad (4)$$

2. Readers: all the readers are IID: $\gamma_r \sim p_\gamma(\gamma)$, and

$$p_\Gamma(\Gamma) = \prod_{r=1}^{R} p_r(\gamma_r). \quad (5)$$

3. Scores: when conditioned on the generating reader and image, scores are IID: the score t for image $\mathbf{g}$ and reader γ is distributed as $p_{t|\mathbf{g},\gamma}(t|\mathbf{g},\gamma)$, and

$$\mathbf{t}_{G\gamma} \sim p_{\mathbf{t}|G,\gamma}(\mathbf{t}|G,\gamma)$$

$$= \prod_{i=1}^{N_0} p_{t|\mathbf{g},\gamma}(t_{0i\gamma}|\mathbf{g}_{0i},\gamma) \prod_{j=1}^{N_1} p_{t|\mathbf{g},\gamma}(t_{1j\gamma}|\mathbf{g}_{1j},\gamma). \quad (6)$$

An example distribution of readers is the set of all US radiologists. Although there is no mathematical formula describing this distribution, we can still sample from it. Another distribution of readers is the one that we use in the simulation. It is a set of linear observers: mathematical observers that score an image with a linear combination of the image pixel values. We can describe this distribution mathematically and sample from it with a random number generator.

Using the probabilistic foundation introduced by BCK, we derive an estimate of the one-shot estimate of the variance of $A(\mathbf{T}_{G\Gamma})$ as a linear combination of sums of step functions (Appendix). The one-shot estimate is

$$\hat{V} = \frac{1}{R}\left(c_1\hat{M}_1 + c_2\hat{M}_2 + c_3\hat{M}_3 + c_4\hat{M}_4\right) + \frac{R-1}{R}\left(c_1\hat{M}_5 + c_2\hat{M}_6 + c_3\hat{M}_7 + c_4\hat{M}_8\right) - \hat{M}_8 \quad (7)$$

where

$$c_1 = 1/N_0N_1, \qquad c_2 = (N_0-1)/N_0N_1,$$
$$c_3 = (N_1-1)/N_0N_1, \quad c_4 = (N_0-1)(N_1-1)/N_0N_1, \quad (8)$$

and the moments $\hat{M}_1 - \hat{M}_8$ are found in Eq A.16–A.23. The estimate $\hat{V}$ can be shown to be the unique, uniformly minimum variance, unbiased estimate of var($A(\mathbf{T}_{G\Gamma})$) (personal communication with Waleed Yousef, who derived the one-shot estimate from U-statistics).

The Jackknife Approach

The jackknife approach of DBM (19) is a method for applying ANOVA model to MRMC ROC data. It constructs a reader-by-case matrix of AUC pseudo-values with the jackknife method. The data are assumed normal, fitting a factorial design with completely crossed factors and one observation per cell. Under the assumptions, the observed mean squares are unbiased estimates of linear combinations of the components of variance related to Eq 1.

We use the single-reader nonparametric AUC estimate (Eq 3) to generate a row of ($N_0 + N_1$) jackknife pseudo-values (one per image) for each observer. The resulting $R \times (N_0 + N_1)$ matrix is denoted $\mathbf{A}$, with elements A_{rk}. The observed mean squares and what they estimate are as follows:

$$\langle M\hat{S}(r)\rangle = \left\langle (N_0+N_1)\sum_{r=1}^{R}\frac{(\bar{A}_{r\cdot} - \bar{A}_{\cdot\cdot})^2}{(R-1)}\right\rangle$$

$$= (N_0+N_1)\,\sigma_r^2 + \sigma_{rc*}^2 \quad (9)$$

$$\langle M\hat{S}(c)\rangle = \left\langle R\sum_{k=1}^{N_0+N_1}\frac{(\bar{A}_{\cdot k} - \bar{A}_{\cdot\cdot})^2}{(N_0+N_1-1)}\right\rangle = R\sigma_{c*}^2 + \sigma_{rc*}^2 \quad (10)$$

$$\langle M\hat{S}(rc)\rangle = \left\langle \sum_{k=1}^{N_0+N_1}\sum_{r=1}^{R}\frac{(\bar{A}_{rk} - \bar{A}_{r\cdot} - \bar{A}_{\cdot k} + \bar{A}_{\cdot\cdot})^2}{(R-1)(N_0+N_1-1)}\right\rangle = \sigma_{rc*}^2 \quad (11)$$

where $\bar{A}_{r\cdot}$ is the mean of the r^{th} row, $\bar{A}_{\cdot k}$ is the mean of the k^{th} column, and $\bar{A}_{\cdot\cdot}$ is the grand mean, or the mean over all the elements in $\mathbf{A}$. In the expression above, $\sigma_{c^*}^2$ and $\sigma_{rc^*}^2$ are normalized with respect to a single case. Thus they differ from σ_c^2 and σ_{rc}^2 by a factor of (N_0 +

N_1). Consequently, we estimate the variance of $A(\mathbf{T}_{G\Gamma})$ with

$$\hat{V}_J = R^{-1}(N_0 + N_1)^{-1}\left(M\hat{S}(r) + M\hat{S}(c) - M\hat{S}(rc)\right). \tag{12}$$

Simulation Experiments

In one MC iteration, we generate a set of images G and a set of readers Γ. The readers score the images $\mathbf{T}_{G\Gamma}$ and produce a reader-averaged AUC, $A_m = A(\mathbf{T}_{G\Gamma})$. The array of scores $\mathbf{T}_{G\Gamma}$ is then sent to two routines. The first routine estimates the total variance of $A(\mathbf{T}_{G\Gamma})$ with $\hat{V}$ given in Eq 7, and the second uses the jackknife approach of DBM ($\hat{V}_J$ given in Eq 12). After N_{mc} = 10,000 MC iterations, we investigate the bias and variance of $\hat{V}$ and $\hat{V}_J$. In what follows, we describe the default MC iteration: how images are generated, how readers are generated, and how the readers score images. Then we describe three experiments in which we vary the default experiment.

Default MC iteration—We start the iteration by generating N_0 signal-absent test images $\mathbf{g}_{0i}$ that are IID, $N_x \times N_y$, Gaussian random vectors with zero mean and covariance matrix equal to the identity matrix times $\sigma_{\mathbf{g}}^2$. Similarly, we generate N_1 signal-present test images $\mathbf{g}_{1j}$, which are the same as the signal-absent images except that the mean image of the distribution contains a bright spot $\mathbf{s}$, the signal to be detected. The bright spot is a sampled Gaussian function located in the center of the image with a spread corresponding to a sigma of 2.5 pixels. The default parameters for the images are as follows: N_0 = 50, N_1 = 50, $N_x = N_y$ = 16, $\sigma_{\mathbf{g}}^2$ = 0.5. The ideal observer for these images is the matched filter: the linear observer with its template equal to $\mathbf{s}$. The height of the signal yields an ideal observer SNR = 1.60, or AUC = 0.87.

The next stage of the MC iteration determines N_r IID linear model observers: the readers. Each model observer requires N_0^* signal-absent training images $\mathbf{g}_{0i}^*$ and N_1^* signal-present training images $\mathbf{g}_{1j}^*$, where the training images are independently sampled from the same distribution as the test images. Specifically, a template for each observer is generated as the mean difference between the signal-present and signal-absent training images, or

$$\mathbf{w}_r = \frac{1}{N_1^*}\sum_{r=1}^{N_1^*}\mathbf{g}_{1i}^* - \frac{1}{N_0^*}\sum_{r=1}^{N_0^*}\mathbf{g}_{0i}^*. \tag{13}$$

Given the i^{th} signal-absent test image and the j^{th} signal-present test image, the scores for the r^{th} reader are thus

$$t_{0ir} = \mathbf{w}_r^t\mathbf{g}_{0i}, \quad t_{1jr} = \mathbf{w}_r^t\mathbf{g}_{1j}. \tag{14}$$

The default parameters for the readers are as follows: N_r = 10, N_0^* = 50, N_1^* = 50.

The reader-averaged AUC for the default setting yields an AUC = 0.86. This is slightly biased from the ideal observer because only 50 pairs of images are used to estimate the observers.

Simulation experiment 1.—In this experiment, we consider three image noise levels, $\sigma_{\mathbf{g}}^2$ = [0.25,0.50,1.00], and add internal noise. The internal noise is realized by adding an IID Gaussian random variable (mean zero, σ_ε^2 = [0.0,1.0,10.0]) to each score. These 3 × 3 parameter values are chosen to get a broad coverage in the reader-averaged AUC estimates (0.60–0.94).

Simulation experiment 2.—In this experiment, we vary the number of signal-absent training images N_0^* = [10,25,50], and the number of signal-absent testing images N_0 = [10,25,50]. The reader template $\mathbf{w}_r$ is equally affected by varying the number of signal-present training images, as by varying the number of signal-absent training images. Likewise, the impact on variance is symmetric for our simulation model with regard to the number of signal-present and signal-absent testing images. Therefore, we do not present results varying the number of signal-present training or testing images. In terms of the reader-averaged AUC estimates, the number of training images changes them from 0.78 to 0.86, whereas the number of testing images has no impact.

Simulation experiment 3.—In this experiment, we vary the number of readers N_r = [5,10,20] and the average number of masked pixels in the reader templates. We introduce masking pixels here to add variability and correlation structure across readers. If two readers have the same pixels masked, they are missing the same information, and their scores will be correlated. We mask pixels independently for each reader by randomly selecting elements of a reader's template and setting them to zero, effectively throwing the information in those pixels away. The selection process simply multiplies the pixels of the reader template by zero with probability μ_{mask} = [0.0,0.5,0.9]. Figure 1 shows a reader that has 90% of its pixels masked. This is the template used to generate scores as in Eq 14.

The selected range of masking probabilities yields reader-averaged AUC estimates between 0.86 and 0.63 with the number of readers having no impact on average AUC, but instead influencing the variance of AUC.

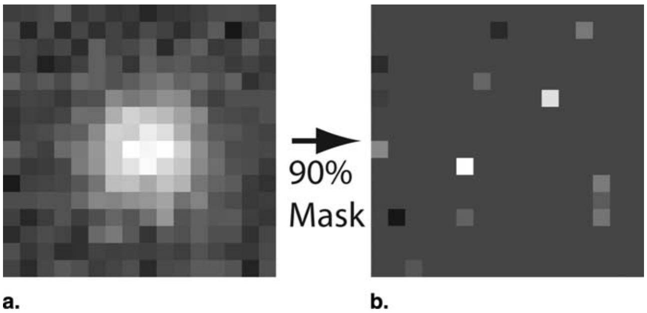

Figure 1. This figure shows a reader template before and after masking 90% of the pixels.

RESULTS

In what follows, we assess the bias and variance of the one-shot estimate of the variance of the reader-averaged AUC, $A(\mathbf{T}_{G\Gamma})$. The gold standard in this comparison is the empirical variance calculated from 10,000 MC trials per simulation configuration; we refer to these as if they are the true population, or ensemble, variances $\mathrm{var}(A(\mathbf{T}_{G\Gamma}))$. Likewise, we let the MC mean and variance of the different variance estimates, $\hat{V}$ and $\hat{V}_J$, represent the true population mean and variance of these estimates. The true mean and variance of the estimates and $A(\mathbf{T}_{G\Gamma})$ can then be used to assess bias, efficiency, and root mean-square error (RMSE). Note that there were 27 simulation configurations: 9 for each experiment.

Bias of the Estimate

We first look at the bias of the one-shot estimate:

$$\mathrm{bias} = \langle \hat{V} - \mathrm{var}(\mathbf{T}_{G\Gamma}) \rangle. \quad (15)$$

In Fig 2A, we plot biases normalized by the corresponding true variances for each simulation configuration and estimator. The dotted line indicates a bias that is 3% of the true variance being estimated. This plot shows that the bias of the one-shot estimate is less than 1% of the true variance. The bias of the jackknife is less than 3% for most of the values but grows to 10% for simulation configurations with very few (10) testers.

Variance of the Estimate

Figure 2B shows the variances of the jackknife divided by the variances of the one-shot estimates for each simulation configuration. These ratios of variances quantify the efficiency of the jackknife relative to the one-shot. The trend in the plot seems to indicate that the one-shot estimate is more efficient than the jackknife estimate. However, the increased efficiency is less than 6% for all the points except those where only a few images are used to test the readers.

RMSE of the Estimate

The final assessment of the one-shot estimate is in terms of root mean-square error, or

$$\mathrm{RMSE} = \sqrt{\mathrm{bias}^2 + \mathrm{variance}}. \quad (16)$$

This measure is the most relevant for the practical user, because it combines the bias and variance into a single number that is in the units of the quantity being estimated. It quantifies the total error that might be encountered in a simulation. Figure 2C plots this summary metric divided by the true variances being estimated for the three estimators and all configurations. Thus the dotted line indicates a RMSE that is 40% of the true variance being estimated. For all but a few configurations, mainly those involving only a few testers, the one-shot and jackknife estimators have the same RMSE, and the RMSEs lie between 10% and 40% of the true variance being estimated.

DISCUSSION

In this article, we present a one-shot estimate of the MRMC variance of the reader-averaged AUC (Eq 7). This reader-averaged AUC is from the fully crossed design in which each reader reads each case and subsequently contains correlations across the readers and cases. In addition to quantifying the uncertainty of a completed experiment, the one-shot estimate can be used to scale a future pivotal study. To scale a pivotal study, one estimates the moments of the step function $s(\cdot)$ as given in Eq A.16–A.23, and investigates the impact of adjusting the number of readers and cases (R,N_0,N_1) in computing $\hat{V}$ (Eq 7).

The derivation of the one-shot estimate uses the probabilistic foundation introduced by BCK (33,34) to write the variance as a linear combination of moments of the step function $s(\cdot)$ (Appendix). This linear combination differs from the BCK variance expression only in notation; they summarize the total variance in terms of linear combinations of conditional variances (ie, central second moments); we use the raw moments.

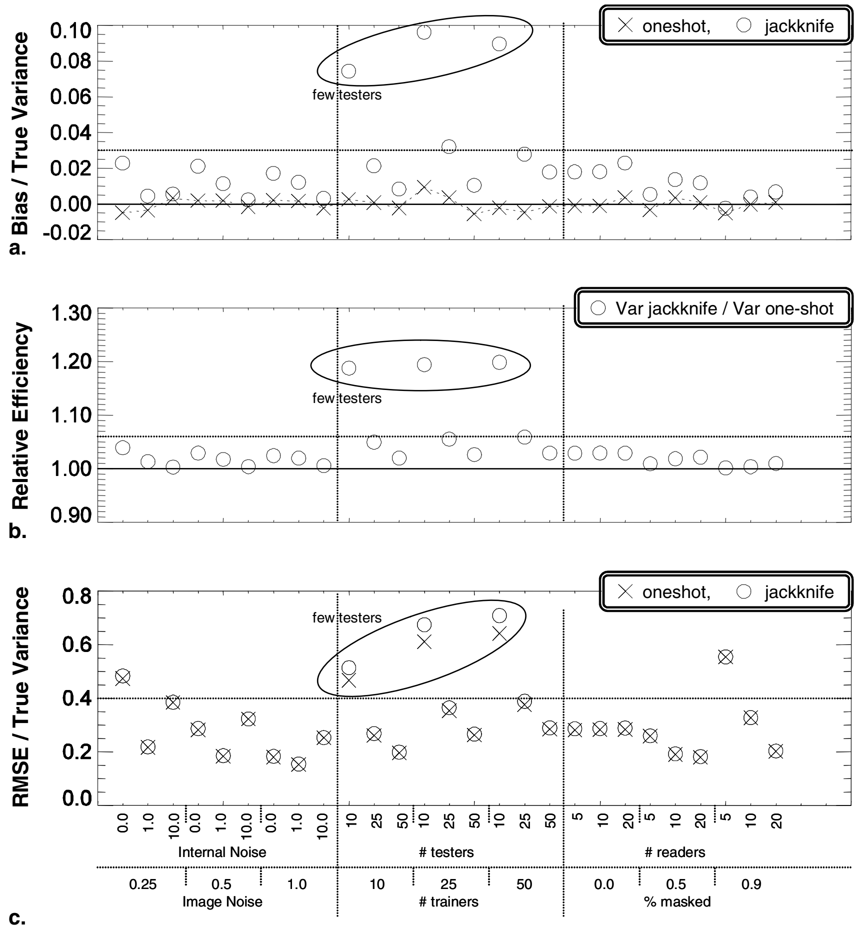

Figure 2. For each simulation configuration we assess the one-shot (diamonds) and jackknife (circles) estimates. **(a)** Biases normalized by the true variances. **(b)** Variances of the jackknife estimates normalized by the variances of the one-shot estimates. **(c)** root mean-square errors normalized by the true variances.

We call the estimate introduced here the one-shot estimate to differentiate it from the estimates that employ the jackknife and bootstrap resampling methods. Additionally, the one-shot estimate is also completely nonparametric. In contrast, the jackknife method and some bootstrap methods have an additive model at their foundation.

The one-shot estimate is unbiased; this trait is demonstrated in Fig 2A. In the same plot, we see that the jackknife is perhaps slightly biased, although the bias is not very large except when very few testers (10) are included in the experiment.

In terms of variance, the one-shot estimate is more efficient than the jackknife by a small amount; the ratio of variances (jackknife over one-shot) is between 1.0 and 1.05 (Fig 2B). Combining the bias and variance to get the RMSE (Fig 2C), we find that the one-shot estimate and the jackknife are similar. Their RMSEs mostly vary between 10% and 40% of the true variance being estimated. It is worth noting that both estimates are challenged when there are very few (10) testers in the experiment.

Although not presented in this article, a bootstrap approach to estimating the MRMC variance of the reader-averaged AUC was also implemented. The bootstrap was found to have considerable bias for some of the simulation configurations considered here (the results were presented at the Medical Image Perception Society Conference 2005, Windermere, UK). In particular, the bias of the bootstrap estimate seems to increase with internal noise and the percent of masked pixels in the reader templates. The bootstrap results were not presented in this manuscript because there were too many unresolved questions that we are actively investigating: Why was the bootstrap biased? Is this bias unique to the reader-averaged AUC? Does the bias generalize to all situations where one is sampling from more than one distribution? Is there a right way to simultaneously bootstrap from two distributions?

Future Work

The probabilistic foundation for the MRMC paradigm by BCK has given us a fresh look at many old problems. The simplest of these is to extend the variance derivations and estimates to reader-averaged sensitivity, specificity, and percent correct. The most critical future work will address the variance of the difference between reader-averaged AUCs from different modalities. Other directions for this work should address other study designs besides the fully crossed design—for example, the various hybrid designs analyzed by Obuchowski et al. (25), and other multiple-reader statistics, such as the generic reader.

ACKNOWLEDGMENTS

I would like to acknowledge Harry Barrett and Eric Clarkson for many discussions that clarified their probabilistic foundation, mechanistic approach, to MRMC analysis. Additionally, I would like to thank Bob Wagner for sharing his broad experience with MRMC analysis, as well as his enthusiastic and critical review of my work every step of the way.

REFERENCES

1. Beam C, Layde P, Sullivan D. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996; 156: 209–213.
2. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol 2004; 11:980–995.
3. Green DM, Swets JA. Signal detection theory and psychophysics. New York, Wiley, 1966 [reprint: Krieger, New York, 1974].
4. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
5. Barrett HH. Objective assessment of image quality: effects of quantum noise and object variability. J Opt Soc Am A 1990; 7:1266–1278.
6. Wilcoxon F. Individual comparisons by ranking methods. Biometrics 1945; 1:80–83.
7. Mann HB, Whitney DR. On a test whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947; 18:50–60.
8. Hoeffding W. A class of statistics with asymptotically normal distribution. Ann Math Stat 1948; 19:293–325.
9. Sen PK. On some convergence properties of U-statistics. Calcutta Stat Assoc Bull 1960; 10:1–18.
10. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psych 1975; 12:387–415.
11. Dorfman DD, Alf E Jr. Maximum-likelihood estimation of parameters of signal detection theory-a direct solution. Psychometrika 1968; 33: 117–124.
12. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. J Math Psychol 1969; 6:487–496.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143: 29–36.
14. DeLong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845.
15. Campbell G, Douglas MA, Bailey JJ. Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve. Proc IEEE Comp Cardiol Conference 1988. Computer Soc IEEE, 267–270.
16. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843.
17. Metz CE, Wang PL, Kronman KB. A new approach for testing the significance of differences between ROC curves measured from correlated data. Inform Proc Med Imaging VIII 432–445.
18. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, Academic Press, 1982.
19. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
20. Dorfman DD, Berbaum KS, Lenth RV, et al. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998; 5:591–602.
21. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: validation with computer simulation. Acad Radiol 1997; 4:298–303.
22. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic (ROC) index estimates. Acad Radiol 1997; 4:587–600.
23. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. Commun Stat 1995; 24:285–308.
24. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. Acad Radiol 1995; 2:S22–S29.
25. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Acad Radiol 1995; 2:709–716.
26. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Stat Med 2005; 24:1579–1607.
27. Toledano A, Gatsonis CA. Regression analysis of correlated receiver operating characteristic data. Acad Radiol 1995; 2:S30–S36; discussion S61–S71.
28. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. Can J Stat 2000; 28:731–750.
29. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad Radiol 2000; 7:341–349.
30. Lee MT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics. Appl Statist 2001; 50:337–344.
31. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics 1997; 53:567–578.
32. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: Lessons from recent experience. Acad Radiol 2002; 9:1264–1277.
33. Barrett HH, Kupinski MA, Clarkson E. Probabilistic foundations of the MRMC method. In: MP Eckstein, Y Jiang, eds. Medical imaging 2005: image perception, observer performance, and technology assessment. Proc SPIE 2005; 5749:21–31.
34. Clarkson E, Kupinski MA, Barrett HH. A probabilistic development of the MRMC method. Stat Med. Submitted.

APPENDIX

One-Shot Derivation

In what follows we derive the first and second moments of $A(\mathbf{t}_G)$ in terms of moments of the step function $s(\cdot)$. The variance of $A(\mathbf{T}_{G\Gamma})$ is the second moment minus the first moment squared. Additionally, at the end of this appendix, we outline how to estimate the moments of the step function.

The mean of $A(\mathbf{T}_{G\Gamma})$ is simply $\langle s(t_{1\gamma} - t_{0\gamma})\rangle$, where the $t_{0\gamma}$, $t_{1\gamma}$ represent scores from an arbitrary reader γ, scoring arbitrary signal-absent and signal-present images $\mathbf{g}_0$, $\mathbf{g}_1$. Because $A(\mathbf{T}_{G\Gamma})$ is just the sample mean of R single-reader AUC estimates, each being unbiased, $A(\mathbf{T}_{G\Gamma})$ is an unbiased estimate of the reader-averaged mean.

Following BCK (33,34), we deal with the second moment of $A(\mathbf{T}_{G\Gamma})$ in terms of $a(\mathbf{t}_{Gr})$ that is,

$$\langle A(\mathbf{T}_{G\Gamma})^2\rangle = \left\langle \frac{1}{R^2}\sum_{r=1}^{R}\sum_{r'=1}^{R} a(\mathbf{t}_{Gr})a(\mathbf{t}_{Gr'})\right\rangle. \quad (A.1)$$

Because $a(\mathbf{t}_{Gr})$ equals $a(\mathbf{t}_{Gr'})$ when $r = r'$ we separate the double sum as follows:

$$\langle A(\mathbf{T}_{G\Gamma})^2\rangle = \frac{1}{R^2}\sum_{r=1}^{R}\langle a(\mathbf{t}_{Gr})^2\rangle + \frac{1}{R^2}\sum_{r=1}^{R}\sum_{r'\neq r}^{R}\langle a(\mathbf{t}_{Gr})a(\mathbf{t}_{Gr'})\rangle. \quad (A.2)$$

When we account for the fact that readers are IID, we get

$$\langle A(\mathbf{T}_{G\Gamma})^2\rangle = \frac{1}{R}\langle a(\mathbf{t}_{G\gamma})^2\rangle + \frac{R-1}{R}\langle a(\mathbf{t}_{G\gamma})a(\mathbf{t}_{G\gamma'})\rangle, \quad (A.3)$$

where γ now refers to an arbitrary reader. In the subsections that follow, we further break down the two terms presently above—the second moment and the cross term—into moments of $s(\cdot)$.

Second Moment

In terms of the step function, the second moment of $a(\mathbf{t}_{G\gamma})$ can be written as

$$\langle a(\mathbf{t}_{G\gamma})^2\rangle = \left\langle \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'=1}^{N_0}\sum_{j'=1}^{N_1} s(t_{1j\gamma} - t_{0i\gamma})s(t_{1j'\gamma} - t_{0i'\gamma})\right\rangle. \quad (A.4)$$

Note that $s(t_{1j\gamma} - t_{0i\gamma})$ is not always independent from $s(t_{1j'\gamma} - t_{0i'\gamma})$; sometimes i equals i', or j equals j'. As such, we break this equation into four parts: 1) $i' = i, j' = j$, 2) $i' \neq i, j' = j$, 3) $i' = i, j' \neq j$, and 4) $i' \neq i, j' \neq j$. These four parts are

$$\begin{aligned}\langle a(\mathbf{t}_{G\gamma})^2\rangle &= \left(\frac{1}{N_0N_1}\right)^2\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\langle s(t_{1j\gamma} - t_{0i\gamma})^2\rangle \\ &+ \left(\frac{1}{N_0N_1}\right)^2\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\langle s(t_{1j\gamma} - t_{0i\gamma})s(t_{1j\gamma} - t_{0i'\gamma})\rangle \\ &+ \left(\frac{1}{N_0N_1}\right)^2\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{j'\neq j}^{N_1}\langle s(t_{1j\gamma} - t_{0i\gamma})s(t_{1j'\gamma} - t_{0i\gamma})\rangle \\ &+ \left(\frac{1}{N_0N_1}\right)^2\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1}\langle s(t_{1j\gamma} - t_{0i\gamma})s(t_{1j'\gamma} - t_{0i'\gamma})\rangle.\end{aligned} \quad (A.5)$$

In the first line, because the images are IID and the scores given the images are IID, we can drop the i and j index in the expected value and replace the summations with N_0N_1. In the second line, the expectation is an average over $t_{1j\gamma}$, $t_{0i\gamma}$, and $t_{0i'\gamma}$, which in integral form is

$$\begin{aligned}&\langle s(t_{1j\gamma} - t_{0i\gamma})s(t_{1j\gamma} - t_{0i'\gamma})\rangle \\ &= \int d\gamma\, p_\gamma(\gamma)\int d\mathbf{g}_{1j}\, p_1(\mathbf{g}_{1j})\int dt_1\, p_{t|\mathbf{g},\gamma}(t_1|\mathbf{g}_{1j},\gamma) \\ &\times \int d\mathbf{g}_{0i}\, p_0(\mathbf{g}_{0i})\int dt_0\, p_{t|\mathbf{g},\gamma}(t_0|\mathbf{g}_{0i},\gamma)s(t_1 - t_0) \\ &\times \int d\mathbf{g}_{0i'}\, p_0(\mathbf{g}_{0i'})\int dt_0'\, p_{t|\mathbf{g},\gamma}(t_0'|\mathbf{g}_{0i'},\gamma)s(t_1 - t_0').\end{aligned} \quad (A.6)$$

Written out as above, we see that the last two lines are identical functions of t_1 given $\mathbf{g}_{1j}$ and γ. Therefore, the expression can be reduced to

$$\begin{aligned}\langle s(t_{1\gamma} - t_{0\gamma})s(t_{1\gamma} - t_{0'\gamma})\rangle &= \int d\mathbf{g}_1\, p(\mathbf{g}_1)\int dt_1\, p_{t|\mathbf{g},\gamma}(t_1|\mathbf{g}_1\gamma,) \\ &\times\left[\int d\mathbf{g}_0\, p(\mathbf{g}_0)\int dt_0\, p_{t|\mathbf{g},\gamma}(t_0|\mathbf{g}_0,\gamma)s(t_1 - t_0)\right]^2,\end{aligned} \quad (A.7)$$

or

$$\langle s(t_{1\gamma}-t_{0\gamma})s(t_{1\gamma}-t_{0'\gamma})\rangle = \langle\langle s(t_{1\gamma}-t_{0\gamma})|t_{1\gamma}\rangle^2\rangle. \quad (A.8)$$

Notice that we have dropped unneeded indices i,j,i' because of IID assumptions on the images. As such, the summations given in Line 2 of Eq A.5 can be replaced by $N_0N_1(N_0-1)$.

The third and fourth lines in Eq A.5 can be treated in a manner similar to above so that the second moment of $a(\mathbf{t}_{G\gamma})$ is

$$\begin{aligned}\langle a(\mathbf{t}_{G\gamma})^2\rangle &= c_1\langle s(t_{1\gamma}-t_{0\gamma})^2\rangle + c_2\langle\langle s(t_{1\gamma}-t_{0\gamma})|t_{1\gamma}\rangle^2\rangle \\ &\quad + c_3\langle\langle s(t_{1\gamma}-t_{0\gamma})|t_{0\gamma}\rangle^2\rangle + c_4\langle\langle s(t_{1\gamma}-t_{0\gamma})|\gamma\rangle^2\rangle, \quad (A.9)\end{aligned}$$

where c_1 through c_4 are given in Eq 8.

Cross Term

We now reduce the second term in Eq A.3, the cross term, to simple moments of $s(\cdot)$. We start with

$$\langle a(\mathbf{t}_{G\gamma})a(\mathbf{t}_{G\gamma'})\rangle = \left\langle \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=0}^{N_0}\sum_{j=0}^{N_1}\sum_{i'=0}^{N_0}\sum_{j'=0}^{N_1} s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j'\gamma'}-t_{0i'\gamma'})\right\rangle, \quad (A.10)$$

and recognize that we must break this equation into four parts: 1) $i'=i, j'=j$, 2) $i'\neq i, j'=j$, 3) $i'=i, j'\neq j$, and 4) $i'\neq i, j'\neq j$. These four parts are

$$\begin{aligned}\langle a(\mathbf{t}_{G\gamma})a(\mathbf{t}_{G\gamma'})\rangle &= \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=0}^{N_0}\sum_{j=0}^{N_1} \langle s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j\gamma'}-t_{0i\gamma'})\rangle \\ &\quad + \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=0}^{N_0}\sum_{j=0}^{N_1}\sum_{i'\neq i}^{N_0} \langle s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j\gamma'}-t_{0i'\gamma'})\rangle \\ &\quad + \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=0}^{N_0}\sum_{j=0}^{N_1}\sum_{j'\neq j}^{N_1} \langle s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j'\gamma'}-t_{0i\gamma'})\rangle \\ &\quad + \left(\frac{1}{N_0N_1}\right)^2 \sum_{i=0}^{N_0}\sum_{j=0}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1} \langle s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j'\gamma'}-t_{0i'\gamma'})\rangle\end{aligned} \quad (A.11)$$

Whereas the moments above are in terms of basic step functions $s(\cdot)$, we shall consider the moment in line two for further simplification. The other lines can be treated in a similar fashion.

The expected value in the second line explicitly contains four random variables $t_{0i\gamma}, t_{1j\gamma}, t_{0i'\gamma'}, t_{1j\gamma'}$ and the subscripts indicate that these depend on five more, $\mathbf{g}_{1j},\mathbf{g}_{0i},\mathbf{g}_{0i'},\gamma,\gamma'$. So, in terms of integrals, the expected value is

$$\begin{aligned}\langle s(t_{1j\gamma}-t_{0i\gamma})s(t_{1j\gamma'}-t_{0i'\gamma'})\rangle &= \int d\mathbf{g}_{1j}\, p(\mathbf{g}_{1j}) \\ &\times \left[\int d\gamma\, p(\gamma)\int d\mathbf{g}_{0i}\, p(\mathbf{g}_{0i})\int dt_1\, p_{t|\mathbf{g},\gamma}(t_1\,|\,\mathbf{g}_{1j},\gamma) \times \int dt_0\, p_{t|\mathbf{g},\gamma}(t_0\,|\,\mathbf{g}_{0i},\gamma)\, s(t_1-t_0)\right] \\ &\times \left[\int d\gamma' p(\gamma')\int d\mathbf{g}_{0i'}\, p(g_{0i'})\int dt'_1\, p_{t|\mathbf{g},\gamma}(t'_1\,|\,\mathbf{g}_{1j},\gamma') \times \int dt'_0\, p_{t|\mathbf{g},\gamma}(t'_0\,|\,\mathbf{g}_{0i'},\gamma')\, s(t'_1-t'_0)\right].\end{aligned} \quad (A.12)$$

The expression above was organized so that the brackets identify two identical functions of $\mathbf{g}_{1j}$, so that

$$\langle s(t_{1\gamma}-t_{0\gamma})s(t_{1'\gamma'}-t_{0'\gamma'})\rangle = \langle\langle s(t_{1\gamma}-t_{0\gamma})|\mathbf{g}_1\rangle^2\rangle, \quad (A.13)$$

where superfluous indices are dropped thanks to IID assumptions. Because the moment does not depend on the image indices, the summation in Line 2 of Eq A.11 can replaced by $N_0N_1(N_0-1)$.

Repeating a similar analysis to the other lines of Eq A.11, we find that

$$\begin{aligned}\langle a(\mathbf{t}_{G\gamma})a(\mathbf{t}_{G\gamma'})\rangle &= c_1\langle\langle s(t_{1\gamma}-t_{0\gamma})|\mathbf{g}_0,\mathbf{g}_1\rangle^2\rangle + c_2\langle\langle s(t_{1\gamma}-t_{0\gamma})|\mathbf{g}_1\rangle^2\rangle \\ &\quad + c_4\langle\langle s(t_{1\gamma}-t_{0\gamma})|\mathbf{g}_0\rangle^2\rangle + c_4\langle s(t_{1\gamma}-t_{0\gamma})\rangle^2, \quad (A.14)\end{aligned}$$

where c_1 through c_4 are given in Eq 8.

The Variance and its Estimate

The objective of this appendix was to present the derivation of the variance of $A(\mathbf{T}_{G\Gamma})$ for those interested. It is just the second moment of $A(\mathbf{T}_{G\Gamma})$ minus the mean squared, where the second moment of $A(\mathbf{T}_{G\Gamma})$ (Eq A.3) is a linear combination of the second moment of $a(\mathbf{t}_{G\gamma})$ (Eq A.9) and the cross term (Eq A.14). Putting all this together we get

$$\begin{aligned}\mathrm{var}(A(\mathbf{T}_{G\Gamma})) &= \frac{1}{R}\left[c_1M_1+c_2M_2+c_3M_3+c_4M_4\right] \\ &\quad + \frac{R-1}{R}\left[c_1M_5+c_2M_6+c_3M_7+c_4M_8\right] - M_8, \quad (A.15)\end{aligned}$$

where M_1 thru M_4 correspond to the moments in Eq. A.9, and M_5 thru M_8 correspond to the moments in Eq. A.14. This expression differs from that from BCK (33,34) only in notation; BCK summarize the total variance in terms of linear combinations of conditional variances (ie, central second moments); we use the raw second moments.

The estimate of Eq A.15 was inspired by the work of Campbell et al. (15). In essence we replace expected values, or more specifically, the integrals over distributions, with the corresponding sums over samples. For example, consider Eq A.12, which is the second moment in Eq A.14. Each integral is an average. First we note that there is no replication; thus, we cannot average over t and t'; therefore, those integrals are dropped. Next we start replacing integrals with summations. The average over the signal-present images gets replaced with the summation over the N_1 signal-present images divided by N_1. Likewise, the averages over the readers and the signal-absent images in the second line of Eq A.12 get replaced by summations over the readers and the signal-absent images divided by R and N_0. In the fourth line, we have to be careful. The primes on the reader and the signal-absent image indicate that they are different from the reader and signal-absent image in Line 2. Thus we replace the integral over γ' with the sum over all readers but the one corresponding to the unprimed reader γ, and we divide by $R - 1$ (there are only $R - 1$ terms in this sum). Likewise, we replace the integral over $\mathbf{g}_{0i'}$ with the sum over all signal-absent images but the one corresponding to the unprimed case $\mathbf{g}_{0i}$, and we divide by $N_0 - 1$. The result for $\hat{M}_6$ is given below with the other moment estimates.

$$\hat{M}_1 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1} \frac{s\left(t_{1jr} - t_{0ir}\right)^2}{RN_0N_1}, \tag{A.16}$$

$$\hat{M}_2 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1jr} - t_{0i'r}\right)}{RN_0N_1\left(N_0 - 1\right)}, \tag{A.17}$$

$$\hat{M}_3 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{j'\neq j}^{N_1} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1j'r} - t_{0ir}\right)}{RN_0N_1\left(N_1 - 1\right)}, \tag{A.18}$$

$$\hat{M}_4 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1j'r} - t_{0i'r}\right)}{RN_0N_1\left(N_0 - 1\right)\left(N_1 - 1\right)}. \tag{A.19}$$

$$\hat{M}_5 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1jr'} - t_{0ir'}\right)}{RN_0N_1\left(R - 1\right)}, \tag{A.20}$$

$$\hat{M}_6 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{r'\neq r}^{R} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1jr'} - t_{0i'r'}\right)}{RN_0N_1\left(R - 1\right)\left(N_0 - 1\right)}, \tag{A.21}$$

$$\hat{M}_7 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{j'\neq j}^{N_1}\sum_{r'\neq r}^{R} \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1j'r'} - t_{0ir'}\right)}{RN_0N_1\left(R - 1\right)\left(N_1 - 1\right)}, \tag{A.22}$$

$$\hat{M}_8 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1}\sum_{r'\neq r}^{R} \times \frac{s\left(t_{1jr} - t_{0ir}\right)s\left(t_{1j'r'} - t_{0i'r'}\right)}{RN_0N_1\left(R - 1\right)\left(N_0 - 1\right)\left(N_1 - 1\right)}. \tag{A.23}$$

We take this opportunity to note that for a single *non-random* reader, the variance of AUC (Eq 3) is

$$\begin{aligned}\operatorname{var}\left(a\left(\mathbf{t}_{G\gamma}\right)\right) &= c_1\langle s\left(t_{1\gamma} - t_{0\gamma}\right)^2\rangle + c_2\langle\langle s\left(t_{1\gamma} - t_{0\gamma}\right)|t_{1\gamma}\rangle^2\rangle \\ &\quad + c_3\langle\langle s\left(t_{1\gamma} - t_{0\gamma}\right)|t_{0\gamma}\rangle^2\rangle + \left(c_4 - 1\right)\langle s\left(t_{1\gamma} - t_{0\gamma}\right)|\gamma\rangle^2.\end{aligned} \tag{A.24}$$

An unbiased estimate is

$$\hat{v} = c_1\hat{M}_{1r} + c_2\hat{M}_{2r} + c_3\hat{M}_{3r} + \left(c_4 - 1\right)\hat{M}_{4r}, \tag{A.25}$$

where $\hat{M}_{1r}$ through $\hat{M}_{4r}$ are the same as $\hat{M}_1$ through $\hat{M}_4$ above, except the sums over R and the R normalization factors are deleted. These expressions can be derived following the notation and examples above, eliminating the averages over the reader distribution. They also can be shown to be equivalent to the expressions given by Bamber (10).