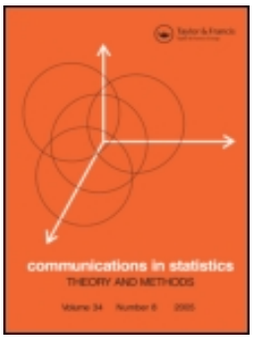

Taylor & Francis
Taylor & Francis Group

Communications in Statistics—Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: www.tandfonline.com/journals/lsta20

A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators

Brandon D. Gallas, Andriy Bandos, Frank W. Samuelson & Robert F. Wagner

To cite this article: Brandon D. Gallas, Andriy Bandos, Frank W. Samuelson & Robert F. Wagner (2009) A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators, Communications in Statistics—Theory and Methods, 38:15, 2586-2603, DOI: 10.1080/03610920802610084

To link to this article: https://doi.org/10.1080/03610920802610084

Communications in Statistics—Theory and Methods, 38: 2586–2603, 2009

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920802610084

Taylor & Francis
Taylor & Francis Group

A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators

BRANDON D. GALLAS[1], ANDRIY BANDOS[2], FRANK W. SAMUELSON[1], AND ROBERT F. WAGNER[1]

[1]NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, Silver Spring, Maryland, USA
[2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

In this article, we analyze the three-way bootstrap estimate of the variance of the reader-averaged nonparametric area under the receiver operating characteristic (ROC) curve. The setting for this work is medical imaging, and the experimental design involves sampling from three distributions: a set of normal and diseased cases (patients), and a set of readers (doctors). The experiment we consider is fully crossed in that each reader reads each case. A reading generates a score that indicates the reader's level of suspicion that the patient is diseased. The distribution of scores for the normal patients is compared to the distribution of scores for the diseased patients via an ROC curve, and the area under the ROC curve (AUC) summarizes the reader's diagnostic ability to separate the normal patients from the diseased ones. We find that the bootstrap estimate of the variance of the reader-averaged AUC is biased, and we represent this bias in terms of moments of success outcomes. This representation helps unify and improve several current methods for multi-reader multi-case (MRMC) ROC analysis.

Keywords Bias; Multi-reader Multi-case (MRMC); Nonparametric AUC; ROC analysis; Three-way bootstrap; Wilcoxon–Mann–Whitney statistic.

Mathematics Subject Classification Primary 62G05; Secondary 62G09, 62C99, 91E45, 92C55, 62J10.

1. Introduction

The text by Swets and Pickett (1982) essentially chartered the modern ROC approach to designing and analyzing experiments with human readers who score diagnostic medical images from patient cases according to their level of suspicion

Received June 11, 2008; Accepted November 7, 2008
Address correspondence to Brandon D. Gallas, NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, Silver Spring, MD 20993-0002, USA; E-mail: brandon.gallas@fda.hhs.gov

that the patient has a specific disease. Although this work provided a model for a general accuracy index that includes key sources of variability, the readers and patient cases, it took another decade before a practical implementation of that model appeared. The first practical solution to the so-called multiple-reader, multiple-case (MRMC) ROC problem was due to "DBM" (Dorfman et al., 1992) who combined leave-one-out jackknife resampling of patient cases with classical ANOVA methods to solve a components-of-variance model. Other ANOVA models and refinements on this basic theme have been presented by several authors (Hillis, 2007; Hillis et al., 2005; Obuchowski and Rockette, 1995; Obuchowski et al., 2004).

The bootstrap was also used for this problem (Dorfman et al., 1995), including an approach due to "BWC" (Beiden et al., 2000) that sets up a family of bootstrap experiments and solves a system of equations to obtain estimates of the variance components in the DBM model. Bandos et al. (2007) considered the *ideal* bootstrap, utilizing empirical distributions, rather than resampling.

Eschewing models, "BCK" (Barrett et al., 2005; Clarkson et al., 2006) derived the variance of an MRMC experiment from first principles of probability theory; their derivation is both model-free and nonparametric. Not surprising, their result is equivalent to what you would get using U-statistics (Randles and Wolfe, 1979). The BCK approach gave rise to two methods for obtaining numerical variance estimates. Kupinski et al. (2006) used the bootstrap and a system of equations, while Gallas (2006) proposed a sample moment estimator that he referred to as the "one-shot". The term "one-shot" indicates that the data is used in a single pass without the need for any resampling strategy.

In this work, we show that the bootstrap-based methods could lead to substantial bias in variance estimates. A central goal of the present work is to exhibit and understand the source of that bias, identify when it could be substantial, and develop a new strategy for reducing or eliminating it. In the process, we shall unify several existing nonparametric MRMC ROC approaches within a common framework.

2. Framework, Models, and Estimates

We consider an ROC experiment where N_0 normal and N_1 diseased cases are interpreted and scored by N_2 readers. The score provided by a reader is their rating of the level of suspicion, or probability, that a given case is diseased. The utilization of all readers across all cases in this way is referred to as a fully crossed design and has become a popular one because it provides the most efficient use of readers and truth-validated cases (Obuchowski, 1995a).

The scores for the ith normal case and the jth diseased case, each read by the rth reader, are given by t_{0ir} and t_{1jr}. This data is mapped to a three-dimensional $(N_0 \times N_1 \times N_2)$ success matrix S whose elements are determined with the following step-function kernel:

$$s_{ijr} = s(t_{1jr} - t_{0ir}) = \begin{cases} 1 & t_{1jr} - t_{0ir} > 0 \\ 1/2 & t_{1jr} - t_{0ir} = 0 \\ 0 & t_{1jr} - t_{0ir} < 0 \end{cases}, \tag{1}$$

which can be interpreted as whether reader r successfully scores the jth diseased case higher than ith normal case.

The step function given above is the U-statistics kernel for the area under the ROC curve (AUC):

$$A = E(s_{ijr}) = P(t_{1jr} > t_{0ir}) + 0.5P(t_{1jr} = t_{0ir}). \tag{2}$$

For a given reader r, the average of the success matrix over the i, j indices (i.e., the average over all comparisons between normal and diseased cases in the sample) can be recognized as the empirical reader-specific AUC

$$\widehat{a}(S \mid r) = \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} \frac{s_{ijr}}{N_0 N_1}. \tag{3}$$

This is the U-statistic estimate that should also be recognized as the well-known Mann–Whitney version of the Wilcoxon statistic. The U-statistic estimate of A is the grand mean of the success matrix

$$\widehat{A}(S) = \sum_{r=1}^{N_2} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} \frac{s_{ijr}}{N_2 N_0 N_1}. \tag{4}$$

The grand mean is often the endpoint of the experiment and the target of a variance analysis, as it is in this manuscript.

In what follows, we shall make the following basic assumptions: readers and cases are independently selected from independent populations. Additionally, given a reader and a case, a score can be a deterministic function of the case, as when the reader is a mathematical classifier, or a score can be a random variable, as might be expected when the reader is a human unable to reproduce the same score on subsequent readings (reader jitter). In either case, we shall assume that scores are independent and identically distributed random variables given either the reader or the case. The difference between scores with and without reader jitter does not impact the present work, because we do not collect replicate readings by the same reader on the same case (cf. Clarkson et al., 2006).

2.1. *Population Variance*

Our principal goal here is to study the variance of $\widehat{A}(S)$ as a measure of uncertainty of the estimated performance of a reader-imaging-technology combination. This will be extended to the difference of estimated performances across competing technologies by a simple step in Sec. 2.7.

In previous work, Gallas (2006) expressed the variance of $\widehat{A}(S)$ as a linear combination of eight moments of the success matrix:

$$V = \mathrm{var}(\widehat{A}(S)) = \underline{c}^t \underline{M} - M_8, \tag{5}$$

where $\underline{c}$ and $\underline{M}$ are given in Table 1 (the underline denotes vector). Note that "success moments" $M_1 - M_7$ are naturally occurring second moments of $s(t_{1jr} - t_{0jr})$ and M_8 is the mean squared. If we subtract the mean squared from each of

Table 1
In this table, we define the success moments and their corresponding coefficients

$\underline{c}$, coefficients	$\underline{M}$, Success moments
$c_1 = \frac{1}{N_0 N_1 N_2}$	$M_1 = E\left(s_{ijr}^2\right)$
$c_2 = \frac{(N_0-1)}{N_0 N_1 N_2}$	$M_2 = E\left(s_{ijr}s_{i'jr} \mid i' \neq i\right)$
$c_3 = \frac{(N_1-1)}{N_0 N_1 N_2}$	$M_3 = E\left(s_{ijr}s_{ij'r} \mid j' \neq j\right)$
$c_4 = \frac{(N_0-1)(N_1-1)}{N_0 N_1 N_2}$	$M_4 = E\left(s_{ijr}s_{i'j'r} \mid i' \neq i, j' \neq j\right)$
$c_5 = \frac{(N_2-1)}{N_0 N_1 N_2}$	$M_5 = E\left(s_{ijr}s_{ijr'} \mid r' \neq r\right)$
$c_6 = \frac{(N_0-1)(N_2-1)}{N_0 N_1 N_2}$	$M_6 = E\left(s_{ijr}s_{i'jr'} \mid i' \neq i, r' \neq r\right)$
$c_7 = \frac{(N_1-1)(N_2-1)}{N_0 N_1 N_2}$	$M_7 = E\left(s_{ijr}s_{ij'r'} \mid j' \neq j, r' \neq r\right)$
$c_8 = \frac{(N_0-1)(N_1-1)(N_2-1)}{N_0 N_1 N_2}$	$M_8 = E\left(s_{ijr}s_{i'j'r'} \mid i' \neq i, j' \neq j, r' \neq r\right)$

the second moments, we generate the (co)variances that can be obtained from a U-statistics approach to this problem. For example:

$$M_6 - M_8 = \operatorname{cov}\left(s_{ijr}, s_{i'jr'} \,\middle|\, \begin{array}{l}\text{case } i' \neq \text{case } i \\ \text{reader } r' \neq \text{reader } r\end{array}\right) \tag{6}$$

$$= \operatorname{var}\left[E(s_{ijr} \mid \text{case } j)\right]. \tag{7}$$

Furthermore, Eq. (5) equals the total variance one would derive following a U-statistics approach.

The derivation of BCK (Barrett et al., 2005; Clarkson et al., 2006) yielded a decomposition of V in reciprocal powers of N_0, N_1, N_2:

$$V = \frac{\alpha_1}{N_0} + \frac{\alpha_2}{N_1} + \frac{\alpha_3}{N_0 N_1} + \frac{\alpha_4}{N_2} + \frac{\alpha_5}{N_0 N_2} + \frac{\alpha_6}{N_1 N_2} + \frac{\alpha_7}{N_0 N_1 N_2}. \tag{8}$$

Here, we'd like to point out that the α's are equivalent to linear combinations of the M's and are themselves variances (see Appendix A and Sec. 2.4), although not the same variances obtained from a U-statistics approach. We refer to the α's as the variance components.

2.2. *Unbiased Estimate*

Gallas (2006) derived an unbiased estimator V that he called the *one-shot* estimator. The one-shot method estimates the success moments with sample moments. These sample moments are in fact the unbiased U-statistic estimators. The estimate of M_6 is:

$$\widehat{M}_{U6} = \sum_{r=1}^{N_2}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1} \frac{s_{ijr}}{N_2 N_0 N_1} \sum_{r'\neq r}^{N_2}\sum_{i'\neq i}^{N_0} \frac{s_{i'jr'}}{(N_2-1)(N_0-1)}, \tag{9}$$

where U indicates the estimate is unbiased, $E(\widehat{M}_{U6}) = M_6$. The sums in $\widehat{M}_{U6}$ average over the two different readers ($r' \neq r$), two different normal cases ($i' \neq i$), and a

single diseased case. It is the skipping of the rth term in the sum over r', the skipping of the ith term in the sum over i', and the corresponding scaling by $(N_2 - 1)$ and $(N_0 - 1)$ that makes the estimator above unbiased. We estimate the other moments in a similar fashion. Please refer to Table 2 for the definition of the full vector $\widehat{\underline{M}}_U$.

Since we have linked the success moments to the U-statistic variances and the variance components ($\underline{\alpha}$), we can estimate these in an unbiased way with the unbiased estimates of the success moments. A nice property of U-statistics are that they have minimum variance among all unbiased estimators.

Unbiasedness sometimes comes at a price in variance estimation; for the *estimates* of the total variance or the variance components outlined above there is no guarantee that they will be positive. This is a common unfortunate consequence of our moment-based approach. We caution that a negative variance estimate usually indicates that there is not enough data to estimate the variance reliably; a negative estimate does not mean that the true variance is zero.

2.3. *Three-Way Bootstrap Estimate*

By assumption, the sets of readers and cases are independent, so their joint distribution separates into factors. Therefore, we can bootstrap (sample with replacement) N_2 readers, N_0 normal cases, and N_1 diseased cases. When we collect the corresponding scores in S_b, we can create a bootstrap replicate $\widehat{A}(S_b)$. We refer to this sampling as the "three-way bootstrap." The sample variance of a set of these bootstrap replicates is the *Monte Carlo (MC) bootstrap* estimate of the variance of $\widehat{A}(S)$. The MC bootstrap is based on the fact that the empirical distribution is the nonparametric maximum likelihood estimate of the actual distribution (Efron, 1982; Efron and Tibshirani, 1993).

We can represent any empirical distribution continuously as a sum of delta functions, $f(x) = \sum_{i=1}^{N} \frac{\delta(x - x_i)}{N}$, or discretely as a multinomial with outcomes $\{x_i\}_{i=1}^{N}$ having equal probabilities $1/N$. Using the empirical distributions of readers and cases, one can derive the *ideal* bootstrap variance of $\widehat{A}(S)$, the variance in the limit of an infinite number of MC bootstraps (Bandos et al., 2007).

Consider, for example, the moment M_6 referred to above. The ideal three-way bootstrap estimate will yield a set of summations similar to those of the unbiased estimator of (Eq. (9)), but the sums will not skip the $r' = r$, $i' = i$ terms and will normalize by $N_2 N_0$ not $(N_2 - 1)(N_0 - 1)$. These differences cause the ideal bootstrap estimate of M_6 to be biased. The full vector of ideal bootstrap estimates $\widehat{\underline{M}}_B$ is defined in Table 2. Each of the bootstrap estimates can be written as a linear combination of the unbiased estimates; the full system can be simply written as $\widehat{\underline{M}}_B = B\widehat{\underline{M}}_U$, where B is given in Appendix B. The matrix B is invertible; we can go back and forth between the bootstrap and unbiased moments as needed.

The variance estimate follows immediately: $\widehat{V}_{3\text{way}} = \underline{c}^t B\widehat{\underline{M}}_U - [B\widehat{\underline{M}}_U]_8$. Since $\widehat{\underline{M}}_U$ is unbiased, we can derive the bias of $\widehat{V}_{3\text{way}}$ in closed-form and show that the bias is positive to leading order and goes to zero only when N_2, N_0, N_1 *all* go to infinity at fixed rates (Appendix B). For example, if the number of cases is fixed, the bias will not go away no matter how many readers we sample! The tradeoff for this bias, however, is that the total variance and variance components tend to be positive. These estimates are, after all, population variances given the empirical distribution.

Table 2
This table gives the different estimators of the vector of success moments M. From left to right there is more bias

	Unbiased $\left(\widehat{\underline{M}}_U\right)$	Unique-reader Hierarchical $\left(\widehat{\underline{M}}_H\right)$	Bootstrap $\left(\widehat{\underline{M}}_B\right)$
M_1 :	s_{ijr}	same as col. 1	same as col. 1
M_2 :	$\sum_{i'\neq i}^{N_0} \frac{s_{i'jr}}{(N_0-1)}$	$\sum_{i'=1}^{N_0} \frac{s_{i'jr}}{N_0}$	same as col. 2
M_3 :	$\sum_{j'\neq j}^{N_1} \frac{s_{ij'r}}{(N_1-1)}$	$\sum_{j'=1}^{N_1} \frac{s_{ij'r}}{N_1}$	same as col. 2
M_4 :	$\sum_{i'\neq i}^{N_0} \sum_{j'\neq j}^{N_1} \frac{s_{i'j'r}}{(N_0-1)(N_1-1)}$	$\sum_{i'=1}^{N_0} \sum_{j'=1}^{N_1} \frac{s_{i'j'r}}{N_0N_1}$	same as col. 2
M_5 :	$\sum_{r'\neq r}^{N_2} \frac{s_{ijr'}}{(N_2-1)}$	same as col. 1	$\sum_{r'=1}^{N_2} \frac{s_{ijr'}}{N_2}$
M_6 :	$\sum_{r'\neq r}^{N_2} \sum_{i'\neq i}^{N_0} \frac{s_{i'jr'}}{(N_2-1)(N_0-1)}$	$\sum_{r'\neq r}^{N_2} \sum_{i'=1}^{N_0} \frac{s_{i'jr'}}{(N_2-1)N_0}$	$\sum_{r'=1}^{N_2} \sum_{i'=1}^{N_0} \frac{s_{i'jr'}}{N_2N_0}$
M_7 :	$\sum_{r'\neq r}^{N_2} \sum_{j'\neq j}^{N_1} \frac{s_{ij'r'}}{(N_2-1)(N_1-1)}$	$\sum_{r'\neq r}^{N_2} \sum_{j'=1}^{N_1} \frac{s_{ij'r'}}{(N_2-1)N_1}$	$\sum_{r'=1}^{N_2} \sum_{j'=1}^{N_1} \frac{s_{ij'r'}}{N_2N_1}$
M_8 :	$\sum_{r'\neq r}^{N_2} \sum_{i'\neq i}^{N_0} \times \sum_{j'\neq j}^{N_1} \frac{s_{i'j'r'}}{(N_2-1)(N_0-1)(N_1-1)}$	$\sum_{r'\neq r}^{N_2} \sum_{i'=1}^{N_0} \times \sum_{j'=1}^{N_1} \frac{s_{i'j'r'}}{(N_2-1)N_0N_1}$	$\sum_{r'=1}^{N_2} \sum_{i'=1}^{N_0} \times \sum_{j'=1}^{N_1} \frac{s_{i'j'r'}}{N_2N_0N_1}$

Note: All estimates above are preceded by $\sum_{r=1}^{N_2} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} \frac{s(t_{1jr}-t_{0ir})}{N_2N_0N_1}$.

2.4. *Three-Way ANOVA*

Let the success matrix be the data for a balanced and complete three-way ANOVA factorial experiment: balanced in the sense that there is only one observation per combination of factors (reader, normal case, diseased case). Then, without appealing to any model, we can write the ANOVA sums of squares as linear combinations of the elements of $\widehat{\underline{M}}_B$, and subsequently, $\widehat{\underline{M}}_U$. When we divide the sums of squares by the appropriate degrees of freedom and take expected values, we find the expected mean squares are linear combinations of the elements of $\underline{M}$, and subsequently, linear combinations of the α variance components. Surprising or not, these variance components correspond exactly to those dictated by the ANOVA model: $\alpha_1 = \sigma_0^2$, $\alpha_2 = \sigma_1^2$, $\alpha_3 = \sigma_{01}^2$, $\alpha_4 = \sigma_2^2$, $\alpha_5 = \sigma_{02}^2$, $\alpha_6 = \sigma_{12}^2$, $\alpha_7 = \sigma_{012}^2 + \sigma_\varepsilon^2$, where the variance components $\sigma_{012}^2, \sigma_\varepsilon^2$ cannot be separated because there is only one observation per combination of random factors. Consequently, the three-way ANOVA total variance estimate is identical to the one-shot variance estimate. Also, all the ANOVA methods and hypothesis tests can be utilized with the caveat that the observations, the success outcomes, are not Gaussian.

2.5. *Two-Way ANOVA and the DBM Method*

The DBM method is based on a balanced and complete three-way ANOVA factorial experiment. The three factors are reader, case, and imaging modality, where the imaging modality is treated as a fixed effect. For each reader and modality, DBM generates the case effect by jackknifing cases (without regard to truth status). Thus the data are the subsequent $N_2 \times (N_0 + N_1)$ AUC pseudovalues for each modality. If we ignore the modality effect, the three-way experiment reduces to a two-way experiment (readers and cases).

The experiment models $a(S \mid r)$ as

$$Y_{rk} = \mu + (\text{single reader effect})_r + (\text{case set effect})_k + \begin{pmatrix} \text{reader} \times \text{case set} \\ \text{interaction \& error} \end{pmatrix}_{rk}, \tag{10}$$

where μ is the overall mean, and the other three terms are independent normal random variables with variances $\underline{\theta} = [\sigma^2_{\text{reader}}, \sigma^2_{\text{case set}}, \sigma^2_{\text{reader}\times\text{set}}]^t$. Since the AUC pseudovalues are per reader and case set, the variances are normalized such that

$$V_{\text{DBM}} = \frac{1}{N_2}\sigma^2_{\text{reader}} + \sigma^2_{\text{case set}} + \frac{1}{N_2}\sigma^2_{\text{reader}\times\text{set}}\ . \tag{11}$$

As above, we can equate the mean squares of the AUC pseudovalues to a linear mapping (B_{MS}) of the bootstrap estimates of the success moments that can then be mapped (B_θ) to estimates of the variance components. Specifically,

$$\widehat{\underline{\theta}}_{\text{DBM}} = \left[\widehat{\sigma}^2_{\text{reader}}, \widehat{\sigma}^2_{\text{case set}}, \widehat{\sigma}^2_{\text{reader}\times\text{set}}\right]^t = B_\theta B_{\text{MS}}\left(B\widehat{\underline{M}}_U\right), \tag{12}$$

where the matrices are shown in Appendix C. The estimate of the total variance $\widehat{V}_{DBM}$ is then obtained by inserting the estimated variance components into Eq. (11).

The framework developed above helps us to analyze the DBM method. First, by comparing Eqs. (5) and (11), we can determine what the variance components should be and then calculate the bias of $\widehat{V}_{\text{DBM}}$ (Appendix C). We find that the bias is positive and asymptotically zero, a result consistent with simulation results (Gallas, 2006). Next, the model given in Eq. (11) hides the separate contributions of the normal and diseased cases:

$$\sigma^2_{\text{case set}} = \frac{1}{N_0}\sigma^2_0 + \frac{1}{N_1}\sigma^2_1 + \frac{1}{N_0 N_1}\sigma^2_{01}, \tag{13}$$

$$\sigma^2_{\text{reader}\times\text{set}} = \frac{1}{N_0}\sigma^2_{02} + \frac{1}{N_1}\sigma^2_{12} + \frac{1}{N_0 N_1}\sigma^2_{012}. \tag{14}$$

The implications of these relations are manifested when scaling the total variance to size a larger experiment (pilot study→pivotal trial). The typical approach is to write

$$\text{var}\left(\widehat{A}(S^{new})\right) = \frac{1}{N_2^{new}}\sigma^2_{\text{reader}} + \lambda\sigma^2_{\text{case set}} + \frac{\lambda}{N_2^{new}}\sigma^2_{\text{reader}\times\text{set}}, \tag{15}$$

where $\lambda = (N_0 + N_1)/(N_0^{new} + N_1^{new})$ and the ratio N_0^{new}/N_1^{new} is similar to N_0/N_1. This scaling essentially ignores σ^2_{01} and σ^2_{012} and assumes $\sigma^2_0 = \sigma^2_1$ and $\sigma^2_{02} = \sigma^2_{12}$. Finally, like the moment-based estimates outlined above, there is no guarantee that the DBM method will yield positive estimates of the total variance or the variance components.

The shortcomings of the DBM method mentioned above do not outweigh its utility. The bias does not seem to be a problem for experiments of moderate size (bias <3% when there are at least 5 readers and 25 normal and diseased cases (Gallas, 2006), the scaling will put you in the right ballpark for sizing a larger trial, and negative variance estimates are an indication of insufficient data. Additionally, the DBM method generalizes to statistics that other than $\widehat{A}(S)$. The

DBM model holds for any reader-averaged statistic, like the reader average of maximum-likelihood estimates from the binormal model (Dorfman and Alf, 1969), or something not related to the detection task, like the reader-averaged expected mean square error in an estimation task.

2.6. *Hierarchical Bootstrap and the BWC Method*

In this section we define a *hierarchical* bootstrap method for estimating the total variance V. This estimator has the advantage of being less biased than the three-way bootstrap and, like the DBM method, does not depend on the step-function kernel of the nonparametric $\widehat{A}(S)$; the method generalizes to other reader-averaged statistics.

In the context of the bootstrap, Davison and Hinkley (1997) considered the problem of bootstrapping *hierarchical* data. This refers to data that result when there is more than a single random effect, as in the simple two-way linear random-effects problem where there is a random group effect plus a random response within a group. They proposed two nonparametric two-stage resampling methods to estimate the second moments of such data. For both methods, the first stage is to sample groups with replacement. At the second stage, one can sample the responses within a group *with* replacement (a "two-way bootstrap" in our terminology) or *without* replacement. They indicated that sampling without replacement at the second stage was less biased than the two-way bootstrap.

Similar to BWC, we consider three observable variances corresponding to three methods of replicating an experiment. The observable variances and their invertible relationship to DBM's variance components are given by

$$\underline{y} = \begin{bmatrix} E\{\mathrm{var}[a(S_{G\gamma}) \mid G]\} \\ E\{\mathrm{var}[a(S_{G\gamma}) \mid \gamma]\} \\ E\{\mathrm{var}[a(S_{G\gamma}) - a(S_{G\gamma'}) \mid \gamma, \gamma']\} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \sigma^2_{\mathrm{reader}} \\ \sigma^2_{\mathrm{case\ set}} \\ \sigma^2_{\mathrm{reader}\times\mathrm{set}} \end{bmatrix} = B_y \underline{\theta}. \quad (16)$$

A consequence of the invertible relationship between $\underline{y}$ and $\underline{\theta}$ and Eq. (12) is that we can also relate $\underline{y}$ to the success moments and to the total variance,

$$V = [1/N_2, 1, 1/N_2] B_y^{-1} \underline{y}. \quad (17)$$

It is helpful to consider BWC's interpretation of these conditional variances. For the variance in the first component of $\underline{y}$, on replication of the experiment the reader is drawn at random but the cases remain fixed. For the variances in the second two components, on replication of the experiment the cases are drawn at random but the reader (or pair of readers) remains fixed. The different conditionings *impose* different hierarchies between the readers and cases that do not otherwise exist between these independent populations. More importantly, the different conditionings allow us to follow a two-stage resampling similar to Davison and Hinkley. Specifically, our hierarchical bootstrap estimate $\widehat{y}_H$ is given by the following:

- $\widehat{y}_{H1}$: calculate the bootstrap reader variance for our single fixed case set;
- $\widehat{y}_{H2}$: calculate the bootstrap case variance for each fixed reader and then average over the readers;
- $\widehat{y}_{H3}$: calculate the bootstrap case variance for each pair of *distinct* readers and then average over these pairs.

The difference between the hierarchical bootstrap and the three-way bootstrap is the treatment of readers. For example, the bootstrap would average over *all* pairs of readers, not just the pairs of *distinct* readers. The ultimate consequence of this is that we can define a set of "unique-reader" estimates of the success moments (see Table 2). Using these, we calculate the bias of our hierarchical bootstrap method and find that to leading order it is indeed less than the bias of the three-way bootstrap (Appendix D).

Finally, the single-modality version of the original BWC method differs from our hierarchical bootstrap in selecting the first component of $\underline{y}$, and the manner of resampling it. BWC used $y_1^* = \mathrm{var}(a(S_{G\gamma}))$ and a three-way bootstrap to estimate y_1^*. Depending on the actual variance components in an experiment, the inclusion of the three-way bootstrapping can add significant bias to the estimate of V, as we show in the simulations below.

2.7. *Comparing Two Modalities*

Extending the representation of the variance of the reader-averaged AUC to the variance of the difference in AUCs across competing modalities is trivial for the fully crossed study design when the same readers and cases are used in both modalities. One simply utilizes the *difference* in success moments instead of the single modality success moments; likewise for variance estimation, one utilizes the *difference* in success outcomes instead of the single modality success outcomes (Bandos et al., 2007; Gallas and Brown, 2008; Gallas et al., 2007).

3. Simulation

We are going to use the same simulation that we have previously described (Gallas, 2006). In brief, the diagnostic task for the reader is to classify whether a noisy image made up of 16×16 pixels contains a lesion (bright spot) or not. In the simulation we generate random images (cases) and readers. An image is the sum of a random vector of noise and a "background alone" or a "background plus lesion". A reader is a linear classifying algorithm with tunable weights, one for each image pixel. Then an ROC score is a deterministic function of an image and a reader (the reader-weighted sum of the image pixels) plus an additional random variable representing reader jitter.

The optimal weights for the simple images form the expected lesion profile when viewed as an image; the weights of this *ideal* reader are tuned to be the expected difference image (the matched filter). In this work *random* reader tunes its weights based on a finite sample of training images; the weights are tuned to the empirical difference between the diseased and normal training images. We add more reader variability with the inclusion of a mask in image space that hides a fraction of the pixels from their view (see Fig. 1). This mask is created independently for each reader, adding variability and correlation structure across readers.

The free parameters in the simulation that we investigate are the image noise level, the reader jitter, the number of training images, the number of testing images, the number of readers, and the probability or fraction of masked pixels. The simulations thus capture the basic structure of the problem in a simple mechanistic model.

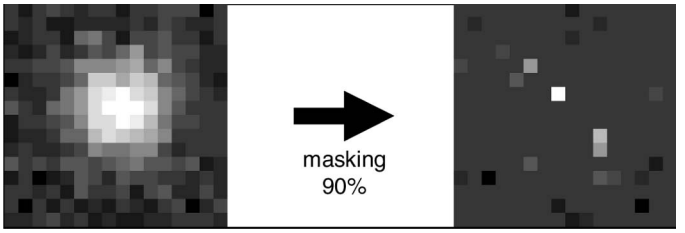

Figure 1. This figure shows a reader template before and after masking 90% of the pixels.

3.1. *Simulation Results*

In what follows, we assess the performance of the variance estimators. The gold standard in this assessment is the empirical variance calculated from 10,000 MC trials per simulation configuration; we refer to these as if they are the true population variances, though some small variability still remains. There were 27 simulation configurations: 9 for each of 3 experiments. Each experiment focuses on two parameters.

Figure 2A shows biases of the ideal versions of the three-way bootstrap and BWC's method normalized by the corresponding true variances for each simulation configuration. The normalization is such that the horizontal dotted line indicates a positive bias that is 10% of the true variance being estimated. Thus, this plot shows

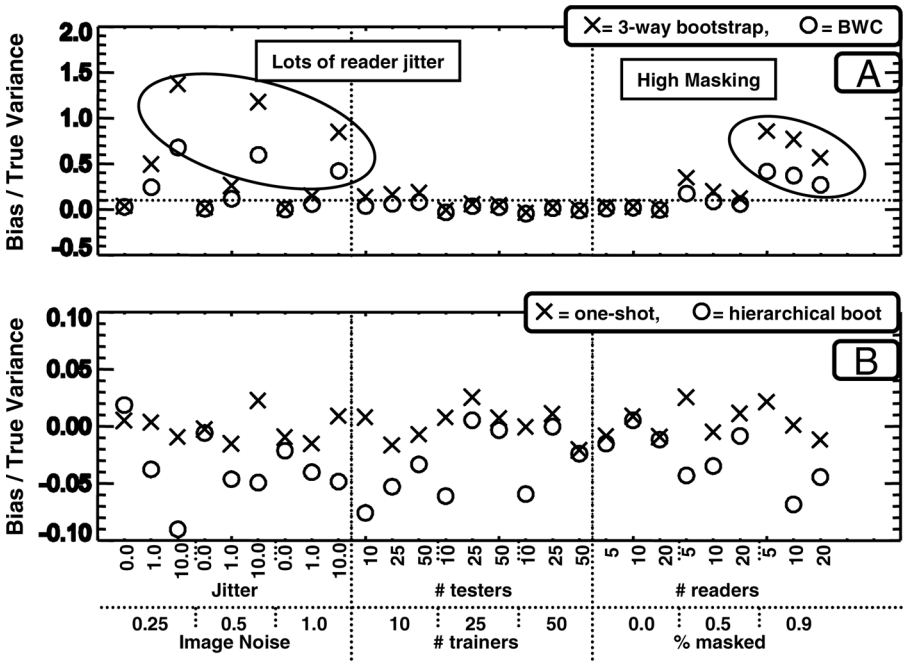

Figure 2. For the 27 simulation configurations, Plot A shows the relative biases of the three-way bootstrap and the BWC estimators, and Plot B shows the relative biases of the one-shot and the hierarchical bootstrap.

that these estimators suffer severe biases for some of the simulation configurations (the encircled points).

Figure 2B shows the normalized biases of the one-shot and our hierarchical bootstrap. Please note that the scale on this plot is about 5% that of Fig. 2A. Thus, as regards bias, these estimators are generally better than those in Fig. 2A.

The root-mean-square error (RMSE) summarizes the total expected error from estimator bias and variability. Figures 3A–B show the RMSE normalized by the corresponding true variances. Plot 3A shows that the severe biases seen in Fig. 2A (the encircled points) dominated the corresponding RMSE calculations; some of these points fall outside the range plotted, as indicated by the arrows. Plot B shows that the one-shot and our hierarchical bootstrap estimators are quite similar in RMSE.

To save space, many results were not shown:

- We confirmed that the biases of the MC versions of the three-way bootstrap and the BWC method (200 replications) are the same as those of the ideal versions; yet, the variances, and thus the RMSEs, are larger. The MC bootstrap is noisy.
- The DBM method has small biases (less than 3%) except for the simulations with only 10 normal cases for testing. The RMSEs were essentially the same as those of the one-shot method.

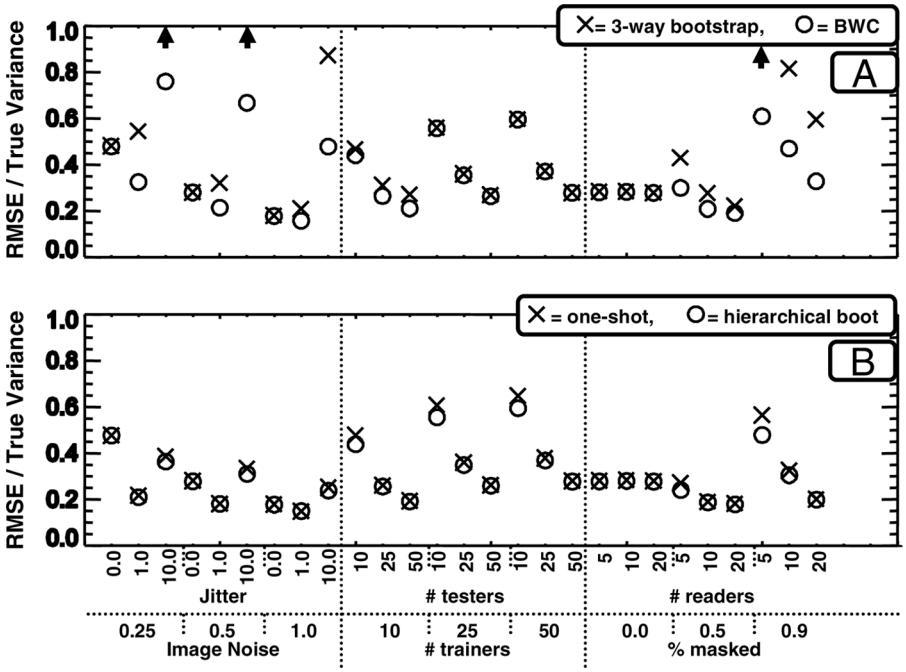

Figure 3. For the 27 simulation configurations, Plot A shows the relative RMSE of the three-way bootstrap and the BWC estimators, and Plot B shows the relative RMSE of the one-shot and the hierarchical bootstrap. The arrows in Plot A indicate that some points were outside the given range.

Finally, the bias and variance of an estimator do not completely characterize *variance* estimators. Variance estimators should be positive and are often not symmetric about their mean. As such, we point out that the one-shot and hierarchical bootstrap estimators can and do go negative for a small number of MC trials. Specifically, fewer than 10 out of 10,000 estimates of total variance were negative for the high internal noise/high case noise configuration, as well as the high masking/few readers configuration.

4. Clinical Observer Study

4.1. *Improving Breast Cancer Diagnosis with Computer-Aided Diagnosis*

We applied the variance estimation methods described in this manuscript to an observer study by Jiang et al. (1999) that asked 10 radiologists to read 104 mammographic images of microcalcification clusters: once without any aid ("unaided"), and once provided with a computer-aided diagnosis (CAD) "likelihood of malignancy" score ("with aid"). In 46 of the cases, the clusters were malignant; in 58, they were benign. The radiologists were asked to report their degree of suspicion that a lesion was malignant by placing a mark on a 5-cm line labeled "benign" at the left and "malignant" at the right. These marks were converted to numerical scores with a ruler and sometimes had ties. Balanced study designs were employed to avoid biases that could result from the sequence of reading the cases under the two reading conditions.

The first two columns of Table 3 show the single-modality estimates of standard error for the AUCs of the unaided radiologists and the radiologists with the computer aid, $\widehat{A}_{\text{unaided}} = 0.6031$, $\widehat{A}_{\text{with aid}} = 0.7489$. The third column shows the standard error in the AUC difference of these two reading conditions $\widehat{A}_{\text{unaided}} - \widehat{A}_{\text{with aid}} = 0.1458$. The last three columns show the square-root of the estimates of the DBM variance components of the AUC difference. The variance components are scaled to the size of the experiment; in other words, they are in units of the standard error, and when squared, they sum to the squared standard error (third column).

In addition to the variance estimation methods described in this manuscript, we also include estimates given by other authors analyzing the same data. The BWC estimates implemented by the original authors (Beiden et al., 2001) are given at the bottom of the table: the last row utilizes the "unsplit" model, and the second-to-last utilizes the "split" model. The unsplit model assumes the variance components do not depend on modality. The split model assumes the components are modality specific. Song and Zhou (2005) estimated the variances of the same data. They assume the variances are "split" in their marginal moment model (MM: row 4), but don't afford the same complexity to their implementation of the DBM model (DBM unsplit by S&Z: row 3). Our implementations of all the variance estimators do not make any assumptions of variances being equal; so they are all of the "split" variety. Differences seen between our AUC and variance estimates and those of the other authors appear to be the result of numerical implementation and precision of the many sums involved.

Table 3 shows that all the nonparametric estimators of variance *for this dataset* give pretty similar results. The three-way and BWC bootstrap estimators are moderately larger than the rest. The biases of these methods keep the estimates of

Table 3

This table shows the estimates of standard error of the mammography-diagnosis study of Jiang et al. (1999). There were 10 readers reading 58 normal and 46 abnormal screen-film (SF) images without and with prompts from a computer aided diagnosis (CAD) algorithm. The last three columns show the square-root of the estimates of the variance components for the AUC difference between observers reading without and with CAD. Observer performance without CAD was 0.6031 and with CAD it was 0.7489. So the difference between the two was 0.1458 in favor of observers reading with CAD. Rows 3 and 4 are taken from S&Z (Song and Zhou, 2005). Rows 11 and 12 are taken from BWC (Beiden et al., 2001)

	Square root variance $\times 10^{-2}$					
	Unaided	With aid	Difference	Components for difference		
	$\sqrt{\widehat{V}_1}$	$\sqrt{\widehat{V}_2}$	$\sqrt{\widehat{V}_{\text{diff}}}$	$\sqrt{\frac{1}{N_2}\widehat{\theta}_1}$	$\sqrt{\widehat{\theta}_2}$	$\sqrt{\frac{1}{N_2}\widehat{\theta}_3}$
one-shot	3.74	3.62	3.24	negative	2.98	1.65
DBM	3.76	3.64	3.26	negative	2.99	1.68
DBM unsplit by S&Z	3.8	3.8	3.5	n/a	n/a	n/a
MM by S&Z	3.6	3.6	3.4	n/a	n/a	n/a
3 way Bootstrap	4.20	4.01	3.87	1.13	3.61	1.62
-ideal	4.18	3.85	3.93	1.21	3.40	1.58
Hierarchical boot	3.74	3.89	3.14	negative	2.90	1.66
-ideal	3.67	3.58	3.19	negative	2.95	1.66
BWC	3.94	4.07	3.45	0.84	2.90	1.66
-ideal	3.95	3.72	3.60	1.21	2.95	1.66
BWC split by BWC	4.06	3.63	3.44	0.95	2.97	1.45
BWC unsplit by BWC	3.86	3.86	3.63	1.24	2.98	1.67

the variance components positive. Finally, making an assumption that the variance components do not depend on modality (or can be pooled), i.e., the unsplit model, does not have a large impact on the estimates of the variance of the difference (as expected by BWC; Beiden et al., 2001).

This dataset shows a statistically significant improvement of using CAD to classify microcalcification clusters regardless of the variance estimate and model used.

5. Discussion and Future Work

In this article, we present a success-moment framework that helps unify the nonparametric MRMC variance estimation methods of DBM and BWC with the more recent probabilistic development by BCK ANOVA and long standing U-statistics. We analyzed these methods and the three-way bootstrap and derived explicit expressions for their bias. The framework was possible because of the simple kernel for AUC, the success outcome for a normal and diseased pair of scores (Eq. (1)).

Practically, the framework allows for efficient programming. This is because all the methods are related to the bootstrap moments, and the bootstrap moments

are perfect squares. Consequently, all the methods can be reduced to order $N_0N_1N_2$ operations, rather than $N_0^2N_1^2N_2^2$.

To assess the different estimators discussed here we ran a simple simulation that captures the basic structure of an MRMC experiment. The simulation results exhibited substantial bias for the bootstrap estimator in some of the simulation configurations. Interestingly, the bias seen in our simulation preceded and motivated the framework developed here. While the framework indicates the bias is possible (Appendix B), it is just as important to know *when* to expect *substantial bias*. The "×" symbols in Fig. 2A show the relative bias increasing with reader jitter and high masking. Both of these effects add to the $\sigma^2_{\text{reader}\times\text{set}}$ variance component, the component that quantifies variability from both reader-case interaction σ^2_{012} and independent reader jitter σ^2_ε. In fact, if we plot the relative bias against the relative contribution of $\sigma^2_{\text{reader}\times\text{set}}$ (relative to the total variance), we find that they increase together (not shown). In particular, the relative bias of the three-way bootstrap is about twice the relative contribution of $\sigma^2_{\text{reader}\times\text{set}}$, and more so as the relative contribution of $\sigma^2_{\text{reader}\times\text{set}}$ increases.

Under the success-moment framework, we improve BWC's method (the hierarchical bootstrap). The main improvement is a reduction in the bias (Figs. 2A–B, Appendix D). This improvement did not destroy the general applicability. The hierarchical bootstrap, like the methods of DBM and BWC, can be implemented with resampling. These methods generalize to any reader-averaged statistic, like the reader average of maximum-likelihood estimates of $\widehat{A}(S)$ from the binormal model (Dorfman and Alf, 1969), or something not related to the detection task, like the reader-averaged expected mean square error in an estimation task.

We also used the framework to determine what, exactly, DBM's method gives as the estimate of total variance. It is a slightly biased estimate (Appendix C). This result, as well as the theoretical work of Hillis et al. (2005), softens criticism based on the jackknifed AUC pseudovalues being correlated (Song and Zhou, 2005).

We have included a short discussion on how to estimate variances of a difference in two AUCs (Sec. 2.7) and demonstrated it with a real dataset (Sec. 4). Future work needs to combine the new variance estimators discussed here with the inference models currently in use today (three-way ANOVA, DBM, OR, and Hillis refinement; Hillis, 2007) to produce confidence intervals and p-values.

While the fully crossed, paired-reader, paired-case study design is the most powerful study design for a specified number of truth-verified cases (Obuchowski, 1995b), it is not always implemented because it may not be possible or practical. Furthermore, if the number of truth-verified cases is not constrained but the number of readings per doctor is, it may be more powerful to have a larger study with each reader reading their own cases. To accommodate such study designs, we have generalized our one-shot variance estimate for AUC (Gallas and Brown, 2008), and presented analogous methods to treat binary performance measures such as sensitivity, specificity, and percent correct (Gallas et al., 2007).

In other work, we have also derived integral expressions for the success moments of the often used Roe and Metz validation/simulation model (Gallas et al., 2007; Roe and Metz, 1997a,b). Future work includes investigating how to force positivity and how well the variance components can be estimated.

Appendix A

While the notation used here is drastically different from that of BCK (Barrett et al., 2005; Clarkson et al., 2006), the variance components (the α's) derived by BCK are linear combinations of the success moments:

$$\underline{\alpha} = B_{\alpha}\underline{M}, \tag{18}$$

where

$$B_{\alpha} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \tag{19}$$

Here, we would like to demonstrate that the terms in the representation given by BCK, the α's, are in fact variances. We demonstrate this for α_6, which BCK showed to be

$$\alpha_6 = M_2 - M_4 - M_6 + M_8. \tag{20}$$

Consider the following variance:

$$\mathrm{var}[E(s_{ijr} \mid \text{case } j,\ \text{reader } r) - E(s_{ijr} \mid \text{reader } r) - E(s_{ijr} \mid \text{case } j)], \tag{21}$$

which equals

$$\mathrm{var}[E(s_{ijr} \mid j, r)] + \mathrm{var}[E(s_{ijr} \mid r)] + \mathrm{var}[E(s_{ijr} \mid j)] - 2\mathrm{cov}[E(s_{ijr} \mid j, r), E(s_{ijr} \mid r)] \tag{22}$$

$$-2\mathrm{cov}[E(s_{ijr} \mid j, r), E(s_{ijr} \mid j)] + 2\mathrm{cov}[E(s_{ijr} \mid j), E(s_{ijr} \mid r)]. \tag{23}$$

Since the readers and cases are all independent, we can write the variance as

$$\mathrm{var}[E(s_{ijr} \mid j, r)] - \mathrm{var}[E(s_{ijr} \mid r)] - \mathrm{var}[E(s_{ijr} \mid j)]. \tag{24}$$

Finally, to get α_6 we can decompose the above three variances into success moments (Table 1) using steps similar to those in Eqs. (6) and (7).

Appendix B

The mapping between the biased and unbiased success moments and their estimates is

$$B = \begin{bmatrix} B_{.25} & 0 \times B_{.25} \\ \frac{1}{N_2} \times B_{.25} & \frac{(N_2-1)}{N_2} \times B_{.25} \end{bmatrix}, \tag{25}$$

where $B_{.25}$ is the 4×4 submatrix given by

$$B_{.25} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{N_0} & \frac{(N_0-1)}{N_0} & 0 & 0 \\ \frac{1}{N_1} & 0 & \frac{(N_1-1)}{N_1} & 0 \\ \frac{1}{N_0 N_1} & \frac{(N_0-1)}{N_0 N_1} & \frac{(N_1-1)}{N_0 N_1} & \frac{(N_1-1)(N_0-1)}{N_0 N_1} \end{bmatrix}. \tag{26}$$

After some unwieldy algebra, we can derive the bias of the three-way bootstrap estimate. In terms of the variance components, the bias is

$$\begin{aligned} E\left(\widehat{V}_{\text{three-way}}\right) - V = &-\frac{1}{N_0^2}\alpha_1 - \frac{1}{N_1^2}\alpha_2 + \frac{2N_0N_1 - 2N_1 - 2N_0 + 1}{N_0^2 N_1^2}\alpha_3 - \frac{1}{N_2^2}\alpha_4 \\ &+ \frac{2N_0N_2 - 2N_2 - 2N_0 + 1}{N_0^2 N_2^2}\alpha_5 + \frac{2N_1N_2 - 2N_2 - 2N_1 + 1}{N_1^2 N_2^2}\alpha_6 \\ &+ \frac{2N_0 + 2N_1 + 2N_2 - 4N_0N_1 - 4N_0N_2 - 4N_1N_2 + 6N_0N_1N_2 - 1}{N_0^2 N_1^2 N_2^2}\alpha_7. \end{aligned} \tag{27}$$

Appendix C

As above, we can equate the mean squares of the AUC pseudovalues to a linear mapping (B_{MS}) of the bootstrap estimates that can then be mapped (B_θ) to estimates of the variance components. Specifically,

$$\underline{\widehat{\theta}}_{\text{DBM}} = \left[\widehat{\sigma}^2_{\text{reader}}, \widehat{\sigma}^2_{\text{case set}}, \widehat{\sigma}^2_{\text{reader}\times\text{set}}\right]^t = B_\theta \; B_{\text{MS}} \left(B \; \underline{\widehat{M}}_U\right), \tag{28}$$

where

$$B_\theta = \begin{bmatrix} \frac{1}{N_0+N_1} & 0 & \frac{-1}{N_0+N_1} \\ 0 & \frac{1}{(N_0+N_1)N_2} & \frac{-1}{(N_0+N_1)N_2} \\ 0 & 0 & \frac{1}{(N_0+N_1)} \end{bmatrix}, \tag{29}$$

$$B_{\text{MS}} = \begin{bmatrix} 0 & 0 & 0 & b_1 & 0 & 0 & 0 & -b_1 \\ 0 & 0 & 0 & 0 & 0 & b_2 & b_3 & -b_2 - b_3 \\ 0 & \frac{b_2}{N_2-1} & \frac{b_3}{N_2-1} & \frac{-b_2-b_3}{N_2-1} & 0 & \frac{-b_2}{N_2-1} & \frac{-b_3}{N_2-1} & \frac{b_2+b_3}{N_2-1} \end{bmatrix}, \tag{30}$$

$$b_1 = \frac{(N_0+N_1)N_2}{N_2-1}, \quad b_2 = \frac{(N_0+N_1-1)N_1N_2}{(N_1-1)^2}, \quad b_3 = \frac{(N_0+N_1-1)N_0N_2}{(N_0-1)^2}. \tag{31}$$

Consequently, the bias of the DBM method is

$$E\left(\widehat{V}_{\text{DBM}}\right) - V = \frac{N_1(N_1-1)\alpha_1 + N_0(N_0-1)\alpha_2 + (N_0+N_1-2)\alpha_3}{(N_0+N_1)(N_0-1)(N_1-1)}. \tag{32}$$

Given the framework we have developed, we can show that the DBM variance components given in Eq. (11) are (according to the normalization)

$$\underline{\theta} = N_2 \begin{bmatrix} 0 & 0 & 0 & N_2^{-1} & 0 & 0 & 0 & -N_2^{-1} \\ 0 & 0 & 0 & 0 & c_1 & c_2 & c_3 & c_4 - N_2^{-1} \\ c_1 & c_2 & c_3 & c_4 - N_2^{-1} & -c_1 & -c_2 & -c_3 & N_2^{-1} - c_4 \end{bmatrix} \underline{M}. \tag{33}$$

Finally, we can show the variance components from DBM's *model* are equivalent to variances of the success moments

$$\underline{\theta} = \begin{bmatrix} \text{var}[E(s_{ijr} \,|\, \text{reader } r)] \\ \text{var}\{E[\widehat{a}(S \,|\, r) \,|\, \text{case set}]\} \\ \text{var}\{\widehat{a}(S \,|\, r) - E[s_{ijr} \,|\, \text{reader } r] - E[\widehat{a}(S \,|\, r) \,|\, \text{case set}]\} \end{bmatrix}. \tag{34}$$

Appendix D

Using the framework that we have developed above, we have calculated the bias of our hierarchical estimate. It is:

$$E\left(\widehat{V}_{\text{H}}\right) - V = -\frac{1}{N_0^2}\alpha_1 - \frac{1}{N_1^2}\alpha_2 + \frac{2N_0N_1 - 2N_1 - 2N_0 + 1}{N_0^2 N_1^2}\alpha_3 \tag{35}$$

$$+ \frac{N_0 - 1}{N_0^2 N_2}\alpha_5 + \frac{N_1 - 1}{N_1^2 N_2}\alpha_6 + \frac{3N_0N_1 - 2N_1 - 2N_0 + 1}{N_0^2 N_1^2 N_2}\alpha_7. \tag{36}$$

Acknowledgments

The authors would like to thank Dr. Yulei Jiang for providing us with his high-quality dataset comparing radiologists with and without a computer aid to classify microcalcification clusters. We'd also like to thank the many colleagues at each institution that have contributed their time and expertise in discussions related to this work. This work is supported in part by grants # EB002106 and EB001694 to the University of Pittsburgh from the National Institute for Biomedical Imaging and Bioengineering, National Institutes of Health, Department of Health and Human Services.

References

Bandos, A. I., Rockette, H. E., Gur, D. (2007). Exact bootstrap variances of the area under the ROC curve. *Commun. Statist. A Theor.* 36(13):2443–2461.

Barrett, H. H., Kupinski, M. A., Clarkson, E. (2005). Probabilistic foundations of the MRMC method. In: Eckstein, M. P., Jiang, Y., eds. *Medical Imaging 2005: Image Perception, Observer Performance, and Technology Assessment*. Vol. 5749. *Proc. SPIE*, pp. 21–31.

Beiden, S. V., Wagner, R. F., Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Acad. Radiol.* 7(5):341–349.

Beiden, S. V., Wagner, R. F., Campbell, G., Metz, C. E., Jiang, Y. (2001). Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. *Acad. Radiol.* 8(7):605–615.

Clarkson, E., Kupinski, M. A., Barrett, H. H. (2006). A probabilistic model for the MRMC method. Part 1. Theoretical development. *Acad. Radiol.* 13(11):1410–1421.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Dorfman, D. D., Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *J. Math. Psychol.* 6:487–496.

Dorfman, D. D., Berbaum, K. S., Metz, C. E. (1992). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest. Radiol.* 27(9):723–731.

Dorfman, D. D., Berbaum, K. S., Lenth, R. V. (1995). Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. *Acad. Radiol.* 2(7): 626–633.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability*. New York: Chapman & Hall.

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

Gallas, B. D., Brown, D. G. (2008). Reader studies for validation of CAD systems. *Neur. Netw.* 21(2–3):387–397.

Gallas, B. D., Pennello, G. A., Myers, K. J. (2007). Multi-reader multi-case variance analysis for binary data. *J. Opt. Soc. Amer. A* 24(12):B70–B80.

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

Hillis, S. L., Obuchowski, N. A., Schartz, K. M., Berbaum, K. S. (2005). A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data. *Statist. Med.* 24(10):1579–1607.

Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Metz, C. E., Giger, M. L., Doi, K. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Acad. Radiol.* 6(1):22–33.

Kupinski, M. A., Clarkson, E., Barrett, H. H. (2006). A probabilistic model for the MRMC method. Part 2. Validation and applications. *Acad. Radiol.* 13(11):1422–1430.

Obuchowski, N. A. (1995a). Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Acad. Radiol.* 2(Suppl 1): S22–S29.

Obuchowski, N. A. (1995b). Multireader receiver operating characteristic studies: a comparison of study designs. *Acad. Radiol.* 2(8):709–716.

Obuchowski, N. A., Rockette, H. E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Commun. Statist. Simul.* 24(2):285–308.

Obuchowski, N. A., Beiden, S. V., Berbaum, K. S., Hillis, S. L., Ishwaran, H., Song, H. H., Wagner, R. F. (2004). Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad. Radiol.* 11(9):980–995.

Randles, R. H., Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley and Sons.

Roe, C. A., Metz, C. E. (1997a). Dorfman–Berbaum–Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: validation with computer simulation. *Acad. Radiol.* 4:298–303.

Roe, C. A., Metz, C. E. (1997b). Variance-component modeling in the analysis of receiver operating characteristic (ROC) index estimates. *Acad. Radiol.* 4:587–600.

Song, X., Zhou, X.-H. (2005). A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. *Biostatistics* 6(2):303–312.

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