Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Exact Bootstrap Variances of the Area Under ROC Curve

Andriy I. Bandos [a], Howard E. Rockette [a] & David Gur [b]
[a] Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[b] Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
Published online: 25 Sep 2007.

To cite this article: Andriy I. Bandos, Howard E. Rockette & David Gur (2007) Exact Bootstrap Variances of the Area Under ROC Curve, Communications in Statistics - Theory and Methods, 36:13, 2443-2461, DOI: 10.1080/03610920701215811

To link to this article: http://dx.doi.org/10.1080/03610920701215811

Communications in Statistics—Theory and Methods, 36: 2443–2461, 2007

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920701215811

Taylor & Francis
Taylor & Francis Group

Survival Analysis

Exact Bootstrap Variances of the Area Under ROC Curve

ANDRIY I. BANDOS[1], HOWARD E. ROCKETTE[1], AND DAVID GUR[2]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[2]Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

The area under the Receiver Operating Characteristic (ROC) curve (AUC) and related summary indices are widely used for assessment of accuracy of an individual and comparison of performances of several diagnostic systems in many areas including studies of human perception, decision making, and the regulatory approval process for new diagnostic technologies. Many investigators have suggested implementing the bootstrap approach to estimate variability of AUC-based indices. Corresponding bootstrap quantities are typically estimated by sampling a bootstrap distribution. Such a process, frequently termed Monte Carlo bootstrap, is often computationally burdensome and imposes an additional sampling error on the resulting estimates. In this article, we demonstrate that the exact or ideal (sampling error free) bootstrap variances of the nonparametric estimator of AUC can be computed directly, i.e., avoiding resampling of the original data, and we develop easy-to-use formulas to compute them. We derive the formulas for the variances of the AUC corresponding to a single given or random reader, and to the average over several given or randomly selected readers. The derived formulas provide an algorithm for computing the ideal bootstrap variances exactly and hence improve many bootstrap methods proposed earlier for analyzing AUCs by eliminating the sampling error and sometimes burdensome computations associated with a Monte Carlo (MC) approximation. In addition, the availability of closed-form solutions provides the potential for an analytical assessment of the properties of bootstrap variance estimators. Applications of the proposed method are shown on two experimentally ascertained datasets that illustrate settings commonly encountered in diagnostic imaging. In the context of the two examples we also demonstrate the magnitude of the effect of the sampling error of the MC estimators on the resulting inferences.

Received August 1, 2006; Accepted December 8, 2006
Address correspondence to Andriy I. Bandos, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, A431 Crabtree Hall, 130 DeSoto Street, Pittsburgh, PA 15261, USA; E-mail: anb61@pitt.edu

Keywords AUC; Bootstrap; Exact variance; Multiple readers; ROC; Variance components.

Mathematics Subject Classification Primary 62P10; Secondary 62G09, 62C99, 92B15.

1. Introduction

A common approach to the evaluation of diagnostic markers, systems, technologies, or practices (often termed modalities) involves an assessment of the inherent ability of a modality to discriminate between subjects with and without the condition of interest (e.g., a specific abnormality). In situations where the output of the diagnostic modality is ordinal and the abnormality status is known for all subjects, an ROC curve is defined as a plot of sensitivity vs. 1-specificity computed at different possible *thresholds*. Receiver operating characteristic (ROC) analysis originated in signal detection theory and presently is a widely used tool for the evaluation of diagnostic modalities (Hanley, 1989; Pepe, 2003; Swets and Picket, 1982; Zhou et al., 2002). The area under the ROC curve (AUC) is a measure of the overall diagnostic performance of a modality and has a practically relevant interpretation as the probability that the diagnostic system correctly distinguishes between randomly selected *normal* (without the condition of interest) and *abnormal* (with the condition of interest) subjects (Bamber, 1975; Hanley and McNeil, 1982). The difference in the AUCs is often used for comparing diagnostic modalities and has been applied in many areas including the regulatory approval process (Wagner et al., 2002).

When modalities are compared it is a common practice to collect the data under a paired design to control for the subjects-related variability. Under such a design, the same set of normal and abnormal subjects are evaluated under different modalities. In the field of diagnostic test evaluation, in general, and in diagnostic imaging, in particular, an observer (reader) is often involved in the diagnostic process and the variability between readers' performance levels is well recognized (Beam, 1992; Hanley, 1989; Metz, 1989; Obuchowski, 1995; Rockette et al., 1999; Swets and Picket, 1982). As a result, data are frequently collected under a multi-reader paired design where the same subjects are rated independently by multiple readers under each modality. Several parametric, semi-parametric, and nonparametric methods have been developed for the analysis of ROC data collected under single- (Bandos et al., 2005; Dorfman and Alf, 1969; Dodd and Pepe, 2003a; DeLong et al., 1988; Metz et al., 1998; Venkatraman and Begg, 1996; Wieand et al., 1983, 1989) and multi-reader designs (Beiden et al., 2000; Bandos et al., 2006; Dorfman et al., 1992; Gallas, 2006; Ishwaran and Gatsonis, 2000; Obuchowski and Rockette, 1995; Song, 1997) and a number of these suggest using the nonparametric estimator of the AUC as a primary statistic.

The bootstrap is a powerful approach that for decades has been used to generate robust nonparametric inferences (Davison and Hinkley, 1997; Efron and Tibshirani, 1993). In the diagnostic test evaluation field the bootstrap is often used for estimating the variability of the AUC and its extensions when the alternative estimators may violate parametric assumptions, are difficult to derive, or unavailable. When the assumption of binormality of the data is questionable, Mossman (1995) proposed to estimate the variability of the AUC by bootstrapping. Dorfman et al. (1995) bootstrapped AUCs in the multi-reader setting in order

to evaluate the conclusions obtained by previously proposed methods in a more robust manner. Beiden et al. (2000) developed an approach to bootstrap estimation of the variance components of the AUC in the multi-reader environment. Rutter (2000) proposed the use of the bootstrap to estimate the mean and variance of the extension of the AUC for patient-clustered data. For the comparison of diagnostic markers with repeated measurements, Emir et al. (2000) developed an extension of the nonparametric AUC and suggested estimating its variance by bootstrapping. Nakas and Yiannoutsos (2004) proposed to use the bootstrap estimator of the variance of the multi-class extension of the AUC. Hillis et al. (2005) suggested that bootstrap variances of the AUC can be used in the multi-reader procedures developed by Obuchowski and Rockette (1995) and Dorfman et al. (1992). When assessing variability of the partial AUC, Dodd and Pepe (2003b) recommended employing the bootstrap technique. For implementation of a proposed goodness of fit test, Zou et al. (2005) suggested using bootstrap for estimation of the variance of the transformed AUC. All these procedures suggest using the Monte Carlo approximation to the exact bootstrap variances.

The exact (*ideal*, Efron and Tibshirani, 1993) bootstrap variance is a variance of the primary statistic over all possible bootstrap samples of the available data. In practice, however, the ideal bootstrap variance is commonly estimated by the Monte Carlo estimator which is the variance over a random sample of the bootstrap samples. Despite its generality and the ease of use the Monte Carlo bootstrap variance estimator has the disadvantages of being computer-intensive and of incorporating an additional error due to resampling. These disadvantages can be eliminated if the exact bootstrap distribution or exact bootstrap variances are derivable. Unfortunately, the exact bootstrap variance is rarely obtainable, notable exceptions being the sample mean, sample median (Maritz and Jarrett, 1978), and L-estimators (Hutson and Ernst, 2000). We have determined that the nonparametric estimator of the AUC also has a structure that permits derivation of the exact bootstrap variance. As a result, many methods that are based on the bootstrap variance of the AUC-related indices can be improved in regard to the precision of estimators and computational effort.

In this article we derive analytical expressions for the ideal bootstrap variances of the several simple AUC-based summaries that are often used in a single- and multi-reader data. In Sec. 3 we derive the ideal bootstrap variance of the AUC computed from single-reader data; then in Sec. 4 we describe how under a multi-reader paired design one can obtain bootstrap variances corresponding to a single random reader and to the average over several fixed or random readers. In Sec. 5, two illustrative examples of the application of the developed formulas are shown. In both examples we generated multiple Monte Carlo estimates in order to illustrate the extent to which the MC sampling error can affect the primary conclusions. We conclude with a discussion of the proposed approach in Sec. 6.

2. Conventions, Notations and Definitions

We assume that the status is known for each subject and hence, every subject can be classified as normal or abnormal. We designate the ordinal output of the diagnostic modality as the subject's *rating* and denote x and y as ratings for normal and abnormal subjects correspondingly. Furthermore, without loss of generality, we will assume that higher values of the ratings are associated with higher probabilities of the presence of abnormality.

The general layout of the data we consider consists of ratings assigned by a sample of R readers to samples of N normal and M abnormal subjects (total $T = N + M$) under each of the modalities. We use subscripts i, k (for normal); j, l (for abnormal); and r, s (for readers) to enumerate subjects and readers. Thus, x_{ir}, y_{jr} denote the ratings assigned to the ith normal and jth abnormal subjects by the rth reader. When using multiple modalities we distinguish between them with the superscript m (e.g., x_{ir}^m). However, when considering a single-reader or a single-modality setting we omit the corresponding indices for the sake of simplicity.

Using the conventions defined above, the nonparametric estimator of the AUC can be written as:

$$\widehat{A} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi(x_i, y_j)}{NM} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{ij}}{NM} = \frac{\psi_{\bullet\bullet}}{NM} = \bar{\psi}_{\bullet\bullet}, \tag{1}$$

where the *order indicator,* ψ, is defined as follows:

$$\psi(x, y) = \begin{cases} 1 & x < y \\ \frac{1}{2} & x = y \\ 0 & x > y \end{cases}. \tag{2}$$

The dot in the place of the index in the subscript of a quantity denotes summation over the corresponding index, and the bar over the quantity, placed in addition to the dot in the subscript, denotes the average over the dotted index.

Under a paired design, the AUC averaged over all the readers can be written as:

$$\widehat{\overline{A}}_{\bullet} = \frac{\sum_{r=1}^{R}\widehat{A}_r}{R} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{ijr}}{RNM} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\bar{\psi}_{ij\bullet}}{NM} = \bar{\psi}_{\bullet\bullet\bullet}. \tag{3}$$

This representation illustrates that the average area has the same structure as the single AUC estimator (1) and allows one to modify expressions derived for a single AUC to those for the average AUC simply by replacing ψ_{ij} with $\bar{\psi}_{ij\bullet}$. Note, that in order to distinguish between different readers in the multi-reader setting we use an additional subscript, r, which we, for brevity, omit in the single reader setting.

We define the *bootstrap space*, B, as conditional on the observed data probability space that is induced by resampling the original data with replacement, and the elements of which are different as long as the sampled observations have different indices in the original dataset. We distinguish random quantities in the bootstrap space, B, by $*$ in the superscript. Thus, i^*, j^*, r^* denote the bootstrap indices for normal, abnormal subject, and reader, correspondingly; $X^*_{i^*r^*}$, $Y^*_{j^*r^*}$ denote the random bootstrap ratings for normal and abnormal subjects; $\Psi^*_{i^*j^*r^*}$ denotes the random bootstrap order indicator (we will use the capital X, Y, and Ψ to highlight the random nature of the corresponding quantities); and A^* denotes the random bootstrap value of AUC (we omit ^ in denoting the bootstrap value of the estimator for brevity).

3. Exact Bootstrap Variances With Single Reader Data

The essence of the bootstrap approach is to generate from a single original data multiple datasets that are then used for inferential purposes. The *bootstrap datasets*

(or *bootstrap samples*) are created by sampling with replacement the independent elements of the original data. The data layout typically used for the evaluation of a diagnostic system is based on a sample of independent subjects. Correspondingly, we treat subjects as units for the bootstrap resampling scheme.

Since the sample of subjects is composed from the two independent samples of normal and abnormal subjects, we resample within the corresponding subsets (normal subjects separately from abnormal). A key concept in our derivation of analytical expressions for the ideal bootstrap quantities is to consider all the possible realizations of the bootstrap values as distinct regardless of their actual values. That is, under the nonparametric bootstrap approach (Efron and Tibshirani, 1993) that we adopted, a normal (abnormal) subject drawn for a bootstrap-sample can with equal probability be one of the normal (abnormal) subjects present in the original data, regardless of the values of the ratings. In this section we consider a single-reader setting, in which the subject-specific datum consists of only one rating. Hence, the distribution of the random bootstrap ratings within a bootstrap space (i.e., conditional on the observed data) can be summarized as follows:

$$\forall i, i^* = 1, \ldots, N \;\; P_B\left(X^*_{i^*} = x_i\right) = \frac{1}{N}, \quad \forall j, j^* = 1, \ldots, M \;\; P_B\left(Y^*_{j^*} = y_j\right) = \frac{1}{M} \tag{4}$$

or

$$X^*_{i^*}\big|_B \overset{i.i.d.}{\sim} Uniform\left[\{x_i\}_{i=1}^N\right] \quad Y^*_{j^*}\big|_B \overset{i.i.d.}{\sim} Uniform\left[\{y_j\}_{j=1}^M\right].$$

Note that since our objective is to derive an expression for the ideal bootstrap quantities we are concerned only with the distributions of the bootstrap random variables within a bootstrap space. These conditional distributions are different from the unconditional distributions of the bootstrap ratings, namely, from the distributions in the larger probability space where observed ratings x_i, y_j are merely the realizations of the random ratings X_i, Y_j.

It is also important to note that the distributions in (4) are the distributions over the observed ratings with different indices, rather than over the observed ratings with different values. Namely, we treat the ratings assigned to different subjects (i.e., indexed differently) as distinct points regardless of whether the numerical values of ratings are equal or different. Such an approach allows for treating ordered-categorical data (where ties are possible) and continuous data (without ties) in the same manner.

As a function of the independent uniformly distributed random quantities $X^*_{i^*}$ and $Y^*_{j^*}$, the bootstrap value, $\Psi^*_{i^*j^*}$, of the order indicator defined in (2), is a random variable uniformly distributed over the values computed from all possible pairs of normal and abnormal ratings (Appendix A). Using the same reasoning, one can describe the bootstrap distributions of the product of the two order indicators that are based on the same normal and abnormal subjects, on the same normal and different abnormal subjects, on the different normal and the same abnormal subjects, or on completely different subjects. While the bootstrap values of the two order indicators based on completely different subjects are independent, it is not the case when these are based on the same subject(s) (Appendix A).

The nonparametric estimator of the AUC is a simple average of the order indicators (1). Thus, its variance, regardless of a specific sample space, can be expressed in terms of generally unknown expectations of the product of two order

indicators (Noether, 1967; van der Vaart et al., 1998). We show in Appendix A that because of the uniform nature of the distributions of the order indicators in the bootstrap sample space, B, the exact bootstrap expectation of the products can be computed directly from the original data. Then, as shown in Appendix A, the ideal bootstrap variance of the nonparametric AUC estimator can be written in the following form:

$$Var_B(A^*) = \frac{\sum_{i=1}^{N}\left(\bar{\psi}_{i\bullet}-\bar{\psi}_{\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\bar{\psi}_{\bullet j}-\bar{\psi}_{\bullet\bullet}\right)^2}{M^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\psi_{ij}-\bar{\psi}_{i\bullet}-\bar{\psi}_{\bullet j}+\bar{\psi}_{\bullet\bullet}\right)^2}{N^2M^2}. \quad (5)$$

The expression in (5) is the ideal bootstrap variance of the AUC that is computed using values of a single diagnostic marker which could represent the ratings obtained from a specific diagnostic modality that incorporates the reader as an integral part of the system. If a reader was indeed involved in the process of assigning the ratings, the above expression can also be referred to as the within specific reader variability.

Under a paired design, the ideal bootstrap variance of the AUC difference can also be easily obtained using the original data. Namely, exploiting the paired structure of the data, the variance of the AUC difference can be obtained from (5) by replacing ψ_{ij} with w_{ij}, where w_{ij} denotes the difference between the order indicators of the two modalities (i.e., $w_{ij} = \psi^1_{ij} - \psi^2_{ij}$).

4. Exact Bootstrap Variances With Multi-Reader Data

In order to better control for subject effect, multi-reader data is commonly constructed from ratings assigned by independent readers to the same set of independent normal and abnormal subjects. For such multi-reader paired data, we consider the bootstrap resampling scheme with independent resampling units consisting of readers, normal subjects, and abnormal subjects. Then, a bootstrap sample includes the ratings corresponding to the units sampled from the original dataset.

One of the frequent uses of multi-reader data is to estimate the variability of the measure of interest (e.g., AUC) computed from the ratings assigned by a randomly selected reader to randomly selected subjects. For this purpose it is sufficient to create a lower-dimensional single-reader bootstrap dataset from the original multi-reader data rather than creating the multi-reader bootstrap datasets. The process of creating a single-reader bootstrap dataset can be described as a bootstrap of hierarchical structure (Davison and Hinkley, 1997) of the ratings. Namely, first we select a set of the ratings assigned by a specific reader to all available subjects (i.e., resample a reader); next, within a selected set of ratings, we bootstrap the ratings separately for normal and abnormal subjects (in the same manner as was done in Sec. 3 for the single-reader data).

We first obtain the ideal bootstrap variance of the nonparametric AUC estimator corresponding to a randomly selected reader (see Appendix B for details). The variance of such AUC can be written as a simple average of the variances computed for each reader according to formula (5) plus the sample-variability of the reader-specific estimates of the AUC computed from the original data $\left(\widehat{A}_r = \bar{\psi}_{\bullet\bullet r}\right)$,

namely:

$$Var_B\left(A^*_{r^*}\right) = \frac{\sum_{r=1}^{R} Var_B\left(A^*_{r^*} \middle| r^* = r\right)}{R} + \frac{\sum_{r=1}^{R} \left(\bar{\psi}_{\bullet\bullet r} - \bar{\psi}_{\bullet\bullet\bullet}\right)^2}{R}. \tag{6}$$

In addition to a single AUC derived from the ratings of a random reader, another multi-reader index that is frequently of interest is the reader-averaged AUC. The interest may lay in the averaging of the AUC over all given readers $\left(\overline{A}^*_{\bullet}\right)$ as well as in averaging over the same number of randomly selected readers $\left(\overline{A}^*_{\bullet^*}\right)$. Under the paired design considered in this article, the ideal bootstrap variance of the AUC averaged over the given readers can then be obtained, similar to deriving (3) from (1), by replacing ψ_{ij} with $\bar{\psi}_{ij\bullet}$ in formula (5), namely:

$$\begin{aligned} Var_B\left(\overline{A}^*_{\bullet}\right) = & \frac{\sum_{i=1}^{N} \left(\bar{\psi}_{i\bullet\bullet} - \bar{\psi}_{\bullet\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M} \left(\bar{\psi}_{\bullet j\bullet} - \bar{\psi}_{\bullet\bullet\bullet}\right)^2}{M^2} \\ & + \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \left(\bar{\psi}_{ij\bullet} - \bar{\psi}_{i\bullet\bullet} - \bar{\psi}_{\bullet j\bullet} + \bar{\psi}_{\bullet\bullet\bullet}\right)^2}{N^2 M^2}. \end{aligned} \tag{7}$$

The variance of the AUC averaged over a random sample of readers is slightly more cumbersome but also derivable. Namely, first this variance is written as a linear combination of the variance of a single AUC and the covariance between two AUCs (ideal bootstrap values of which are the same for all readers). The ideal bootstrap variance of a single random-reader's AUC is provided in (6), and the ideal bootstrap covariance between two random-reader's AUCs can be shown to be equal to the variance of the AUC averaged over all given readers. This is not surprising since the average over all given readers is the exact bootstrap expectation of the AUC for a selected sample of cases. Thus, the ideal bootstrap variance of the random readers-averaged AUC is a simple combination of the exact bootstrap variances for AUC corresponding to a single random and fixed reader and is described by:

$$Var_B\left(\overline{A}^*_{\bullet^*}\right) = \frac{Var_B\left(A^*_{r^*}\right)}{R} + \frac{R-1}{R} \times Var_B\left(\overline{A}^*_{\bullet}\right). \tag{8}$$

Finally, under the multi-reader paired design considered here, the ideal bootstrap variances of the difference in AUCs can be obtained from those of a single AUC (6)–(8) by replacing ψ_{ijr} with w_{ijr}.

5. Examples

Example 5.1 (Single Reader Data)**.** We now apply the proposed approach to the comparison of two diagnostic modalities under the single-reader paired design. Specifically, we compute the ideal bootstrap variance of the difference in AUC and use it to construct an asymptotic confidence interval. One of the alternatives to the bootstrap variance is the widely used estimator developed by DeLong et al. (1988). Although this variance estimator possesses good large sample properties, for small samples it may lead to a suboptimal procedure (Bandos et al., 2003). In addition, DeLong's variance estimator is equivalent to the two-sample jackknife variance estimator, and since jackknife variance provides a linear approximation to the bootstrap, the DeLong's procedure might be expected to be inferior compared to the

bootstrap. The data for this example was taken from an observer performance study performed as part of a general effort to understand various aspects of the transition of radiology to a digital environment (Gur et al., 2005; Thaete et al., 1994). In that study the performance of radiologists in detecting interstitial disease on conventional PA chest films was compared with their performance using computed radiography (CR)-acquired images displayed on a high-resolution workstation. The likelihood of the presence or absence of each abnormality was reported by using a "continuous" rating scale (from 0–100).

We begin by computing asymptotic confidence intervals based on the two different variance estimators—the ideal bootstrap variance derived in this article and the two-sample jackknife estimator proposed by DeLong et al. (1988). Since typically asymptotic approaches exhibit differences only for small samples, for the present example we take a subset of the original rating data corresponding to a single reader and a random sample of $N = 20$ normal and $M = 20$ abnormal subjects. For this sample dataset the areas under the ROC curves for the conventional films and workstation are 0.8588 and 0.7100, correspondingly. The difference between AUCs for conventional film and workstation is 0.1488 with a conventional (DeLong et al., 1988) jackknife-based asymptotic 95% confidence interval of (0.0166, 0.2809). As discussed at the end of Sec. 4, the ideal bootstrap variance of the AUC difference can be obtained by replacing ψ_{ij} with w_{ij} in formula (5) which results, in this case, in a slightly narrower asymptotic 95% confidence interval of (0.0180, 0.2795).

For the task of testing equality of two AUCs in the setting commonly encountered in diagnostic imaging, Bandos (2005) demonstrated that the bootstrap variance has a smaller upward bias and leads to a statistical test with the Type I error rate comparable to that of DeLong et al. (1988). In other words, in the problems similar to the one considered in this example, the bootstrap variance is likely to provide, in general, a tighter confidence interval with comparable coverage. In order to provide a quantitative justification for the choice of the bootstrap variance in the present example, and as an additional way of illustrating potential applications of the developed closed-form solutions, we conducted a small simulation study designed to reflect the values of the parameters that are close to those observed in the current example. Each rating dataset was simulated from the multivariate normal distribution with the range of parameters (means, variances, covariance) chosen to reflect the pre-specified characteristics (average AUC, AUC difference, correlation) of rating distributions that include those observed in the data. For each parameter pattern 10,000 rating datasets were generated. The estimates of the width and coverage of the asymptotic 95% confidence intervals based on bootstrap and two-sample jackknife variance estimators are summarized in the Table 1. The results demonstrate that the bootstrap variance estimator usually leads to slightly tighter confidence intervals without sacrificing coverage.

Although the ideal bootstrap estimators, being sampling-error free and easier-to-compute, are undoubtedly better than the Monte Carlo bootstrap estimators, the gain in precision of inferences achieved as a result of using the ideal instead of MC estimators depends on individual situation. To illustrate the magnitude of the gain in the context of the present example we first generated 10,000 sets of 200 bootstrap samples. Next, for each set of 200 bootstrap samples we computed the Monte Carlo bootstrap estimate of the variance of the AUC difference. Finally, using 10,000 computed MC variance estimates, we constructed the distributions

Table 1
Average width and estimated coverage of the 95% asymptotic confidence intervals

		$\rho^* = 0$		$\rho = 0.4$		$\rho = 0.6$	
Average AUC	AUC difference	J_2**	B	J_2	B	J_2	B
0.7	0.00	0.4658 (0.94)	0.4599 (0.94)	0.3746 (0.95)	0.3721 (0.95)	0.3140 (0.96)	0.3139 (0.96)
	0.10	0.4632 (0.95)	0.4573 (0.94)	0.3728 (0.95)	0.3703 (0.95)	0.3151 (0.95)	0.3149 (0.95)
	0.15	0.4584 (0.94)	0.4526 (0.94)	0.3716 (0.95)	0.3692 (0.95)	0.3172 (0.95)	0.3169 (0.95)
0.75	0.00	0.4344 (0.95)	0.4292 (0.95)	0.3500 (0.95)	0.3481 (0.95)	0.2956 (0.95)	0.2959 (0.95)
	0.10	0.4309 (0.94)	0.4259 (0.94)	0.3507 (0.95)	0.3487 (0.95)	0.2983 (0.95)	0.2985 (0.95)
	0.15	0.4277 (0.94)	0.4226 (0.94)	0.3508 (0.95)	0.3488 (0.95)	0.3013 (0.95)	0.3012 (0.95)
0.8	0.00	0.3938 (0.95)	0.3897 (0.95)	0.3214 (0.95)	0.3201 (0.95)	0.2710 (0.96)	0.2718 (0.96)
	0.10	0.3912 (0.95)	0.3871 (0.95)	0.3220 (0.94)	0.3206 (0.94)	0.2756 (0.95)	0.2761 (0.95)
	0.15	0.3878 (0.95)	0.3837 (0.94)	0.3236 (0.95)	0.3221 (0.95)	0.2819 (0.94)	0.2820 (0.94)
0.85	0.00	0.3415 (0.95)	0.3388 (0.95)	0.2808 (0.96)	0.2805 (0.96)	0.2408 (0.96)	0.2420 (0.96)
	0.10	0.3396 (0.95)	0.3368 (0.95)	0.2857 (0.94)	0.2851 (0.94)	0.2488 (0.94)	0.2496 (0.94)
	0.15	0.3370 (0.94)	0.3340 (0.94)	0.2905 (0.93)	0.2894 (0.93)	0.2596 (0.93)	0.2597 (0.93)

The numbers in parenthesis are the estimated coverage of the confidence intervals. For each parameter pattern, the estimates are based on 10,000 simulations.

*ρ denotes the correlation between the ratings of the same subject.

**Columns J_2 and B contain width and coverage of the asymptotic CIs that are based on the two-sample jackknife variance (DeLong et al., 1988) and on exact bootstrap variance, correspondingly.

of the MC variance estimator and widths of the corresponding asymptotic 95% confidence intervals (Fig. 1).

From Fig. 1a we can see that values of the MC bootstrap variance are distributed around the ideal bootstrap variance (“mean”≈“ideal”), which is expected since the MC bootstrap variance is an unbiased estimator of the ideal (or infinite) bootstrap variance within the bootstrap space (Efron and Tibshirani, 1993). The width of the distribution is a result of the Monte Carlo sampling error and is directly related to the degree of imprecision of the final inferences. Figure 1b indicates that from one realization of MC estimator to another the CI width can change by as much as 0.1 (from 0.213–0.311) or 38% of its ideal value (0.261) and that it is not unlikely to observe widths in the range from 0.239 (5%)–0.283(95%). These changes in the confidence intervals around the difference in AUC can be substantial considering the fact that differences of 0.05 in AUC are often considered “clinically significant”. The magnitude of the sampling error depends on the number of the bootstrap samples and for any given problem one could always choose a large enough number to make this sampling error tolerable. Our initial selection of 200 bootstrap samples was motivated by a recommendation made by Efron and Tibshirani (1993) for variance estimation. Increasing the number of bootstrap samples to 1,000 results in a more precise MC variance estimator however, the width of the confidence interval still varies from 0.240 to 0.286.

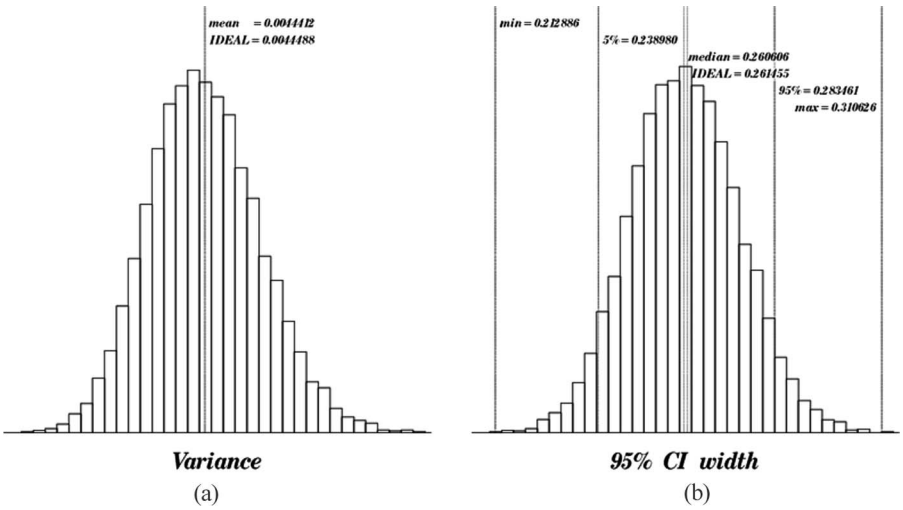

Figure 1. Distributions of the Monte Carlo bootstrap estimators of the variance and corresponding asymptotic 95% confidence interval for the difference in AUCs. (a) The distribution of the Monte Carlo estimator of the variance; (b) the distribution of the width of the asymptotic 95% confidence interval which is based on the MC bootstrap variance. All distributions are based on 10,000 Monte Carlo bootstrap estimates. Each Monte Carlo bootstrap variance estimator is based on 200 random (bootstrap) samples selected from the original dataset with replacement. On both graphs, “IDEAL” denotes the quantity corresponding to the value of the ideal (infinite) bootstrap variance estimator. Labels “mean”, “min”, “5%”, “median”, “95%” and “max” denote the corresponding summaries of generated distributions.

Example 5.2 (Multi-Reader Data)**.** In the next example, we illustrate the application of the proposed approach for planning a multi-reader study. First, using the derived formulae for the ideal bootstrap variances (5)–(7) we compute the ideal values of the bootstrap-based variance components that were originally proposed to be estimated by a MC bootstrap (Beiden et al., 2000). Next, we use the ideal values in estimating the required sample size for a proposed future study. Finally, we generate the distributions of the sample sizes predicted when using the Monte Carlo estimators of the same bootstrap-based variance components.

The data for the example were taken from the original multi-observer reader performance study in which observer performance was measured for three luminance levels and three resolution levels (Herron et al., 2000). Board-certified radiologists interpreted chest radiographs under different reading modes. Readers rated each image with regard to the likelihood of the presence of the specified abnormality using a "continuous" rating scale. The pilot study for the present example was constructed from the ratings for the presence (or absence) of pneumothorax assigned by 5 readers to a random sample of $N = 20$ normal and $M = 20$ abnormal subjects (total number $T = 40$). The estimates of the AUCs are summarized in the Table 2.

The present example focuses on planning a multi-reader study with an objective of detecting the differences in average AUCs of the two modalities. Since subjects and readers represent different sources of variability of the AUC, to develop the sample size of a future study we need to know the estimates of the corresponding variance components (Swets and Picket, 1982). Using the formulation proposed by Roe and Metz (1997) the variability of the difference between AUC under the paired design can be partitioned into the three variance components in the following manner:

$$Var\left(\widehat{A}^1_{rc} - \widehat{A}^2_{rc}\right) = 2 \times \left(\sigma^2_{mr} + \sigma^2_{mc} + \sigma^2_{\varepsilon}\right) \tag{9}$$

where $\widehat{A}^m_{rc}$ is the AUC obtained from the ratings assigned by a random reader (r) to the sample of cases (c) under the mth modality, and $\sigma^2_{mr}, \sigma^2_{mc}, \sigma^2_{\varepsilon}$ are the mode-by-reader, mode-by-case, and the residual variance components, correspondingly.

In 2000, Beiden et al. (BWC) proposed estimating the variance components in (9) by using repeated bootstrap resampling (Monte Carlo bootstrap). Employing

Table 2
AUC estimates

	AUC		
Readers	Low luminance	High luminance	Difference
1	0.900	0.950	0.050
2	0.916	0.950	0.034
3	0.964	0.993	0.029
4	0.791	0.896	0.105
5	0.790	0.863	0.073
Average	0.880	0.942	0.062

the formulas that we developed, the ideal values of such bootstrap-based estimators can be computed directly from the data avoiding the need for resampling and consequently eliminating the Monte Carlo sampling error. Namely, first, the ideal bootstrap variance of the difference in (9) can be obtained from formula (6) derived in Sec. 3.2, by replacing ψ_{ijr} with w_{ijr}. Next, using the formula for the within-reader variability (5) and following the derivation in Appendix B it can be shown that the reader-related variability of the AUC difference ($2\sigma^2_{mr}$) is the sample-variability of the reader-specific differences. Finally, the separation of case- and residual error-related variance components can be achieved by using formula (7) for the ideal bootstrap variance of the reader-average AUC. For the sample dataset used in this example the ideal values of the BWC estimators for mode-by-case, σ^2_{mc}, mode-by-reader, σ^2_{mr}, and error, σ^2_e, variance components are 0.00090, 0.000393, and 0.001420, respectively.

Knowledge of the estimates of variance components can now be used to design a future multi-reader study. It was proposed (Beiden et al., 2000; Obuchowski et al., 2004) that the total number of cases for the study with the same number of readers be estimated based on the following formula:

$$T' = T \times \frac{\sigma^2_{mc} + \frac{\sigma^2_e}{R}}{\frac{1}{2}\left(\frac{\Delta}{z_{1-\alpha}+z_{1-\beta}}\right)^2 - \frac{\sigma^2_{mr}}{R}} \tag{10}$$

where $z_{1-\alpha}$, $z_{1-\beta}$ are the normal percentiles and Δ is the difference between two AUCs that one desires to detect with power $1-\beta$ at significance level α.

Using the computed ideal values of the BWC variance components estimators we can estimate the required sample size directly from the data avoiding repeated sampling. Thus, for a 5-reader study the required number of subjects to detect a difference in average AUCs of $\Delta = 0.05$ at significance level $\alpha = 0.05$ with power of $(1-\beta) = 0.8$ is $T' = 122$.

To illustrate the magnitude of the MC sampling error that is eliminated when using the ideal bootstrap variances instead of MC bootstrap variances in the context of the present example, we generated the distribution of the sample sizes computed from the MC estimates of the bootstrap based variance components. The distributions in Fig. 2 were constructed using 10,000 triplets of MC variance component estimates where each triplet was based on 1,000 bootstrap samples. We chose here 1,000 bootstrap samples since the conventionally recommended number of 200 was clearly too low, sometimes leading to unreasonable sample sizes (e.g., less than 0 or more than 4,000).

Figures 2a–c illustrate that MC bootstrap estimates of the variance components are distributed around the corresponding ideal values ("mean" $\approx$ "ideal"). The width of the distribution characterizes the degree of imprecision caused by the MC sampling error for a corresponding variance component. The effect of the MC sampling error on the primary inferences considered in this example can be seen from the distribution of the resulting sample sizes shown in Fig. 2d. While the sample size computed using the ideal bootstrap values of the variance components is 122, MC estimates of sample size can be as low as 76 and as high as 459, with any value between 96 (5%) and 167 (95%) being not unlikely to be obtained. Increasing the number of the bootstrap samples from 1,000 to 10,000 improves the precision of the MC estimators although there still remains a substantial sampling error that results in sample sizes in the range of 101–155.

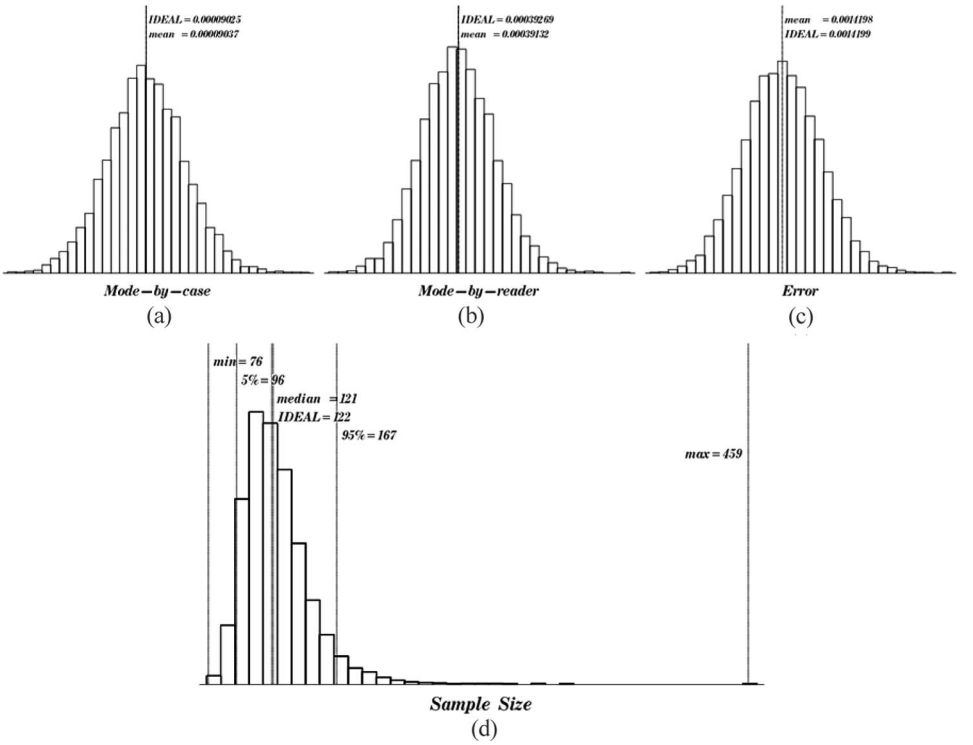

Figure 2. Distributions of the Monte Carlo bootstrap estimators of the variance components and corresponding estimator of the sample size. (a) The distribution of the MC estimator of the "mode-by-case" variance component; (b) the distribution of the MC estimator of the "mode-by-reader" variance component; (c) the distribution of the MC estimator of the "error" variance component; (d) the distribution of the sample size estimator constructed based on the MC estimators of the variance components. All distributions are based on 10,000 Monte Carlo estimates. Each triplet of Monte Carlo variance components estimates is based on 1,000 random (bootstrap) samples selected from the original dataset with replacement. On the graphs, "IDEAL" denotes the ideal (infinite) estimate of the corresponding quantity and "mean", "min", "5%", "median", "95%", and "max" denote the summaries of the generated distributions.

6. Discussion

In this article we derived exact bootstrap variances of the several AUC-based summaries that are often used with single- and multi-reader data. The obtained formulas provide an easy algorithm for computing the ideal bootstrap variances exactly instead of the commonly employed approach of approximating it with Monte Carlo bootstrap estimators. This simplifies and improves the precision of existing bootstrap based methods that use single AUC, AUC difference, and reader-average AUCs computed from fixed or random readers. Moreover, although we focused in this article on the conventional AUC, similar formulas are likely to be derivable for the ideal bootstrap variances of modifications of the AUC index such as patient-clustered AUC (Rutter, 2000), repeated-measures AUC (Emir et al.,

2000), partial AUCs (Dodd and Pepe, 2003b), and multi-class volume (Nakas and Yiannoutsos, 2004).

The two examples that we provided demonstrate that the number of bootstrap samples required to obtain adequately precise Monte Carlo bootstrap estimators is highly dependent on the specific problem. This results in the necessity of either assessing the appropriate number of the bootstrap samples for each specific problem in question, or of using a very large number that may be computationally burdensome as well as unnecessarily large for many problems. Alternatively, in the problems where the closed-form solutions for the ideal bootstrap quantities exist, these can be used to obtain precise results while saving computational resources and eliminating the need for estimating the appropriate number of the bootstrap samples.

Although the bootstrap statistical moments generally possess good statistical properties they should not be used indiscriminately. For instance, the validity of the symmetric confidence intervals for the single AUC for small samples may be questionable because of the asymmetry of the distribution. On the other hand, the symmetric confidence intervals for the AUC difference have better properties (Obuchowski and Lieber, 1998) than those for a single AUC (a fact that can be partially attributed to the general symmetry of the distribution of the difference). Also, depending on the parameter being estimated, the bootstrap estimators may be biased or inferior compared to other estimators (e.g., unbiased estimators for the variances of the AUC in the multi-reader setting were recently developed by Gallas (2006)). The closed-form solutions we developed will facilitate further investigation of the properties of the bootstrap estimators relative to the truth and to other available estimators.

Appendix A

Exact Bootstrap Variance for a Single Fixed Reader AUC

We start by writing a bootstrap value of the nonparametric estimator of AUC as an average of $N * M$ identically (but not independently) distributed random variables (random bootstrap order-indicators) $\Psi^*_{i^*j^*}$, i.e., we replace ψ_{ij} with $\Psi^*_{i^*j^*}$ in expression (1). Next, following the uniform distribution of the bootstrap ratings for normal ($X^*_{i^*}$) and abnormal ($Y^*_{j^*}$) subjects (4), we can derive the distribution of order-indicators in the bootstrap-space B. Note that, as with bootstrap ratings, we consider the distribution of the bootstrap order-indicator, Ψ^*, over the ψ's with different indices rather than over the ψ's with arithmetically different values, namely, we treat the results of the function $\psi(\bullet, \bullet)$ as different as long as at least one of the two arguments are different observations in the original dataset:

$$P_B\left[\Psi^*_{i^*j^*} = \psi_{ij}\right] = P_B\left[\psi\left(X^*_{i^*}, Y^*_{j^*}\right) = \psi\left(x_i, y_j\right)\right] = P_B\left[X^*_{i^*} = x_i, Y^*_{j^*} = y_j\right]$$
$$= P_B\left[X^*_{i^*} = x_i\right] \times P_B\left[Y^*_{j^*} = y_j\right] = \frac{1}{NM}.$$

Similarly, for the product of two bootstrap order-indicators we have:

$$j^* \neq l^* \quad P_B\left[\Psi^*_{i^*j^*} \times \Psi^*_{i^*l^*} = \psi_{ij} \times \psi_{il}\right] = P_B\left[X^*_{i^*} = x_i, Y^*_{j^*} = y_j, Y^*_{l^*} = y_l\right] = \frac{1}{NM^2}$$

$$i^* \neq k^* \quad P_B\left[\Psi^*_{i^*j^*} \times \Psi^*_{k^*j^*} = \psi_{ij} \times \psi_{kj}\right] = P_B\left[X^*_{i^*} = x_i, X^*_{k^*} = x_k, Y^*_{l^*} = y_l\right] = \frac{1}{N^2 M}$$

$$i^* \neq k^*,\ j^* \neq l^* \quad P_B\left[\Psi^*_{i^*j^*} \times \Psi^*_{k^*l^*} = \psi_{ij} \times \psi_{kl}\right]$$

$$= P_B\left[X^*_{i^*} = x_i, X^*_{k^*} = x_k, Y^*_{j^*} = y_j, Y^*_{l^*} = y_l\right] = \frac{1}{N^2 M^2}.$$

Alternatively, the distributions described above can be written as:

$$\begin{aligned}
\Psi^*_{i^*j^*}\big|_B &\sim Uniform\left[\left\{\psi_{ij}\right\}_{i=1,j=1}^{N,M}\right] \\
j^* \neq l^* \quad \Psi^*_{i^*j^*} \times \Psi^*_{i^*l^*}\big|_B &\sim Uniform\left[\left\{\psi_{ij}\psi_{il}\right\}_{i=1,j=1,l=1}^{N,M,M}\right] \\
i^* \neq k^* \quad \Psi^*_{i^*j^*} \times \Psi^*_{k^*j^*}\big|_B &\sim Uniform\left[\left\{\psi_{ij}\psi_{kj}\right\}_{i=1,k=1,j=1}^{N,N,M}\right] \\
i^* \neq k^*, j^* \neq l^* \quad \Psi^*_{i^*j^*} \times \Psi^*_{k^*l^*}\big|_B &\sim Uniform\left[\left\{\psi_{ij}\psi_{kl}\right\}_{i=1,k=1,j=1,l=1}^{N,N,M,M}\right]. \qquad \text{(A.1)}
\end{aligned}$$

Now, we derive the statistical moments of the order indicators in B. Following the uniformity of the distribution of a single-order indicator, its first and second moments can be obtained as:

$$\begin{aligned}
E_B\left(\Psi^*_{i^*j^*}\right) &= \sum_{\psi_{ij}} \psi_{ij} \times P_B\left(\Psi^*_{i^*j^*} = \psi_{ij}\right) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{ij} = \bar{\psi}_{\bullet\bullet} \\
E_B\left(\Psi^{*^2}_{i^*j^*}\right) &= \sum_{\psi_{ij}} \left(\psi_{ij}\right)^2 \times P_B\left(\Psi^*_{i^*j^*} = \psi_{ij}\right) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\psi_{ij}^2 \\
Var_B\left(\Psi^*_{i^*j^*}\right) &= E_B\left(\Psi^{*^2}_{i^*j^*}\right) - E_B\left(\Psi^*_{i^*j^*}\right)^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\psi_{ij} - \bar{\psi}_{\bullet\bullet}\right)^2. \qquad \text{(A.2)}
\end{aligned}$$

The covariances between two order indicators can be derived using the first moment of the corresponding uniform distribution of a product in (A.1). Namely, the covariance of the two order indicators based on the same normal subjects but on different abnormal subjects is:

$$\begin{aligned}
\forall j^* \neq l^* \quad E_B\left(\Psi^*_{i^*j^*} \times \Psi^*_{i^*l^*}\right) &= \frac{1}{NM^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{M}\psi_{ij}\psi_{il} = \frac{1}{N}\sum_{i=1}^{N}\bar{\psi}_{i\bullet}^2 \\
Cov_B\left(\Psi^*_{i^*j^*}, \Psi^*_{i^*l^*}\right) &= \frac{1}{N}\sum_{i=1}^{N}\bar{\psi}_{i\bullet}^2 - \bar{\psi}_{\bullet\bullet}^2 = \frac{\sum_{i=1}^{N}\left(\bar{\psi}_{i\bullet} - \bar{\psi}_{\bullet\bullet}\right)^2}{N}. \qquad \text{(A.3)}
\end{aligned}$$

The covariance of the two order indicators that are based on different normal subjects but on the same abnormal subject can be derived in the same manner, namely:

$$\begin{aligned}
\forall i^* \neq k^* \quad E_B\left(\Psi^*_{i^*j^*} \times \Psi^*_{k^*j^*}\right) &= \frac{1}{N^2 M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{N}\psi_{ij}\psi_{kj} = \frac{1}{M}\sum_{j=1}^{M}\bar{\psi}_{\bullet j}^2 \\
Cov_B\left(\Psi^*_{i^*j^*}, \Psi^*_{k^*j^*}\right) &= \frac{\sum_{j=1}^{M}\bar{\psi}_{\bullet j}^2}{M} - \bar{\psi}_{\bullet\bullet}^2 = \frac{\sum_{j=1}^{M}\left(\bar{\psi}_{\bullet j} - \bar{\psi}_{\bullet\bullet}\right)^2}{M}. \qquad \text{(A.4)}
\end{aligned}$$

Finally, the order indicators based on completely different normal and abnormal subjects are uncorrelated by construction of the bootstrap space (can also be verified by direct calculations).

Using the moments of the order indicators in (A.2)–(A.4), the formulae for the ideal bootstrap variance of the nonparametric estimator of AUC can be obtained as follows:

$$\begin{aligned} Var_B(A^*) &= \frac{1}{(NM)^2}\left\{\sum_{i^*=1}^{N}\sum_{j^*=1}^{M} Var_B\left(\Psi^*_{i^*j^*}\right) + \sum_{i^*=1}^{N}\sum_{j^*=1}^{M}\sum_{\substack{l^*=1\\ l^*\neq j^*}}^{M} Cov_B\left(\Psi^*_{i^*j^*}, \Psi^*_{i^*l^*}\right)\right. \\ &\qquad \left. + \sum_{j^*=1}^{M}\sum_{i^*=1}^{N}\sum_{\substack{k^*=1\\ k^*\neq i^*}}^{N} Cov_B\left(\Psi^*_{i^*j^*}, \Psi^*_{k^*j^*}\right)\right\} \\ &= \frac{1}{N^2M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\psi_{ij} - \bar{\psi}_{\bullet\bullet}\right)^2 + \frac{(M-1)}{N^2M}\sum_{i=1}^{N}\left(\bar{\psi}_{i\bullet} - \bar{\psi}_{\bullet\bullet}\right)^2 \\ &\quad + \frac{(N-1)}{NM^2}\sum_{j=1}^{M}\left(\bar{\psi}_{\bullet j} - \bar{\psi}_{\bullet\bullet}\right)^2. \end{aligned} \tag{A.5}$$

Formula (5) follows from the above after simplification.

Appendix B

Exact Bootstrap Variance for a Single Random Reader AUC

The ideal bootstrap variance of the AUC corresponding to a random reader can be partitioned according to the following formula:

$$Var_B\left(A^*_{r^*}\right) = E_B\left\{Var_B\left(A^*_{r^*} \mid r^*\right)\right\} + Var_B\left\{E_B\left(A^*_{r^*} \mid r^*\right)\right\}. \tag{B.1}$$

Since under the bootstrap resampling scheme a random reader can with equal probability be one of the readers in the original dataset, the first term of the partitioning in (B.1) can be written as:

$$E_B\left\{Var_B\left(A^*_{r^*} \mid r^*\right)\right\} = \frac{\sum_{r=1}^{R} Var_B\left(A^*_{r^*} \mid r^* = r\right)}{R}. \tag{B.2}$$

The right-hand expression (B.2) is a simple average of the within-specific-reader variances shown in (5). Next, to derive the formula for the second term in (B.1), we consider the expectation of the bootstrap value of the AUC within the selected reader $\left(E_B\left(A^*_{r^*} \mid r^*\right)\right)$. Once the reader is selected for the bootstrap sample, the remainder of the bootstrap process is equivalent to bootstrapping subjects in a single-reader problem, therefore, using (A.2), we have:

$$\begin{aligned} E_B\left(A^*_{r^*} \mid r^* = r\right) &= \frac{\sum_{i^*=1}^{N}\sum_{j^*=1}^{M} E_B\left(\Psi^*_{i^*j^*r^*} \mid r^* = r\right)}{NM} \\ &= \frac{\sum_{i^*=1}^{N}\sum_{j*=1}^{M} E_B\left(\Psi^*_{i^*j^*r}\right)}{NM} = \bar{\psi}_{\bullet\bullet r} = \widehat{A}_r. \end{aligned} \tag{B.3}$$

Once again using equal probability of selecting any reader under the bootstrap scheme we obtain:

$$Var_B\left\{E_B\left(A^*_{r^*} \mid r^*\right)\right\} = \frac{\sum_{r^*=1}^{R}\left\{E_B\left(A^*_{r^*} \mid r^*\right) - E_B\left(A^*_{r^*}\right)\right\}^2}{R} = \frac{\sum_{r=1}^{R}\left(\widehat{A}_r - \widehat{\overline{A}}_\bullet\right)^2}{R}. \quad \text{(B.4)}$$

Formula (6) follows immediately after substituting (B.2) and (B.4) into (B.1).

Acknowledgments

This work is supported in part by Public Health Service grants EB002106 and EB001694 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Health Institutes, Department of Health and Human Services.

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Mathemat. Psychol.* 12:387–415.

Bandos, A. (2005). Nonparametric Methods in Comparing Two Correlated ROC Curves. Ph.D. dissertation, University of Pittsburgh, Pittsburgh, PA (http://www.pitt.edu/~graduate/etd/).

Bandos, A. I., Rockette, H. E., Gur, D. (2003). Small sample size properties of the nonparametric comparison of the area under two ROC curves. *Medical Image Perception Society Conference X*, September, Durham, NC.

Bandos, A. I., Rockette, H. E., Gur, D. (2005). A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statist. Med.* 24(18):2873–2893.

Bandos, A. I., Rockette, H. E., Gur, D. (2006). A permutation test for comparing ROC curves in multireader studies. *Academic Radiol.* 13:414–420.

Beam, C. A. (1992). Strategies for improving power in diagnostic radiology research. *Amer. J. Roentgenol.* 159:631–637.

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. *Academic Radiol.* 7:341–349.

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

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845.

Dodd, L. E., Pepe, M. S. (2003a). Semiparametric regression for the area under the receiver operating characteristic curve. *J. Amer. Statist. Assoc.* 98:409–417.

Dodd, L. E., Pepe, M. S. (2003b). Partial AUC estimation and regression. *Biometrics* 59:614–623.

Dorfman, D. D., Alf, Jr. E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating-method data. *J. Mathemat. 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. *Investigative Radiol.* 27:723–731.

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

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Emir, B., Wieand, S., Jung, S. H., Ying, Z. (2000). Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach. *Statist. Med.* 19:511–523.

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

Gur, D., Rockette, H. E., Glenn, S. M., King, J. L., Klym, A. H., Bandos, A. I. (2005). Variability in observer performance studies: experimental observations. *Academic Radiol.* 12:1527–1533.

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: state of the art. *Crit. Rev. Diagnost. Imaging* 29:307–335.

Hanley, J. A., McNeil, B. J. (1982). The meaning and use of the area under receiver operating characteristic (ROC) curve. *Radiology* 143:29–36.

Herron, J. M., Bender, T. M., Campbell, W. L., Sumkin, J. H., Rockette, H. E., Gur, D. (2000). Effects of luminance and resolution on observer performance with chest radiographs. *Radiology* 215:169–174.

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.* 24:1579–1607.

Hutson, A. D., Ernst, M. D. (2000). The exact bootstrap mean and variance of an *L*-estimator. *J. Roy. Statist. Soc. Ser. B (Statist. Methodol.)* 62(1):89–94.

Ishwaran, H., Gatsonis, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canad. J. Statist.* 28:731–750.

Maritz, J. S., Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *J. Amer. Statist. Assoc.* 73(361):194–196.

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiol.* 24:234–245.

Metz, C. E., Herman, B. A., Shen, J. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statist. Med.* 17:1033–1053.

Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROC data. *Med. Decision Making* 15:358–366.

Nakas, C. T., Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statist. Med.* 23:3437–3449.

Noether, G. E. (1967). *Elements of Nonparametric Statistics*. New York: Wiley & Sons Inc.

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

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

Obuchowski, N. A., Lieber, M. L. (1998). Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiol.* 5:561–571.

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. *Academic Radiol.* 11:980–995.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Test for Classification and Prediction.* Oxford: Oxford University Press.

Rockette, H. E., Campbell, W. L., Britton, C. A., Holbert, J. M., King, J. L., Gur, D. (1999). Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Academic Radiol.* 6:723–729.

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

Rutter, C. M. (2000). Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Academic Radiol.* 7:413–419.

Song, H. H. (1997). Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 53(1):370–382.

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

Thaete, F. L., Fuhrman, C. R., Oliver, J. H., Britton, C. A., Campbell, W. L., Feist, J. H., Straub, W. H., Davis, P. L., Plunkett, M. B. (1994). Digital radiography and conventional imaging of the chest: a comparison of observer performance. *Amer. J. Roentgenol.* 162:575–581.

van der Vaart, A. W., Gill, R., Ripley, B. D., Ross, S., Silverman, B., Stein, M. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Venkatraman, E. S., Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 83(4):835–848.

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

Wieand, H. S., Gail, M. M., Hanley, J. A. (1983). A nonparametric procedure for comparing diagnostic tests with paired or unpaired data. *I.M.S. Bull.* 12:213–214.

Wieand, H. S., Gail, M., James, B., James, K. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 76:585–592.

Zhou, X. H., Obuchowski, N. A., McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley & Sons Inc.

Zou, K. H., Resnic, F. S., Talos, I. F., Goldberg-Zimring, D., Bhagwat, J. G., Haker, S. J., Kikinis, R., Jolesz, F. A., Ohno-Machado, L. (2005). A global goodness-of-fit test for receiver operating characteristic curve analysis via the bootstrap method. *J. Biomed. Informatics* 38:395–403.