Resampling Methods for the Area Under the ROC Curve

Andriy I. Bandos ANB61@PITT.EDU
Howard E. Rockette HERBST@PITT.EDU
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, U.S.A

David Gur GURD@UPMC.EDU
Department of Radiology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, U.S.A

Abstract

Receiver Operating Characteristic (ROC) analysis is a common tool for assessing the performance of various classification tools including biological markers, diagnostic tests, technologies or practices and statistical models. ROC analysis gained popularity in many fields including diagnostic medicine, quality control, human perception studies and machine learning. The area under the ROC curve (AUC) is widely used for assessing the discriminative ability of a single classification method, for comparing performances of several procedures and as an objective quantity in the construction of classification systems. Resampling methods such as bootstrap, jackknife and permutations are often used for statistical inferences about AUC and related indices when the alternative approaches are questionable, difficult to implement or simply unavailable. Except for the simple versions of the jackknife, these methods are often implemented approximately, i.e. based on the random set of resamples, and, hence, result in an additional sampling error while often remaining computationally burdensome. As demonstrated in our recent publications, in the case of the nonparametric estimator of the AUC these difficulties can sometimes be circumvented by the availability of closed-form solutions for the ideal (exact) quantities. Using these exact solutions we discuss the relative merits of the jackknife, permutation test and bootstrap in application to a single AUC or difference between two correlated AUCs.

1. Introduction

Many different fields are faced with the practical problems of detection of a specific condition or classification of findings – the tasks that can be collectively described as classification of the subjects into categories. The system that defines the specific manner of a classification process is termed differently depending on the field and task at hand (e.g. diagnostic marker, diagnostic system, technology or practice, predictive model, etc.). In this manuscript we will use the terms classification system or tool to refer to such a system regardless of the field and the task.

Since the ultimate goal is an application of the classification system to subjects from the general "target" population the performance in the target population is one of the important characteristics of the classification system. Since in practice it is usually impossible to apply the classification system to the whole population it is applied to a sample of subjects from the target population. Based on such a sample the performance of the classification system in the target population can be assessed using statistical methods.

For classification problems, performance is typically assessed in terms of the multiple probabilities of the possible outputs conditional on the true status of subjects (for binary classification - sensitivity or true positive rate and specificity or false positive rate). Multiple probabilities are considered in order to avoid specification of the relative costs and conditioning on the true class is performed in order to eliminate a dependence on the class distribution within the sample.

Some classification systems can be supervised to produce different classification rules. Most commonly such classification systems produce a quantitative output (e.g. probability of belonging to a specific class) and a decision rule is determined by a specific threshold. Another example is an unlabelled classification tree where a decision rule is determined by a specific labeling of the terminal nodes (Ferri, Flach, & Hernandez-Orallo 2002). For such classification systems an operating mode (threshold, labeling etc.) is often chosen considering the class distribution in the target population and relative cost and benefits of the specific decisions. Because of that, when assessing the performance of the classification system using a sample from the population it is often

Key words: ROC, AUC, bootstrap, permutations, jackknife, exact variances

desirable to have a performance measure that is also independent from a specific operating mode.

For binary classification tasks (subjects are classified into the two classes), conventional ROC analysis provides a tool to assess the performance of a classification system simultaneously for all operating thresholds and independently of the class distribution in the sample and costs and benefits of various decisions. The conventional ROC analysis originated in signal detection theory and presently is a widely used tool for the evaluation of classification systems (Swets & Picket, 1982; Zhou, Obuchowski and McClish, 2002; Pepe, 2003). The keystone of ROC analysis is the ROC curve which is defined as a plot of sensitivity (true positive rate) versus 1-specificity (false positive rate) computed at different possible operating modes. It illustrates the tradeoff between the two classification rates and enables the assessment of the inherent ability of a classification system to discriminate between subjects from different classes (e.g. with and without a specific disease or abnormality). Another beneficial feature of the ROC curve is its invariance to monotone transformations of the data. For example, the ROC curve corresponding to a pair of normal distributions representing classification scores (binormal ROC) is the same as the ROC curve for any pair of distribution that is monotonically transformable to the original pair.

Because its construction requires the probabilities of various classifications conditional on the true class of the subjects, a conventional Receiver Operating Characteristic (ROC) analysis is only applicable in situations where the true class is known for all subjects. On the other hand this feature enables ROC analysis to be used for studies where a fixed number of subjects have been selected from each class separately as opposed to taking a sample from the total population. Selection of subjects from each class separately eliminates problem resulting from low frequency of a specific class (e.g. low prevalence of a specific disease) and permits more efficient study design in regard to statistical considerations.

Although the ROC curve is quite a comprehensive measure of performance, because it is a whole curve there is often a desire to obtain a simpler summary index. Thus, for summarizing the performance of a classification system, more simple indices such as the area under the ROC curve (AUC), or partial AUC are typically used. The area under the ROC curve (AUC) is a widespread measure of the overall diagnostic performance and has a practically relevant interpretation as the probability of a correct discrimination in a pair of randomly selected representatives of each class (Bamber, 1975; Hanley & McNeil, 1982). In the presence of a continuous classification score the AUC is the probability of stochastic dominance of an "abnormal' class versus "normal" class, where "abnormal" class is expected to have greater scores on average.

The AUC is used for assessing the performance of a single classification system, comparing several systems and as an objective quantity for constructing a classifier (Verrelst et al 1998; Pepe & Tompson 2000; Ferri, Flach, & Hernandez-Orallo 2002; Yan et al 2003; Pepe, 2006).

An assessment of the performance of a single or a comparison of several classification systems is often initiated by computing the AUCs from the sample selected from the target population ("sample AUC"). Since the performances in the sample might differ from that in the target population, inferences about the population performance should incorporate assessment of the sample-related uncertainty. A common approach to evaluate the sample-related uncertainty is to estimate the variance of the AUC estimator. The variance estimator can than be used to place confidence intervals, test hypothesis or plan future studies.

When comparing two classification systems, an attempt is often made to control for variability by design. Namely, the data is collected under a paired design where the same set of subjects is evaluated under different classification systems, reducing the effect of heterogeneity of the samples of subjects. On the one hand the paired design leads to correlated estimators of the AUCs, requiring specific analytic methods, but on the other hand, similar to the paired t-test, because of the completely paired structure the variance for the difference of the correlated AUCs can be obtained from the variance of a single AUC by direct substitution.

Many nonparametric estimators of the variance of a single AUC and the difference between two correlated AUCs have been proposed. The methods proposed by Bamber in 1975 (based on formula from Noether 1967) and Wieand, Gail & Hanley (1983) provide unbiased estimators of the variance of a single AUC and the covariance of two correlated AUCs correspondingly. Hence, these estimators are useful for assessing the magnitude of the variability but may provide no advantages in hypothesis testing. The estimator proposed by Hanley & McNeil (1982) explicitly depends only upon the AUC and sample size and thus enables simple estimation of the sample size for a planned study. However, this estimator is known to underestimate or overestimate variance depending on the underlying parameters (Obuchowski 1994; Hanley & Hajian-Tilaki 1997) and thus is not optimal for either variance estimation or hypothesis testing (an improved estimator of the same kind was proposed by Obuchowski in 1994). Perhaps the most widely used estimator which offers both relatively accurate estimator of the variability and leads to acceptable hypothesis testing is the estimator proposed by DeLong, DeLong and Clarke-Pearson (1988). This estimator possesses an upward bias which on the one hand results in an improved (compared to the unbiased estimator) type I error of the statistical test for equality of the AUCs when AUCs are small, but on the other hand results in loss of statistical power when AUCs are large (Bandos 2005; Bandos, Rockette & Gur 2005).

Absence of a uniformly superior method, potentially poor small-sample properties of the asymptotic procedures; complexity or unavailability of the variance formulas for generalized indices (such as for AUC extensions for clustered, repeated and multi-class data) have lead many investigators to suggest using the resampling methods such as jackknife, bootstrap and permutations in applications to the AUC and its extensions (Dorfman, Berbaum & Metz, 1992; Mossman 1995; Song, 1997; Beiden, Wagner, & Campbell, 2000; Emir et al, 2000; Rutter, 2000; Hand & Till, 2001; Nakas & Yiannoutsos 2004; Bandos, Rockette, & Gur, 2005, 2006a,b).

Because of the variety of methods for assessing variability of a single AUC estimate or comparing several AUCs it is important to know their relative advantages and limitations. Previously we developed a permutation test for comparing AUCs with paired data, constructed a precise approximation based on the closed-form solution for the exact permutation variance and investigated its properties relative to the conventional approach (Bandos et al 2005). The closed-form solutions for the exact (ideal) resampling variances that we derived in that as well as in our other works permit a better understanding of the relationships and relative advantages of resampling procedures and other methods for the assessment of AUCs (Bandos 2005; Bandos et al. 2006b). In this paper we discuss the relative merits of the jackknife, bootstrap and permutation procedures applied to a single AUC or difference between two correlated AUCs.

2. Preliminaries

We assume that the true class ("normal" or "abnormal") is uniquely determined and known for each subject. Hence, according to the true status, every subject in the population can be classified as normal or abnormal. We term the ordinal output of the classification as the subject's classification score and denote $\boldsymbol{x}$ and $\boldsymbol{y}$ as scores for normal and abnormal subjects correspondingly. Furthermore, without loss of generality, we will assume that higher values of the scores are associated with higher probabilities of the presence of "abnormality".

The general layout of the data we consider consists of scores assigned to samples of N "normal" and M "abnormal" subjects by each of the classification systems. We enumerate subjects with subscripts i, k (for normal); j, l (for abnormal). Thus, x_i, y_j denote the classification scores assigned to the i^{th} "normal" and j^{th} "abnormal" subjects. When operating with more than one classification system we distinguish between them with the superscript m (e.g. x_i^m). However, when the discussion concerns primarily a single-system setting we omit the corresponding index for the sake of simplicity.

Using the conventions defined above, the nonparametric estimator of the AUC or "sample AUC" (equivalent to the Wilxocon-Mann-Whitney statistic) can be written as:

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

where the *order indicator,* $\boldsymbol{\psi}$, is defined as follows:

$$\psi_{ij} = \psi(x_i, y_j) = \begin{cases} 1 & x_i < y_j \\ \frac{1}{2} & x_i = y_j \\ 0 & x_i > y_j \end{cases} \qquad (2)$$

Also, 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 doted index.

Under a paired design, the difference in AUCs can be written as:

$$\hat{A}^1 - \hat{A}^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\psi(x_i^1, y_j^1) - \psi(x_i^2, y_j^2)\right]}{NM} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}}{NM} = \overline{w}_{\bullet\bullet} \qquad (3)$$

where

$$w_{ij} = w(x_i, y_j) = \psi(x_i^1, y_j^1) - \psi(x_i^2, y_j^2) = \psi_{ij}^1 - \psi_{ij}^2 \qquad (4)$$

This representation illustrates that the difference in areas under a paired design 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 AUC difference simply by replacing ψ_{ij} with w_{ij}.

3. Resampling approaches

Resampling approaches such as jackknife, bootstrap, permutations and combination thereof are widely used whenever conventional solutions are questionable, difficult to derive or unavailable. Major advantages of these methods include offering reliable statistical inferences in small sample problems and circumventing the difficulties of deriving the statistical moments of complex summary statistics.

3.1 Jackknife

Jackknife is a simple resampling approach that is often attributed to Quenouille (1949) and Tukey (1958). Many different varieties of the jackknife can be implemented in practice. The performance of several of them in hypothesis testing about AUC was considered by Song (1997). Although often forgotten, the variance estimators used in the procedure proposed by the DeLong et al. (1989) is also a jackknife variance estimator for the two-sample U-statistics (Arvesen, 1969). This procedure, which we will often term as "two-sample jackknife", is perhaps the most commonly used nonparametric method for comparing several correlated AUCs. In a more complex multi-reader setting a conventional "one-sample" jackknife was employed by Dorfman, Berbaum & Metz (1992) within an ANOVA framework.

The general idea of the jackknife is to generate multiple samples from the single original one by eliminating a fixed number of observations. The jackknife samples are then used as a base for calculation of the pseudo-values of a summary statistic, that are later used for inferential purposes. Since the nonparametric estimator of the AUC is an unbiased statistic, the one-sample and two-sample jackknife estimator (averages of the pseudovalues) are equal to the original one. Thus, the difference in these jackknife approaches occurs in the variances. A one-sample jackknife computes the variability of the pseudovalues regardless of the class of the eliminated subject while the two-sample jackknife computes a stratified variance. Both variances can be expressed in a closed-form and thus permit an easy comparison of these (Bandos 2005). Namely, the two-sample jackknife variance for the AUC (DeLong et al) can be written as:

$$V_{J2}(A)=\frac{\sum_{i=1}^{N}\left(\overline{\psi}_{i\bullet}-\overline{\psi}_{\bullet\bullet}\right)^2}{N(N-1)}+\frac{\sum_{j=1}^{M}\left(\overline{\psi}_{\bullet j}-\overline{\psi}_{\bullet\bullet}\right)^2}{M(M-1)} \qquad (5)$$

A one-sample jackknife variance has the following form:

$$V_{J1}(A)=\left[\frac{\sum_{i=1}^{N}\left(\overline{\psi}_{i\bullet}-\overline{\psi}_{\bullet\bullet}\right)^2}{(N-1)^2}+\frac{\sum_{j=1}^{M}\left(\overline{\psi}_{\bullet j}-\overline{\psi}_{\bullet\bullet}\right)^2}{(M-1)^2}\right]\times\frac{N+M-1}{N+M} \qquad (6)$$

A straightforward comparison of formulas (5) and (6) reveals that a one-sample jackknife variance is always larger than the two-sample one. This fact limits the usefulness of a one-sample variance since the two-sample jackknife variance is already greater than the Bamber-Wieand unbiased estimator and thus has an upward bias (Bandos 2005).

Although the jackknife approach is straightforward to implement and possesses good asymptotic properties, it is generally considered to be inferior compared to more advanced resampling techniques such as bootstrap. In application to the difference between AUCs the bootstrap variance estimator was also found to have lower mean squared error than the jackknife (Bandos, 2005). However, under certain conditions the jackknife can be considered as a linear approximation to the bootstrap (Efron & Tibshirani, 1993) and for some problems the jackknife might result in a statistical procedure that is practically indifferent from the bootstrap-based one.

3.2 Bootstrap

A good summary of the general bootstrap methodology can be found in the book by Efron & Tibshirani (1993). In ROC analysis bootstrap is commonly used for estimation of variability or for construction of confidence intervals. In recent years it has gained increased popularity in connection with its ability to obtain insight into the components of the variability of the indices estimated in multi-reader data (Beiden, Wagner & Campbell, 2000). The bootstrap was also proposed to be used for estimation of the variance of the partial AUC (Dodd & Pepe, 2003b), variance of the AUC computed from patient-clustered (Rutter, 2000) and repeated measures data (Emir et al., 2000).

The concept of the bootstrap is to build a model for the population sample space from the resamples (with replacement) of the original data. The nonparametric bootstrap completes the formation of the bootstrap sample space by assigning equal probability to all bootstrap samples. Next, a value of the summary statistic (called its bootstrap value) is calculated from every bootstrap sample and the set of all bootstrap values determines a bootstrap distribution. Such a bootstrap distribution of the summary statistic is a nonparametric maximum likelihood estimator of the distribution of the statistic computed on a sample randomly selected from a target population and serves as the basis for the bootstrap estimators of distributional parameters.

Since, even for a moderately sized problem, it may not be computationally feasible to draw all possible bootstrap samples, the conventional approach is to approximate the bootstrap distribution by computing the bootstrap values corresponding to a random sample of the bootstrap samples. Such a procedure is often called Monte Carlo or approximate bootstrap and the quantities computed from an approximate bootstrap distribution are called Monte Carlo bootstrap estimators in contrast to the quantities of the exact bootstrap distribution which are called ideal bootstrap estimators. The Monte Carlo bootstrap might still be computationally burdensome and also leads to an additional sampling error in the resulting estimators.

Some summary statistics permit circumventing the drawbacks of the Monte Carlo approach by allowing computation of ideal (exact) bootstrap quantities directly from the data. Unfortunately, the exact bootstrap variance is rarely obtainable except for simple statistics such as the sample mean. Some other estimators for which the exact bootstrap moments have been derived include sample median (Maritz & Jarret, 1978) and L-estimators (Hutson & Ernst, 2000).

In our recent work (Bandos 2005; Bandos, Rockette & Gur, 2006b) we have shown that the nonparametric estimator of the AUC permits the derivation of the analytical expression for the ideal bootstrap variance for several commonly used data structures (the bootstrap expectation of the AUC is equal to the original estimate). These results not only eliminate the need of the Monte Carlo approximation to the bootstrap of the AUC in existing methods, but can also be extended to the bootstrap applications for the patient-clustered data, repeated measure data, partial areas and potentially to a multi-class AUC extension (Hand & Till, 2001; Nakas & Yiannoutsos, 2004). For the single AUC the exact bootstrap variance has the following form:

$$V_B(A) = \frac{\sum_{i=1}^{N}\left(\overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{\psi}_{\bullet j} - \overline{\psi}_{\bullet\bullet}\right)^2}{M^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\psi_{ij} - \overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet j} + \overline{\psi}_{\bullet\bullet}\right)^2}{N^2 M^2} \quad (7)$$

Unfortunately, there is no uniform relationship between the bootstrap variance and that of any of the considered jackknife variances. The Monte Carlo investigations indicate that the bootstrap variance has uniformly smaller mean squared error. It also has a smaller bias except for very large AUC. Thus, the bootstrap often provides a better estimator of the variability than the jackknife. However, the estimator of Bamber (1975) and Wieand et al. (1983), because of its unbiasedness, might be preferred by some investigators.

Although the nonparametric bootstrap is a powerful approach that produces nonparametric maximum likelihood estimators, it is not uniformly the best resampling technique. Davison & Hinkley (1997) indicate that for hierarchical data a combination of resampling with and without replacement may better reflect the correlation structure in the general population. Furthermore, although the bootstrap can be implemented for a broad range of problems, in situations where there is something to permute (e.g. single index hypothesis testing, comparison of several indices) the permutation approach may be preferable because of the exact nature of the inferences (Efron & Tibshirani, 1993).

3.3 Permutations

Permutation procedures are usually associated with the early works of Fisher (1935). In ROC analysis permutation tests have been employed for comparison of the diagnostic modalities (Venkatraman & Begg, 1996; Venkatraman 2000; Bandos, Rockette & Gur, 2005).

Permutation based procedures are resampling procedures that are specific to hypothesis testing. Similar to the bootstrap, a permutation procedure constructs a permutation sample space, which consists of the equally likely permutation samples. The permutation samples are created by interchanging the units of the data that are assumed to be “exchangeable” under the null hypothesis. However, unlike the bootstrap sample space, the permutation sample space is the exact probability space of the possible arrangements of the data under the null hypothesis given the original sample.

The same permutation scheme can be used with different summary statistics resulting in different statistical tests. The choice of the summary statistic determines the alternatives that are more likely to be detected, but may not affect the null hypothesis. In this respect, permutation tests are similar to the tests of trend which, still assuming overall equality under the null hypothesis, aim to detect specific alternatives in the complementary hypothesis, e.g. a specific trend (linear, quadratic).

For example, when two diagnostic systems are to be compared with paired data, the natural permutation scheme consists of exchanging the paired units. Several reasonable permutation tests are possible under such a permutation scheme. One of these was developed by Venkatraman & Begg (1996) for detecting any differences between two ROC curves. For this purpose the authors used a measure specifically designed to detect the differences at every operating point. In our recent work (Bandos, Rockette & Gur, 2005) on a test that is especially sensitive to the difference in overall diagnostic performance we used the differences in nonparametric AUCs as a summary measure. Both of these tests assume the same condition of exchangeability of the diagnostic results under the null hypothesis, but differ with respect to their sensitivity to specific alternatives and the availability of an asymptotic version. Namely our permutation test better detects different ROC curves if they differ with respect to the AUC, and it has an easy-to-implement and precise approximation which is unavailable for the test of Venkatraman & Begg.

The availability of the asymptotic approximation to the permutation test can be an important issue since the exact permutation tests are practically impossible to implement with even moderate sample sizes and the Monte Carlo approximation to the permutation test is associated with a sampling error. Fortunately, in some cases the asymptotic approximation can be constructed by appealing to the asymptotic normality of the summary statistic and using the estimator of its variance, if the latter is derivable. For the nonparametric estimator of the difference in the AUC we demonstrated (Bandos, Rockette & Gur, 2005) that the exact permutation variance can be calculated directly without actually permuting the data, i.e.:

$$V_\Omega\left(A^1 - A^2\right) = \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet}^{1\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j}^{\bullet 1}\right)^2}{M^2} \quad (8)$$

where

$$w_{ij}^{p,q} = \psi\left(x_i^p, y_j^q\right) - \psi\left(x_i^{3-p}, y_j^{3-q}\right)$$

denotes the difference in the order indicators computed over the scores combined over the two systems.

The availability of an analytical expression for the exact permutation variance not only permits constructing an easy-to-compute approximation, but also makes such an approximation very precise even with small samples. Because of the restriction to the null hypothesis, the permutation variance is not directly comparable to previously mention estimation methods which provide estimators of the variance regardless of the magnitude of the difference. However, the properties of the statistical tests can be compared directly with Monte Carlo and the availability of the closed-form solution for the permutation variance greatly alleviates the computational burden of this task. The comparison of the asymptotic permutation test with the widely used procedure of DeLong et al. indicate the advantages of the former for the range of parameters common in diagnostic imaging , i.e. AUC greater than 0.8 and correlation between scores greater than 0.4 (Bandos et al., 2005).

4. Discussion

In this paper we discussed the relative merits of basic resampling approaches and outline some recent developments in the resampling-based procedures focused on the area under the ROC curve. The major drawbacks of the advanced resampling procedures are computational burden and sampling error. Sampling error results from the application of the Monte Carlo approximation to the resampling process, and adds to the uncertainty of the obtained results. Although alleviated by the development of faster computers the computational burden can still be substantial especially in the case of iteratively obtained estimators such as m.l.e. of AUC (Dorfmann & Alf 1969; Metz, Herman & Shen 1998) or when assessing the uncertainty of the resampling-based estimators (e.g. jackknife- or bootstrap-after-bootstrap). In our previous works we showed that for the nonparametric estimator of the AUC presented here all of the considered resampling procedures permit derivation of the ideal variances directly avoiding implementation of the resampling process or its approximation. Such closed-form solutions greatly reduce computational burden, eliminate a sampling error associated with the Monte Carlo approximation to the resampling variances, permit construction of precise approximations to the exact methods and facilitate assessment and comparison of the properties of various statistical procedures based on resampling.

In general jackknife provides a somewhat simplistic method that, depending on the problem, may still offer valuable solutions. In application to estimation of the nonparametric AUC, the two-sample jackknife is preferable over the one-sample due to a smaller upward bias. Bootstrap is a more elaborate resampling procedure that provides nonparametric maximum likelihood estimators by offering an approximation to the population sample space. Bootstrap is usually preferred over the jackknife because of cleaner interpretation and sometimes better precision. Exploiting a formula for the exact bootstrap variance of the AUC we demonstrated that it provides an estimator of the variance that is more accurate in terms of the mean squared error than the two-sample jackknife variance and is often more efficient than the unbiased estimator. In the case of comparing two AUCs the asymptotic tests based on the bootstrap and jackknife variances have very similar characteristics. However, for more complex problems the bootstrap may perform better than the jackknife. The permutations explore the properties of the population sample space assuming the exchangeability satisfied under the null hypotheses. For the comparison of the performances under a paired design the permutation test can be considered as preferable over the bootstrap and jackknife due to the exact nature of the permutation inferences. The availability of the exact permutation variance permits construction of an easy-to–implement and precise approximation and facilitates investigation of the properties of the permutation test. Compared to the two-sample jackknife asymptotic test for comparing two correlated AUCs, the asymptotic permutation test was shown to have greater statistical power for the range of parameter common in diagnostic radiology.

Although this paper focuses on the most commonly used summary index, AUC, the availability of the analytical expression for the exact variances is not limited to this relatively simple case. Formulas for ideal variances may also appear derivable for other AUC related indices and for different types of data (multi-reader, clustered, repeated measures and multi-class data) as well as under other, more complex, resampling schemes or study designs.

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

Arvesen, J.N. (1969). Jackknifing U-statistics. *Annals of Mathematical Statistics* 40(6), 2076-2100.

Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 12, 387-415.

Bandos, A. (2005). Nonparametric methods in comparing two ROC curves. Doctoral dissertation, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh. (http://etd.library.pitt.edu /ETD/available/etd-07292005-012632/)

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. *Statistics in Medicine* 24(18), 2873-2893.

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

Bandos, A.I., Rockette, H.E., Gur, D. (2006b). Components of the bootstrap variance of the areas under the ROC curve. *IBS ENAR* 2006, Tampa, FL.

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 Radiology* 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. *Journal of the American Statistical Association* 98, 409-417.

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

Dorfman, D.D., Alf JrE. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology* 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 Radiology* 27, 723-731.

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*. 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. *Statistics in Medicine* 19, 511-523.

Ferri, C., Flach, P., Hernandez-Orallo, J. (2002). Learning decision trees using area under the ROC curve. *Proceedings of ICML-2002*.

Fisher, R.A. (1935). *Design of experiments*. Oliver and Boyd, Edinburgh

Hand, D.J., Till, R.J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning* 45, 171-186.

Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology* 143, 29-36.

Hanley, J.A., Hajian-Tilaki, K.O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Academic Radiology* 4, 49-58.

Hutson, A.D., Ernst, M.D. (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 62(1), 89-94.

Maritz, J.S., Jarrett, R.G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association* 73(361), 194-196.

Metz, C.E., Herman, B.A., Shen, J. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* 17, 1033-1053.

Mossman, D. (1995). Resampling techniques in the analysis of non-binormal ROC data. Medical decision making 15, 358-366.

Nakas, C.T., Yiannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* 23, 3437-3449.

Noether GE. Elements on Nonparametric Statistics. *Wiley & Sons Inc.:* New York 1967.

Obuchowski, N.A. (1994). Computing sample size for receiver operating characteristics studies. *Investigative Radiology* 29, 238-243.

Obuchowski, N.A. (1997). Nonparametric analysis of clustered ROC curve data. *Biometrics* 53, 567-578.

Pepe, M.S., Thompson, M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* 1, 123-140.

Pepe, M.S. (2003). *The statistical evaluation of medical test for classification and prediction*. Oxford: Oxford University Press.

Pepe, M.S., Cai, T., Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 62, 221-229.

Quenouille, M.H (1949). Approximate tests of correlation in time series. Journal of Royal Statistical Society, Series B 11, 18-84.

Rutter, C.M. (2000). Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Academic Radiology* 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.

Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics* 29, 614.

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.

Venkatraman, E.S. (2000) A permutation test to compare receiver operating characteristic curves. *Biometrics* 56, 1134-1136.

Verrelst, H., Moreau, Y., Vandewalle, J., Timmerman, D. (1998). Use a multi-layer perceptron to predict malignancy in ovarian tumors. *Advances in Neural Information Processing Systems*, 10.

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. Bulletin* 12, 213-214.

Yan, L., Dodier, R., Mozer, M.C., Wolniewicz, R. (2003). Optimizing Classifier performance via an approximation to the Wilcoxon-Mann-Whitney Statistic. *Proceedings of ICML-2003*.

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