NONPARAMETRIC METHODS IN COMPARING TWO CORRELATED ROC CURVES

by

Andriy Bandos

M.S., Kharkiv National University, 2000

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Andriy Bandos

It was defended on

July 25, 2005

and approved by:

Stewart Anderson, PhD
Associate Professor
Department of Biostatistics,
Graduate School of Public Health
University of Pittsburgh

Vincent C. Arena, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

David Gur, ScD
Professor
Department of Radiology
School of Medicine
University of Pittsburgh

Dissertation Director: Howard E. Rockette, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

NONPARAMETRIC METHODS IN COMPARING TWO CORRELATED ROC CURVES

Andriy Bandos, PhD

University of Pittsburgh, 2005

Receiver Operating Characteristic (ROC) analysis is one of the most widely used methods for summarizing intrinsic properties of a diagnostic system, and is often used in evaluation and comparison of diagnostic technologies, practices or systems. These methods play an important role in public health since they enable researchers to achieve a greater insight into the properties of diagnostic tests and eventually to identify a more appropriate and beneficial procedure for diagnosing or screening for a specific disease or condition. The topic of this dissertation is the nonparametric testing of hypotheses about ROC curves in a paired design setting. Presently only a few nonparametric tests are available for the task of comparing two correlated ROC curves. Thus we focus on this basic problem leaving the extensions to more complex settings for future research. In this work, we study the small-sample properties of the conventional nonparametric method presented by DeLong *et al.* and develop three novel nonparametric approaches for comparing diagnostic systems using the area under the ROC curve. The permutation approach that we present enables conducting an exact test and allows for an easy-to-use asymptotic approximation. Next, we derive a closed-form bootstrap-variance, construct an asymptotic test, and compare them to the existing competitors. Finally, exploiting the idea of "discordances" we develop a conceptually new conditional approach that offers advantages in certain types of studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my gratitude to all of my supervisors who directed and guided me through the whole process of my study and research. I am greatly thankful to the members of the committee for their help and support in preparation of the thesis. I also want to express my sincere appreciation of the efforts of all the professors whose courses I have been taken for they helped to create a solid foundation for my current and future research. Finally I want to thank my family and friends for their encouragement, love and support.

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

I. INTRODUCTION

The performance of a diagnostic system is frequently characterized by its ability to discriminate between subjects with and without an abnormality of interest. A Receiver Operating Characteristic (ROC) curve is one of the most commonly used methods for summarizing the intrinsic discriminative abilities of a diagnostic system. Additionally ROC analysis is often used in the evaluation and comparison of diagnostic technologies, practices or systems (often termed as *modalities*) [1,2,3,4,5,6].

As a simplification to considering the entire curve, a variety of summary indices have been proposed [1,2,3,4,7,8]. One of the most common measures used for summarizing the overall performance of diagnostic modalities is the Area Under the ROC Curve (AUC). The AUC measure is conveniently interpretable as the probability of correct discrimination between "abnormal" (with the condition) and "normal" (without the condition) subjects [1,2,13]. The AUC as well as other indices derived from the ROC curve can be estimated using both parametric [2,4,10,11] and nonparametric [13,14,16,20,19] approaches.

The comparison of diagnostic systems is often performed by comparing various ROC indices. To control for additional sources of variability a paired design, in which a selected population of subjects is evaluated by both modalities being compared, is often implemented. This type of design, however, leads to correlated estimates which then require an appropriate analysis. A number of parametric, semi-parametric, and completely nonparametric approaches have been developed to compare diagnostic modalities under a paired design [17,18,19,20,21,23]. The relative benefit of a paired design compared to an unpaired design depends on the correlation between the observations for the two modalities being compared. In a review of a large number of experimental studies to compare different imaging modalities Rockette *et al.* [27] found that the average correlation between two modalities in paired experiments ranged from 0.35 to 0.59 depending on the specific abnormality in question.

Appropriate use of a paired design also requires that the experimenter has adequately controlled in the design for the effects of order of the administration of the two diagnostic systems being compared. Finally, if the number of normal and abnormal cases is fixed, as we have assumed here, then careful attention must be given to the purpose of the study and potential biases that might result due to the selection process.

A. OBJECTIVES

The primary purpose of this dissertation is to improve upon existing methods of comparing two ROC curves in a paired design setting. Although the approaches we develop appear to be extendable to analysis of more general problems such as comparing more than two modalities, using multiple readers [27,35,38,37] or comparing partial areas [20,28] we consider these more complex problems to be beyond the scope of this dissertation.

1. Properties of the conventional nonparametric AUC test

The test proposed by DeLong *et al.* [19] is the conventional nonparametric procedure for comparing correlated AUCs. It uses a consistent variance estimator and relies on asymptotic normality of the AUC estimator. Although it is generally recognized that convergence to the asymptotic properties depends on the underlying parameters, and several Monte Carlo studies include the conventional procedure in their investigation [38,39,40], there have not been extensive simulations characterizing the effects of relevant parameters on the small-sample properties of the this procedure.

We study the behavior of the type I error and the statistical power of the conventional nonparametric test for comparing two AUCs over a wide range of relevant parameters and against various alternatives. These investigations provide useful information in regard to how and to what extent various underlying parameters affect small-sample statistical inferences. Part of the results of this investigation was presented at the Medical Image Perception Society conference X [31].

2. A permutation test for comparing diagnostic modalities

Using the permutation scheme previously employed in the paper by Venkatraman and Begg [24] we construct a permutation test for detecting differences between two AUCs in a paired design setting. Such a permutation procedure not only provides an exact (suitable for small samples) and powerful test for detecting differences in overall performances but also permits developing a precise and easy-to-apply approximation. The availability of a simple and precise approximation to the permutation test is a desirable property since with increasing sample size exact permutation tests quickly become very demanding computationally. The properties of the nonparametric AUC estimator permit the derivation of the exact variance in the permutation space and therefore facilitate the development of a precise approximation. We also conduct simulations to investigate properties of the new procedure. This part of the dissertation was accepted for publication in Statistics in Medicine [32].

3. Bootstrap-variance, asymptotic test and their properties

The bootstrap is a powerful nonparametric approach [41] and the ideas of exploiting the bootstrap procedure in ROC analysis have been previously proposed [39,37,43]. Unfortunately, the intensity of the computations required to create all bootstrap-samples or an additional error associated with incomplete sampling of the bootstrap-space reduce the attractiveness of the approach.

The conventional procedure for comparing correlated AUCs developed by DeLong *et al.* [19] is equivalent to the two-sample jackknife procedure [22]. Since the bootstrap approach is usually considered to be superior to the jackknife [42], it is reasonable to investigate the properties of the asymptotic bootstrap test compared to the conventional test. For a specific statistic such as the nonparametric estimator of the AUC, the closed-form bootstrap-variance can be derived allowing one to construct an easy-to-compute asymptotic test. We compare the properties of the variance estimators and the corresponding asymptotic procedures based on jackknife and bootstrap approaches using computer simulations.

4. Conditioning on discordances between two diagnostic modalities

When comparing the AUCs in a paired design setting, each pair of normal and abnormal cases can be classified based on whether the two modalities agreed in regard to the relative orderings between normal and abnormal subjects' ratings (concordant) or whether the two modalities had different relative orderings for the normal-abnormal pair (discordant). While the orderings of ratings that are the same in both modalities ("concordant" orderings) are important for assessing performance of each modality separately, these could mask the true difference between two modalities in a paired design. The orderings that differ in two modalities ("discordant" orderings) on contrary contain information about the discrepancies between the performances of diagnostic systems.

We develop a novel approach for statistical comparison of the overall performance of the two modalities in a paired design setting. The difference between the overall performances of two modalities is assessed by the fraction of the discordant orderings observed in favor of one of them. The corresponding statistical test is similar in spirit to McNemar's procedure [44] which conducts the analysis only on discordant pairs. Simulations are conducted to verify the small-samples properties of the conditional test. This portion of the research is published in Academic Radiology [33].

II. ROC METHODOLOGY

Many statistical problems address binary outcomes that are associated with an ordinal variable. ROC curves represent one of the most popular and powerful tools in the analysis of the relationship between two such variables.

Although ROC analysis is applicable to a variety of disciplines one of its most common uses is in the area of diagnostic test evaluation. In this field, the binary outcome usually indicates presence or absence of a specific abnormality where the status is determined based on an accepted "gold standard". The ordinal variable associated with such a binary outcome can represent a continuous measure based on a quantitative clinical test or the confidence of a rater in the subject's abnormality based on the result of a diagnostic test.

A. CONVENTIONS AND DEFINITIONS

We will treat the binary outcome as the indicator of presence or absence of an abnormality (sometimes called "true status") and assume that it is uniquely determined and known for each subject. Hence the population of subjects can be divided into normal and abnormal subpopulations according to the true status of each subject. We will designate the ordinal variable related to the presence of abnormality as the rating of the subject and denote $\boldsymbol{X}$ and $\boldsymbol{Y}$ as ratings for normal and abnormal subjects correspondingly. Furthermore, without loss of generality, we assume that higher values of ratings are associated with higher probabilities of the presence of abnormality.

For any real-valued threshold, $\boldsymbol{c}$, the population of subjects can be classified into the two groups according to their ratings being greater or less than $\boldsymbol{c}$. If a diagnostic procedure is reasonable then the group with higher ratings will include proportionally more abnormal than

normal subjects. The agreement between the classification obtained and the real status of the subjects can be characterized using two quantities: *sensitivity* (*True Positive Fraction*) and *specificity* (*True Negative Fraction*) defined as follows:

$$sens(c) = TPF(c) = P(Y > c)$$

$$spec(c) = TNF(c) = P(X \leq c).$$

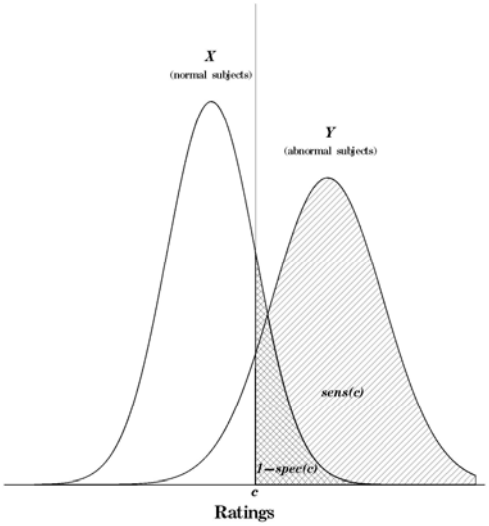

Figure II.1 **Distribution of ratings**

The Receiver Operating Characteristic (ROC) approach allows considering the agreement between ratings and the presence of abnormality for all thresholds simultaneously. The ROC curve is the plot of *sensitivity* versus *1-specificity* where each point on the graph corresponds to a specific threshold ***c***. (See Figure II.2)

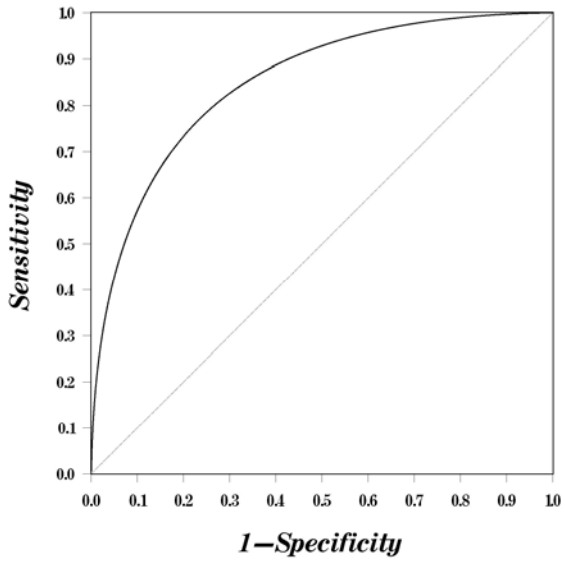

Figure II.2 **"Binormal" ROC curve**

Note that for every distribution of ratings in the groups of normal and abnormal subjects there is a unique ROC curve. However, a single ROC curve corresponds to an infinite class of bivariate distributions any two of which are monotonically transformable to each other. In other words, an ROC curve is invariant to any monotone transformation of the underlying bivariate distribution.

An ROC curve is useful in many tasks related to accuracy of diagnostic tests such as selecting an optimal threshold for a diagnostic procedure or in determining which diagnostic procedure is better on average or at a particular operating point [1,2,3,4]. Although the ROC curve is employed in describing a diagnostic test there is frequently a desire to have a simple summary index. In diagnostic radiology as well as in many other fields one of the most useful measures derived from the ROC curve is the area under the ROC curve (AUC).

The AUC index reflects the inherent discriminative ability of a diagnostic procedure and has a nice interpretation of the probability of correct discrimination between randomly chosen normal and abnormal subjects [13]. To illustrate this concept consider the 2-Alternative Forced Choice (2AFC) experiment in which for a pair of normal and abnormal subjects the "rater" has to select the abnormal subject. The probability of a correct decision in a 2AFC experiment equals the AUC of the diagnostic procedure. In the presence of an ordinal variable (rating) representing

the confidence of abnormality the selection is guided by the value of the rating. Namely if ratings of two subjects do differ then the subject with greater rating is declared abnormal; otherwise when both have equal ratings, any of two subjects can be diagnosed as abnormal with equal probability. Hence the area under the ROC curve can be expressed in the following way:

$$A = P(X < Y) + \tfrac{1}{2} P(X = Y)$$

As the formula indicates, the AUC can be interpreted as the probability that a randomly selected abnormal subject has greater rating than a randomly selected normal subject plus half of the probability of equality of the ratings for the pair of subjects.

Throughout this work we will denote $\{x_i^r\}_{i=1}^N$ and $\{y_j^r\}_{j=1}^M$ as the ratings observed for the $\boldsymbol{N}$ normal and $\boldsymbol{M}$ abnormal subjects in the $\boldsymbol{r}^{th}$ modality ($\boldsymbol{r=1,..,K}$). In these notations the unbiased estimator of AUC in the $\boldsymbol{r}^{th}$ modality is given by:

(II.A.1) $$\hat{A}^r = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi(x_i^r, y_j^r)}{NM} = \overline{\psi}_{\bullet\bullet} \text{ where, } \psi(x,y) = \begin{cases} 1 & x < y \\ \frac{1}{2} & x = y \\ 0 & x > y \end{cases} .$$

In the completely paired design the difference between two nonparametric AUC estimators derived from the same cases can be written in a similar way to a single AUC, namely:

(II.A.2) $$\hat{A}^1 - \hat{A}^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi(x_i^1, y_j^1)}{NM} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\psi(x_i^2, y_j^2)}{NM} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}}{NM} = \overline{w}_{\bullet\bullet}$$

where $\boldsymbol{w_{ij}}$ is a *"joint order indicator"* and is defined as:

(II.A.3) $$w_{ij} = w(x_i, y_j) = \psi(x_i^1, y_j^1) - \psi(x_i^2, y_j^2) = \psi_{ij}^1 - \psi_{ij}^2 =$$

$$= \begin{cases} 1 & x_i^1 < y_j^1, x_i^2 > y_j^2 \\ \frac{1}{2} & x_i^1 < y_j^1, x_i^2 = y_j^2 \quad or \quad x_i^1 = y_j^1, x_i^2 > y_j^2 \\ 0 & x_i^1 < y_j^1, x_i^2 < y_j^2 \quad or \quad x_i^1 > y_j^1, x_i^2 > y_j^2 \quad or \quad x_i^1 = y_j^1, x_i^2 = y_j^2 \\ -\frac{1}{2} & x_i^1 > y_j^1, x_i^2 = y_j^2 \quad or \quad x_i^1 = y_j^1, x_i^2 < y_j^2 \\ -1 & x_i^1 > y_j^1, x_i^2 < y_j^2 \end{cases}$$

B. METHODS OF ANALYSIS

1. General

Several different methods have been developed for the analysis of ROC curves. The parametric methods usually model the ROC curves by assuming a particular underlying distribution of subject ratings (usually assuming that a bivariate distribution of ratings is transformable to a binormal). The "binormal" ROC curves were shown to be quite robust for a wide class of curves encountered in practice [9], a property that is in part due to variety of distributions that can be approximated by a monotone transformation of a binormal distribution. One of the best known parametric approaches to the analysis of the ROC curves is the maximum likelihood approach introduced by Dorfman and Alf Jr. [10]. C. Metz *et al.* have developed computer software ROCKIT that implements the original [10] and a modified [11] maximum likelihood estimation approaches. The software permits the analysis of two ROC curve in the presence of categorical or continuous ratings data.

Nonparametric methods utilize empirical ROC points by connecting them with straight lines, step functions or sometimes by fitting a smooth curve [1,6,12,13,14]. The main advantage of nonparametric methods compared to parametric ones is the absence of specific assumptions about the shape of the curve or the underlying distribution of ratings. Furthermore, unlike many parametric procedures, iterative algorithms are not needed for implementation of most nonparametric methods. A wide family of nonparametric statistics is described by Wieand *et al.* [20].

2. AUC index

As previously mentioned, one of the most popular and convenient indices is the Area Under the ROC Curve (AUC). The nonparametric estimate of the AUC is easy to compute and its numerical value is equal to the actual area under the estimated ROC curve where empirical points are connected by straight lines [13]. If $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{M}$ are the ratings observed for the samples of $\boldsymbol{N}$ normal and $\boldsymbol{M}$ abnormal subjects then the estimate of the AUC is given in (II.A.1)

The nonparametric AUC estimator as presented in (II.A.1) is a generalized U-statistic and therefore is approximately normally distributed under quite general assumptions [26]. Hence, knowing the variance of the estimator is essential for constructing simple asymptotic procedures. The nonparametric AUC estimator is related to the Mann-Whitney two-sample test statistic [16,13] and many of the nonparametric approaches to variance derivation are related to the formula derived by Noether [15] for the Wilcoxon statistic. Using previously introduced notation, the formula of Noether when applied to the AUC can be written as follows:

(II.B.2.1)
$$Var(\hat{A}) = \frac{N-1}{NM}\xi_{10} + \frac{M-1}{NM}\xi_{01} + \frac{1}{NM}\xi_{11},$$
$$A = E\left[\psi(X_i, Y_j)\right] = E\left[\hat{A}\right]$$
$$\xi_{10} = Cov\left[\psi(X_i, Y_j), \psi(X_i, Y_l)\right] = E\left[\psi(X_i, Y_j) \times \psi(X_i, Y_l)\right] - A^2, \quad j \neq l$$
$$\xi_{01} = Cov\left[\psi(X_i, Y_j), \psi(X_k, Y_j)\right] = E\left[\psi(X_i, Y_j) \times \psi(X_k, Y_j)\right] - A^2, \quad i \neq k \text{ and}$$
$$\xi_{11} = Var\left[\psi(X_i, Y_j)\right] = E\left[\psi(X_i, Y_j)^2\right] - A^2$$

Bamber [16] proposed an unbiased variance estimator that is based on expressing unknown expectations using probabilities which can be estimated by proportions. Hanley and McNeil [17] used the parametric assumption to estimate certain variance elements. The consistent, completely nonparametric estimators of the covariance matrix for several nonparametric AUC estimators were developed by Wieand *et al.* [18] in 1983 and by DeLong *et al.* [19] in 1988.

The conventional variance estimator proposed by DeLong *et al.* [19] can also be shown to be equivalent to the two-sample jackknife estimator of the variance [22]. Because of the structure of the nonparametric estimator of AUC its variance estimator is easy to compute, i.e.:

a) Compute the X- and Y-components:

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

b) The components ξ_{10} and ξ_{01} are estimated as:

$$s_{10} = \frac{1}{N-1}\sum_{i=1}^{N}\left[\overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet\bullet}\right]^2, \qquad s_{01} = \frac{1}{M-1}\sum_{j=1}^{M}\left[\overline{\psi}_{\bullet j} - \overline{\psi}_{\bullet\bullet}\right]^2$$

c) The consistent estimator of the variance is:

$$V(\hat{A}) = \frac{s_{10}}{N} + \frac{s_{01}}{M} = \frac{\sum_{i=1}^{N}\left[\overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet\bullet}\right]^2}{N(N-1)} + \frac{\sum_{j=1}^{M}\left[\overline{\psi}_{\bullet j} - \overline{\psi}_{\bullet\bullet}\right]^2}{M(M-1)} \quad \text{(II.B.3.1)}$$

The estimation approach employed by Wieand *et al.* [18] when implemented for a single AUC produces the biased and unbiased estimators that are equivalent to that proposed by Bamber [16]. In our notations the unbiased estimator has the following form (both estimators are shown in the Appendix C in application to AUC difference):

$$V_W(\hat{A}) = \frac{\sum_{i=1}^{N}\left[\overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet\bullet}\right]^2}{N(N-1)} + \frac{\sum_{j=1}^{M}\left[\overline{\psi}_{\bullet j} - \overline{\psi}_{\bullet\bullet}\right]^2}{M(M-1)} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\overline{\psi}_{ij} - \overline{\psi}_{i\bullet} - \overline{\psi}_{\bullet j} + \overline{\psi}_{\bullet\bullet}\right]^2}{NM(N-1)(M-1)} \quad \text{(II.B.3.2)}$$

3. Comparing diagnostic modalities

To simplify the discussion we will use the term *modality* to designate a diagnostic system, practice or technology. The between subject heterogeneity is recognized as substantial in the field of diagnostic imaging as well as in many other fields. Hence, the paired design is often used to improve the precision of the analysis. In a paired design each subject is independently evaluated by all modalities and the ratings obtained in such a way are used for the analysis. The correlation between the ratings for the same subjects can be substantial [27] and should be accounted for in the analysis.

4. Comparing ROC curves in a paired design

One of the nonparametric procedures for comparing ROC curves is a permutation test developed by Venkatraman and Begg [24]. The test they proposed is designed to compare two correlated ROC curves at every operating point using the specially developed measure denoted as ***E***. The significance of the observed difference is then evaluated using the permutation space. Namely the ***E***-index is calculated for every permutation and the ***p***-value is calculated as the proportion of times when more extreme values than the ***E***-index computed from the observed data are obtained.

The ***E***-statistic is composed of so called "empirical errors" [24]. The "error" indicators are defined for each empirical operating point and for every normal and abnormal subject using ranks. Namely, if $\{x_i^r\}_{i=1}^N$ and $\{y_j^r\}_{j=1}^M$ are the ratings observed for the ***N*** normal and ***M*** abnormal subjects in the r^{th} modality and $\{rank(x_i^r)\}_{i=1}^N$ and $\{rank(y_j^r)\}_{j=1}^M$ are corresponding ranks then the "errors" indicators are defined as follows:

$$e_k(x_i) = \begin{cases} 1 & if\ rank(x_i^1) \le k\ and\ rank(x_i^2) > k \\ -1 & if\ rank(x_i^1) > k\ and\ rank(x_i^2) \le k \\ 0 & otherwise \end{cases}$$

$$e_k(y_j) = \begin{cases} 1 & if\ rank(y_j^1) > k\ and\ rank(y_j^2) \le k \\ -1 & if\ rank(y_j^1) \le k\ and\ rank(y_j^2) > k \\ 0 & otherwise \end{cases}$$

Using computed "errors" indicators, the measure of "closeness" of two ROC curves at the k^{th} operating point is computed as:

$$e_{.k} = \sum_{i=1}^{N} e_k(x_i) + \sum_{j=1}^{M} e_k(y_j)$$

Finally the ***E***-statistic which provides a measure of "closeness" over all operating points is defined as:

$$E = \sum_{k=1}^{N+M-1} |e_{.k}|$$

As was indicated previously, the significance of the observed difference between two ROC curves is assessed by the significance of the computed ***E***-statistic in the permutation space. The permutation space is created by permuting the ratings assigned to the same subjects for the different modalities. Namely, consider the vector $q^t = (q_1^t, ..., q_{N+M}^t)$ consisting of 0s and 1s. The set of all such vectors can be used to enumerate all 2^{N+M} permutations. In the t^{th} permutation of the original data the values of the ratings for each subject can be determined using the $\boldsymbol{q_t}$ vector. For instance the ratings of the i^{th} normal subject in the t^{th} permutation of the data are:

$$X_i^{1t} = q_i^t x_i^1 + (1 - q_i^t) x_i^2 \qquad X_i^{2t} = (1 - q_i^t) x_i^1 + q_i^t x_i^2$$

Since q_i^t is either 0 or 1, the vector (X_i^{1t}, X_i^{2t}) equals either (x_i^1, x_i^2) or (x_i^2, x_i^1). If all the permutations are equally likely then the values of the ***E***-statistics computed for all permutations constitute the "reference" distribution of the ***E***-statistic. The constructed permutations are equally likely under the null hypothesis of equality of the ROC curves and the additional assumption of exchangeability.

To make the procedure appropriate for comparing modalities with different underlying scales (when ratings are not directly exchangeable even under the null hypothesis), the rank transformation is suggested. If the transformation is applied then the permutations are conducted on the rank of the ratings instead of raw ratings (the ties that appeared during the process of permutation of the ranks are suggested to be uniformly broken).

Venkatraman and Begg evaluated operating characteristics of their procedure on simulated datasets. Due to the computational burden, the p-values were evaluated by sampling from a permutation distribution. They found, that compared to the nonparametric "area test" proposed by DeLong *et al.* [19], their procedure possesses more power against alternatives of crossing ROC curves with equal AUC but less power against alternatives of difference in AUCs.

5. Comparing AUCs with paired data

Both parametric and nonparametric methods for comparing correlated AUC indices are available. The parametric analysis assuming the *binormal model* was developed by Dorfman and Alf Jr. [10] and later implemented and further developed by Metz *et al.* [11]. Hanley and McNeil [17] suggested using the binormal assumption only for estimation of the covariance between two area estimators.

Wieand, Gail, James B and James K [20] described a general class of nonparametric statistics for comparison of two diagnostic markers based on a weighted average of sensitivities. Earlier, Wieand, Gail and Hanley developed a nonparametric procedure for comparing diagnostic tests with paired or unpaired data [18]. DeLong *et al.* [19] developed a consistent nonparametric estimator of the covariance matrix for several AUC estimators in a paired design. This method, which is described below, is a natural extension to ***K***-samples of the formulas given in Section II.2.

Let $\{x_i^r\}_{i=1}^N$ and $\{y_j^r\}_{j=1}^M$ be the ratings assigned by the r^{th} modality ($r=1,..,K$) to N normal and M abnormal subjects. Then the vector of the AUC estimators can be computed as a simple average of the order indicators, i.e.:

$$(\hat{A}^1,...,\hat{A}^K) = \left(\overline{\psi}_{\bullet\bullet}^1,...,\overline{\psi}_{\bullet\bullet}^K\right)$$

The covariance matrix for a vector the estimators $(\hat{A}^1,...,\hat{A}^K)$ can be computed as follows:

a) Compute the X and Y components of the r^{th} modality,

$$\overline{\psi}_{i\bullet}^r = \frac{1}{M}\sum_{j=1}^M \psi(x_i^r, y_j^r), \qquad \overline{\psi}_{\bullet j}^r = \frac{1}{N}\sum_{i=1}^N \psi(x_i^r, y_j^r)$$

b) Compute the matrices $S_{10} = \{s_{10}^{r,s}\}_{r,s=1}^K$ and $S_{01} = \{s_{01}^{r,s}\}_{r,s=1}^K$, where

$$s_{10}^{r,s} = \frac{1}{N-1}\sum_{i=1}^N \left[\overline{\psi}_{i\bullet}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{i\bullet}^s - \overline{\psi}_{\bullet\bullet}^s\right], \qquad s_{01}^{r,s} = \frac{1}{M-1}\sum_{j=1}^M \left[\overline{\psi}_{\bullet j}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{\bullet j}^s - \overline{\psi}_{\bullet\bullet}^s\right]$$

c) A consistent estimator of the covariance matrix is:

$$\hat{Cov}(\hat{A}^1,...,\hat{A}^K) = \frac{S_{10}}{N} + \frac{S_{01}}{M} \text{ the (r,s)}^{\text{th}} \text{ element of which is}$$

$$\hat{Cov}(\hat{A}^r,\hat{A}^s) = \frac{\sum_{i=1}^N \left[\overline{\psi}_{i\bullet}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{i\bullet}^s - \overline{\psi}_{\bullet\bullet}^s\right]}{N(N-1)} + \frac{\sum_{j=1}^M \left[\overline{\psi}_{\bullet j}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{\bullet j}^s - \overline{\psi}_{\bullet\bullet}^s\right]}{M(M-1)}$$

Using our notation, the unbiased estimator proposed by Wieand *et al.* [18] takes the following form:

$$\hat{Cov}(\hat{A}^r,\hat{A}^s) = \frac{\sum_{i=1}^N \left[\overline{\psi}_{i\bullet}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{i\bullet}^s - \overline{\psi}_{\bullet\bullet}^s\right]}{N(N-1)} + \frac{\sum_{j=1}^M \left[\overline{\psi}_{\bullet j}^r - \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{\bullet j}^s - \overline{\psi}_{\bullet\bullet}^s\right]}{M(M-1)} - \frac{\sum_{i=1}^N\sum_{j=1}^M \left[\overline{\psi}_{ij}^r - \overline{\psi}_{i\bullet}^r - \overline{\psi}_{\bullet j}^r + \overline{\psi}_{\bullet\bullet}^r\right]\times\left[\overline{\psi}_{ij}^s - \overline{\psi}_{i\bullet}^s - \overline{\psi}_{\bullet j}^s + \overline{\psi}_{\bullet\bullet}^s\right]}{NM(N-1)(M-1)}$$

Note that in a completely paired design, the variance of the difference between the nonparametric estimators of AUC can be found using formulas (II.B.3.1-2) but employing the difference of the order indicators $w_{ij} = \psi_{ij}^1 - \psi_{ij}^2$ (II.A.3) instead of the original indicators (Appendix C).

III. PROPERTIES OF THE CONVENTIONAL NONPARAMETRIC TEST

The conventional nonparametric test for comparing correlated AUCs proposed by DeLong *et al.* [19] uses a consistent variance estimator and relies on asymptotic normality of the AUC estimator. Although it is generally recognized that convergence to the asymptotic properties depends on the underlying parameters, and several Monte Carlo studies include the conventional procedure in their investigation [38,39,40], there have not been extensive simulations characterizing the effects of relevant parameters on the small-sample properties of the this procedure.

We study the behavior of the type I error and the statistical power of the conventional nonparametric test for comparing two AUCs over a wide range of relevant parameters and against various alternatives. These investigations provide useful information on the effect of selected underlying parameters on small-sample statistical inferences. Part of the results of this investigation was presented at the MIPS conference [31].

A. GENERAL SIMULATION DESCRIPTION

To model the ratings assigned to a sample of subjects by two diagnostic modalities we simulate the data from two correlated bivariate (normal and abnormal subjects') distributions. For our simulations we use the "binormal" ROC model because of its simplicity and robustness [9] Thus, within the r^{th} modality, subjects' ratings are generated from binormal distributions namely, $X_i^r \overset{i.i.d.}{\sim} N(\mu_X^r, \sigma_X^r)$, for the ratings of the normal subjects and $Y_j^r \overset{i.i.d.}{\sim} N(\mu_Y^r, \sigma_Y^r)$, for the ratings of the abnormal subjects. Furthermore, to model a paired data structure a correlation of magnitude, ρ, is induced for the ratings of the same subject in different modalities

($Cov(X^1, X^2) = Cov(Y^1, Y^2) = \rho$). Note that the use of the binormal distribution to model subjects' ratings provides considerable flexibility since the ROC curve and ROC techniques that we consider are invariant with respect to order-preserving transformation of the data.

The binormal ROC curve corresponding to the distribution of ratings within the r^{th} modality can be parameterized using the following quantities:

$A^r = P\left(X^r < Y^r\right)$ - the Area Under the ROC Curve, and

$b_r = \frac{\sigma_X^r}{\sigma_Y^r}$ - the shape-parameter

By varying the parameters of the distributions of the ratings we model various patterns of the correlation between the ratings of the same subjects (ρ), average of two AUCs (A), difference between two areas (Δ) and shapes of the ROC curve (b). The scenario of non-crossing ROC curves is modeled by setting b=1 for both modalities while crossing ROC curves were simulated by setting b<1 (corresponds to a greater variability among ratings of abnormal subjects) for one of the modalities. We also considered different values of the total number of normal and abnormal subjects (T=N+M) and of the proportion of subjects with an abnormality (p=M/(N+M)). For each considered scenario 10,000 datasets were simulated.

B. SIMULATION STUDY

The effects of the selected parameters on the type I error of the conventional test for comparing correlated AUCs are summarized in Figure III.1 and Table III.1. Figure III.2 and Table III.2 depict the effect of selected parameters on the statistical power of the procedure. Each figure is only able to summarize the trend in the rejection rate for two parameters and therefore the other parameters are kept fixed at what is considered reasonable values. Specifically, when the value of a parameter is not specified on the graph it is set to one of the following: sample size (T) of 80, an average AUC (A) of 0.85, a correlation between ratings (ρ) of 0.4, a shape parameter (b) of 1 in both modalities and "prevalence" of the abnormal subjects (p) of ½.

All graphs in Figure III.1 demonstrate the substantial effect of the underlying AUC on the false rejection rate of the conventional test. Namely, the type I error decreases with increasing average AUC (***A***) shifting from being slightly elevated above the nominal level to being substantially lower. Although other parameters can slightly change the rate of the relationship the general decreasing pattern remains the same.

From the Figure III.1.a, one can note a moderate but distinct effect of the correlation (adjusted for the effect of AUC). The graph suggests that increasing correlation may decrease the type I error independently from the AUC. The difference in shapes of the ROC curves that have equal AUCs does not greatly affect the false rejection rate (type I error) of the statistical test (Figure III.1.b). However the complete results of our investigation of the type I error (Table III.1) suggests a small increase of the false rejection rate when the ROC curves cross.

The effect of prevalence of abnormal subjects in a selected sample is depicted on Figure III.1.c. It can be noted that imbalance of the selected sample affects the behavior of the type I error by strengthening its dependence on the underlying AUC.

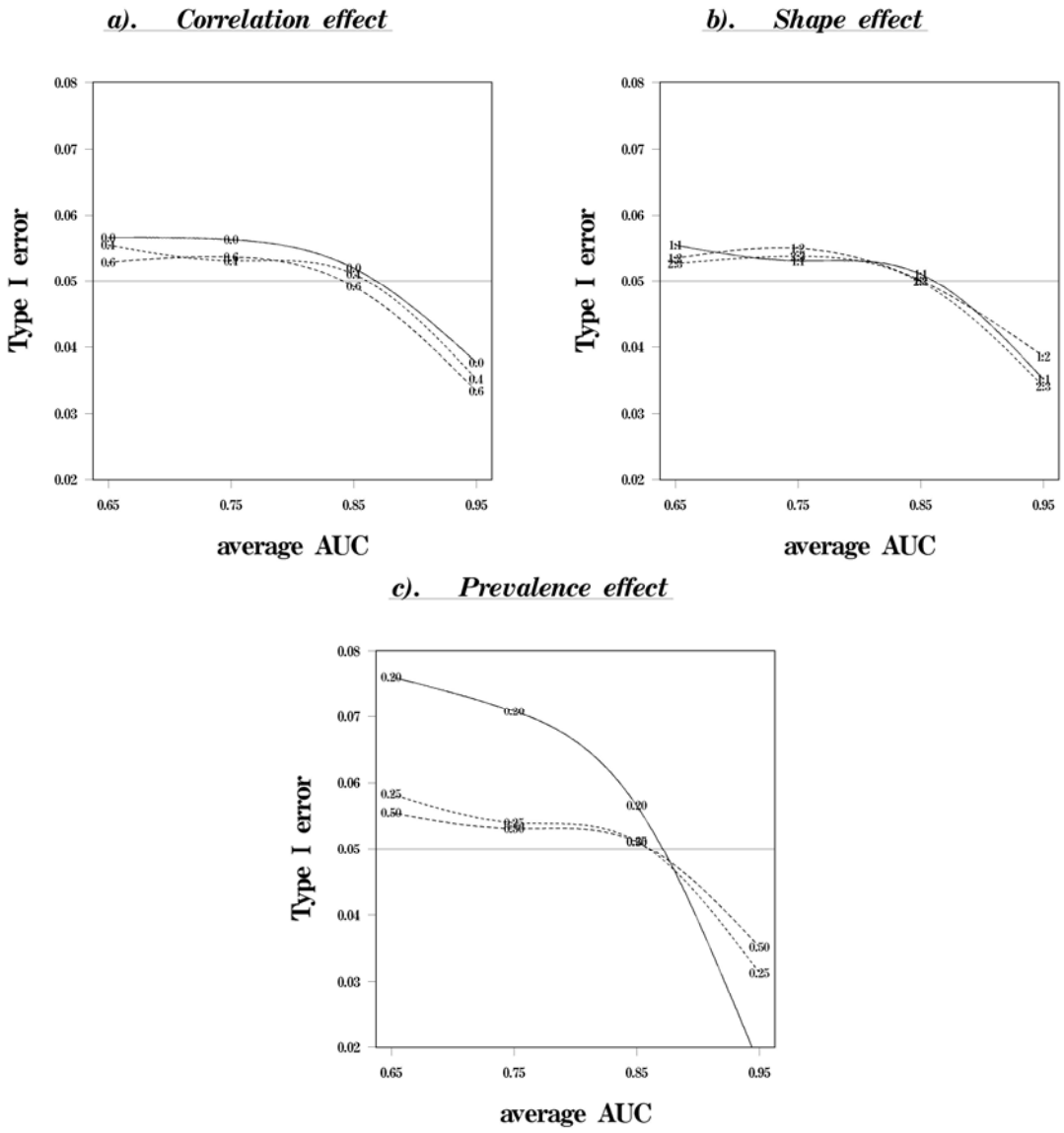

Figure III.1 **Effects of the selected parameters (type I error)**

a). Different levels of the correlation (sample size T=80, shape parameters b_1:b_2=1:1, prevalence p=1/2); b). Difference in shapes indicated by the ratio of the shape parameters b of the two ROC curves (sample size T=80, correlation ρ=0.4, prevalence p=1/2); c). The prevalence of the abnormal subjects in the sample indicated by the proportion (sample size T=80, shape parameters b_1:b_2=1, correlation ρ=0.4)

Table III.1 includes the estimates of the type I error over the complete range of parameters we considered. From presented estimates, it can be seen that for a sample size as large as 80 subjects, the type I error of the conventional procedure can vary from 0.027 to 0.067 depending on underlying parameters.

Table III.1 **Conventional test: type I error**

Prevalence	Correlation	AUC	Total sample size (T) 40 subjects: The same ROC ($b_1=b_2=1$)	40 subjects: Crossing ROCs ($b_1=1,b_2=1/2$)	80 subjects: The same ROC ($b_1=b_2=1$)	80 subjects: Crossing ROCs ($b_1=1,b_2=1/2$)	120 subjects: The same ROC ($b_1=b_2=1$)	120 subjects: Crossing ROCs ($b_1=1,b_2=1/2$)
p=0.5	ρ=0.0	0.65	0.061	0.060	0.057	0.056	0.054	0.054
		0.75	0.062	0.060	0.056	0.058	0.054	0.054
		0.85	0.051	0.054	0.052	0.051	0.053	0.053
		0.95	0.021	0.023	0.038	0.041	0.043	0.047
	ρ=0.4	0.65	0.056	0.060	0.056	0.054	0.053	0.053
		0.75	0.054	0.056	0.053	0.055	0.053	0.050
		0.85	0.045	0.047	0.051	0.050	0.051	0.049
		0.95	0.015	0.017	0.035	0.039	0.039	0.043
	ρ=0.6	0.65	0.050	0.054	0.053	0.053	0.051	0.053
		0.75	0.051	0.053	0.054	0.055	0.051	0.051
		0.85	0.039	0.042	0.049	0.049	0.046	0.049
		0.95	0.012	0.014	0.033	0.035	0.036	0.041
p=0.25	ρ=0.0	0.65	0.062	0.071	0.063	0.067	0.059	0.057
		0.75	0.061	0.073	0.060	0.065	0.056	0.057
		0.85	0.054	0.070	0.056	0.065	0.053	0.057
		0.95	0.017	0.022	0.039	0.052	0.042	0.060
	ρ=0.4	0.65	0.058	0.064	0.058	0.063	0.056	0.055
		0.75	0.056	0.067	0.054	0.062	0.056	0.057
		0.85	0.044	0.055	0.051	0.064	0.053	0.058
		0.95	0.012	0.019	0.031	0.048	0.038	0.058
	ρ=0.6	0.65	0.051	0.062	0.053	0.059	0.055	0.054
		0.75	0.052	0.059	0.052	0.060	0.055	0.054
		0.85	0.037	0.048	0.049	0.060	0.052	0.056
		0.95	0.012	0.022	0.027	0.045	0.036	0.059

The effects of the selected parameters on the statistical power of the conventional test are summarized in Figure III.2 and Table III.2. The relative order of the effects of the parameters remains similar to that observed for the type I error with the average AUC having the largest effect and the difference in shapes of the ROC curves having the smallest effect. However the direction of the relationships does differ. Namely increasing the average AUC or correlation tend increase the statistical power of the conventional test for large AUC differences in contrast to decreasing its type I error (Figure III.2.a,d). Increasing balance between the numbers of subjects in the selected sample not only improves the rate of false rejection (type I error) of the statistical test making it closer to the nominal level but also tend to increase the rate of its true rejections (power) for large AUC differences.

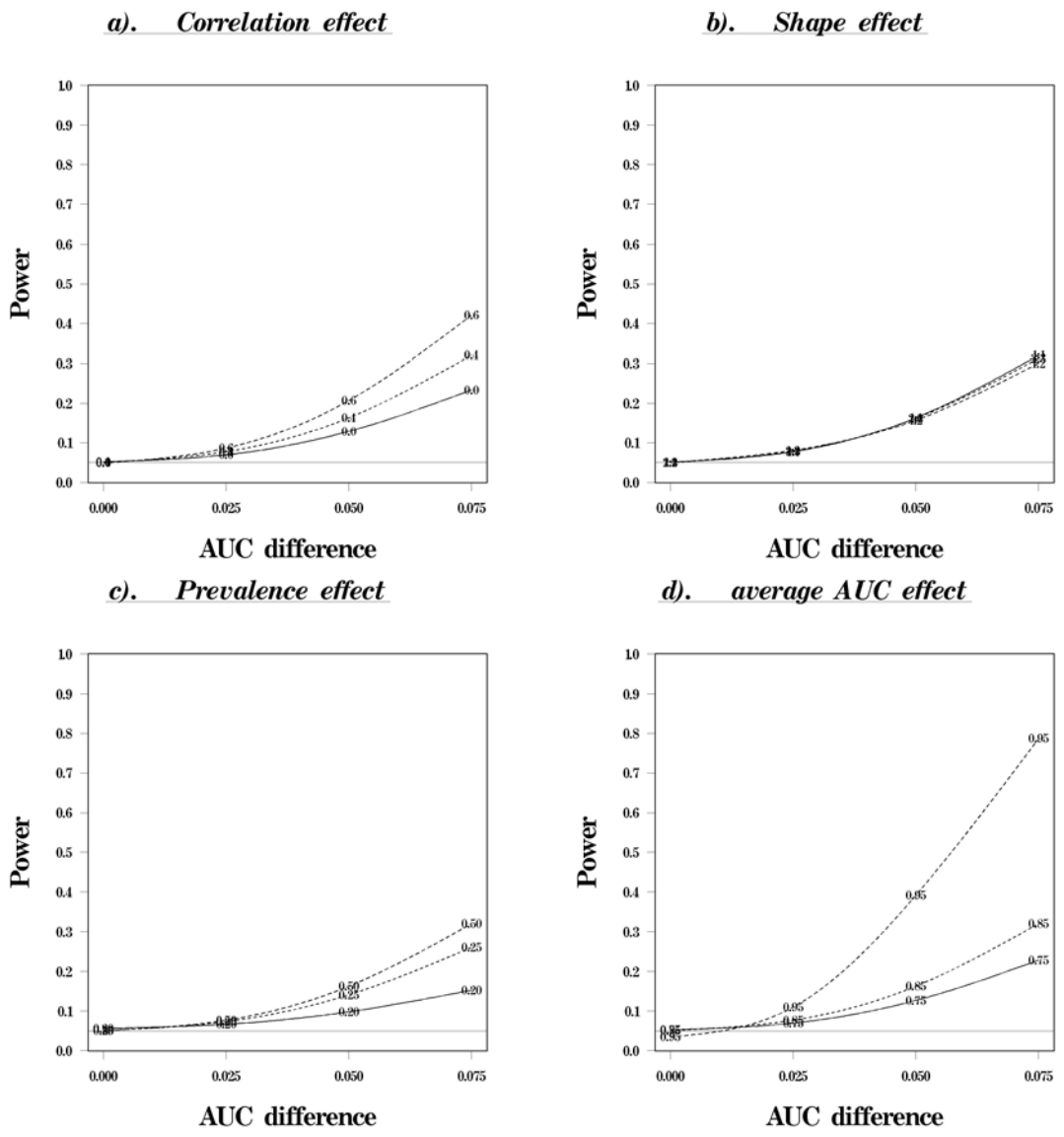

Figure III.2 **Effects of the selected parameters (statistical power)**

a). Different levels of the correlation (sample size T=80, average AUC A=0.85, shape parameters b_1:b_2=1:1, prevalence p=1/2); b). Difference in shapes indicated by the ratio of the shape parameters b of the two ROC curves (sample size T=80, average AUC A=0.85, correlation ρ=0.4, prevalence p=1/2); c). The prevalence of the abnormal subjects in the sample is indicated by the proportion (sample size T=80, average AUC A=0.85, correlation ρ=0.4, shape parameters b_1:b_2=1:1); d). Magnitudes of the underlying average AUC (sample size T=80, correlation ρ=0.4, shape parameters b_1:b_2=1:1,prevalence p=1/2)

Table III.2 **Conventional test: statistical power**

Prevalence	Correlation	Average AUC	AUC difference	Total sample size (T) 40 subjects: The same ROC ($b_1=b_2=1$)	40 subjects: Crossing ROCs ($b_1=1, b_2=1/2$)	80 subjects: The same ROC ($b_1=b_2=1$)	80 subjects: Crossing ROCs ($b_1=1, b_2=1/2$)	120 subjects: The same ROC ($b_1=b_2=1$)	120 subjects: Crossing ROCs ($b_1=1, b_2=1/2$)
p=0.5	ρ=0.0	0.75	0.025	0.064	0.066	0.065	0.066	0.070	0.069
			0.050	0.082	0.085	0.103	0.101	0.125	0.127
			0.075	0.112	0.114	0.166	0.160	0.226	0.220
		0.85	0.025	0.060	0.068	0.070	0.071	0.078	0.079
			0.050	0.085	0.095	0.129	0.130	0.172	0.171
			0.075	0.132	0.142	0.234	0.231	0.331	0.320
		0.95	0.025	0.037	0.047	0.096	0.120	0.144	0.156
			0.050	0.104	0.121	0.317	0.325	0.479	0.469
			0.075	0.247	0.264	0.684	0.675	0.888	0.855
	ρ=0.4	0.75	0.025	0.063	0.065	0.070	0.071	0.081	0.077
			0.050	0.088	0.090	0.128	0.120	0.167	0.164
			0.075	0.131	0.132	0.229	0.212	0.325	0.301
		0.85	0.025	0.058	0.063	0.077	0.081	0.091	0.090
			0.050	0.096	0.105	0.164	0.158	0.235	0.220
			0.075	0.161	0.165	0.321	0.301	0.465	0.429
		0.95	0.025	0.037	0.042	0.110	0.131	0.176	0.181
			0.050	0.109	0.125	0.393	0.393	0.602	0.569
			0.075	0.277	0.284	0.787	0.755	0.953	0.918
	ρ=0.6	0.75	0.025	0.061	0.061	0.075	0.075	0.091	0.085
			0.050	0.098	0.095	0.157	0.144	0.224	0.202
			0.075	0.159	0.151	0.301	0.266	0.436	0.385
		0.85	0.025	0.058	0.062	0.086	0.086	0.107	0.103
			0.050	0.109	0.114	0.207	0.191	0.310	0.276
			0.075	0.195	0.193	0.421	0.374	0.600	0.531
		0.95	0.025	0.036	0.044	0.131	0.150	0.223	0.216
			0.050	0.126	0.135	0.487	0.463	0.727	0.657
			0.075	0.317	0.312	0.868	0.821	0.983	0.956
p=0.25	ρ=0.0	0.75	0.025	0.067	0.085	0.068	0.076	0.071	0.073
			0.050	0.079	0.099	0.095	0.104	0.115	0.115
			0.075	0.106	0.124	0.149	0.152	0.187	0.179
		0.85	0.025	0.061	0.087	0.071	0.090	0.076	0.089
			0.050	0.082	0.115	0.118	0.136	0.148	0.155
			0.075	0.116	0.153	0.198	0.215	0.271	0.257
		0.95	0.025	0.030	0.044	0.083	0.131	0.121	0.176
			0.050	0.068	0.087	0.239	0.301	0.384	0.413
			0.075	0.150	0.175	0.549	0.582	0.777	0.738
	ρ=0.4	0.75	0.025	0.062	0.078	0.069	0.079	0.079	0.079
			0.050	0.084	0.102	0.111	0.123	0.145	0.136
			0.075	0.121	0.138	0.192	0.191	0.260	0.228
		0.85	0.025	0.054	0.079	0.074	0.099	0.088	0.096
			0.050	0.085	0.119	0.142	0.163	0.193	0.191
			0.075	0.136	0.168	0.261	0.261	0.368	0.338
		0.95	0.025	0.025	0.041	0.088	0.145	0.145	0.207
			0.050	0.069	0.089	0.296	0.350	0.484	0.488
			0.075	0.167	0.185	0.643	0.647	0.869	0.818
	ρ=0.6	0.75	0.025	0.060	0.076	0.073	0.084	0.083	0.085
			0.050	0.089	0.108	0.137	0.142	0.182	0.165
			0.075	0.135	0.153	0.249	0.229	0.347	0.291
		0.85	0.025	0.050	0.076	0.079	0.107	0.099	0.111
			0.050	0.089	0.120	0.174	0.191	0.247	0.231
			0.075	0.161	0.189	0.340	0.318	0.483	0.414
		0.95	0.025	0.023	0.041	0.097	0.155	0.177	0.237
			0.050	0.079	0.099	0.359	0.399	0.591	0.564
			0.075	0.189	0.200	0.726	0.709	0.931	0.880

C. SUMMARY

Using the conventional nonparametric procedure for comparing correlated AUCs developed by DeLong *et al.* [19], we attempted to characterize the effects of various parameters on the statistical inferences with small samples. The parameter with the greatest effect on both the type I error and power of the conventional nonparametric test was found to be the average AUC ($\boldsymbol{A}$). When $\boldsymbol{A}$ increases, the type I error decreases making the test overly conservative for large AUCs. However, for small AUCs the type I error of the conventional procedure is elevated above the nominal level. Thus, while the conventional test might be underpowered for large AUCs, it may be inappropriate if the average AUC and sample size are small. This effect can be partially explained by the non-normality of the distribution of the nonparametric estimator of the area. However the decrease of the type I error and hence, potential reduction in the statistical power of the test might also be in part attributed to the increasing bias (with increasing AUC) of the conventional variance estimator (Chapter V Section B).

The correlation between the ratings of the same subjects (ρ) also appears to have a distinct effect on the type I error and power of the conventional test. The direction of the effect of this parameter is similar to that of the average AUC, however the magnitude of the influence is not as large over the considered range of scenarios.

The balance between the number of subjects with and without the abnormality was also shown to be relevant for small-sample inferences. We observed that for the considered ranges of parameters, increasing the balance of the sample improves properties of the conventional statistical test. Namely, the type I error is closer to the nominal level and the statistical power to detect large AUC differences tends to be larger in more balanced (prevalence closer to 0.5) than in less balanced samples.

The difference in shapes of the ROC curves (difference in $\boldsymbol{b}$'s) has little effect on the statistical power or the type I error of the conventional test, although there is some indication that increasing the discrepancy between shapes of two ROC curves slightly elevates the type I error.

IV. PERMUTATION TEST

In this chapter we develop the permutation test for detecting differences between two AUCs in a paired design setting. Such a permutation procedure not only provides an exact (suitable for small samples) and powerful test for detecting differences in overall performances but also permits developing a precise and easy-to-apply approximation. The availability of a simple and precise approximation to the permutation test is a desirable property since, with increasing sample size the exact permutation tests quickly become very demanding computationally. We also conduct simulations to investigate properties of the new procedure. The material in this chapter is accepted for publication in Statistics in Medicine [32].

A. EXACT PERMUTATION TEST

In order to compare the AUCs of the two correlated ROC Curves we propose a permutation test in which the values of the estimator of the AUC difference computed from all possible permutations constitute the distribution of the estimator under the null hypothesis. If the two modalities had the same underlying scale of ratings we could justify directly permuting the actual ratings for each subject. However, since the ROC curves are invariant with respect to monotone transformations of the data, without loss of generality we can permute the rank of the ratings (or appropriate monotonic transformation of the ratings) as if they were actual ratings on the same underlying scale. Hence the use of the transformed ratings allows us to compare the modalities with different underlying scales as well. We will refer to monotonically transformed ratings or ranks of the ratings as rank-ratings.

The proposed test is conducted by permuting the subject specific rank-ratings between the two modalities within the structure of given pairs. The 2^{N+M} permutations are created by

exchanging the rank-rating observed for each subject for the two modalities, and permutations for different subjects are done independently of each other. Thus if for the $\boldsymbol{i}^{\text{th}}$ normal subject ($\boldsymbol{X_i}$) and $\boldsymbol{j}^{\text{th}}$ abnormal subject ($\boldsymbol{Y_j}$) the rank-ratings observed in first and second modality are $x_i^1, x_i^2, y_j^1 \text{ and } y_j^2$ respectively then all possible permutations that can be performed with those two subjects are as follows:

$$
\begin{array}{cc}
\textit{I mod} & \textit{II mod} \\
(x_i^1, y_j^1) & (x_i^2, y_j^2) \qquad X's - \textit{not exchanged}, Y's - \textit{not exchanged} \\
(x_i^2, y_j^1) & (x_i^1, y_j^2) \qquad X's - \textit{exchanged}, Y's - \textit{not exchanged} \\
(x_i^1, y_j^2) & (x_i^2, y_j^1) \qquad X's - \textit{not exchanged}, Y's - \textit{exchanged} \\
(x_i^2, y_j^2) & (x_i^1, y_j^1) \qquad X's - \textit{exchanged}, Y's - \textit{exchanged}
\end{array}
\qquad \text{(IV.A.1)}
$$

where the pairs in the first column are assumed to be observed for the first modality and the pairs in the second column are assumed to be observed for the second modality.

To justify equal probability of all permutations under the null hypothesis, we assume the exchangeability of the subject specific rank-ratings between the two modalities. Exchangeability means that the joint distribution of the rank-ratings is symmetric with respect to its arguments (separately for the normal and abnormal subjects) [24]. The exchangeability assumption is a stricter assumption than the equality of the ROC curves. We consider our procedure to have as a null hypothesis equality of ROC curves under the assumption of exchangeability. The distribution of the differences in the estimated Areas under the ROC Curves over all permutations is readily obtained and the rejection region can be selected based on $\alpha/2$ and $1-\alpha/2$ percentile values. The two-sided p-value can be defined as:

$$
p = P_{\Omega}\left(\left|\hat{A}^1 - \hat{A}^2\right| \geq \left|\hat{A}_0^1 - \hat{A}_0^2\right| \right) = \frac{\#\left\{ \left|\hat{A}_t^1 - \hat{A}_t^2\right| \geq \left|\hat{A}_0^1 - \hat{A}_0^2\right| \right\}}{2^{N+M}} \quad t = 1,\ldots,2^{N+M}
$$

where $2^{\boldsymbol{N+M}}$ is the total number of all possible permutations, $\hat{A}_0^1 - \hat{A}_0^2$ - is the observed AUC difference and $\hat{A}_t^1 - \hat{A}_t^2$ is the AUC difference computed from the $\boldsymbol{t}^{\text{th}}$ permutation.

Properties of the difference between two nonparametric AUC estimators allow for the construction of a simple asymptotic procedure. As a member of the class of U-statistics the parametric estimator of the AUC difference is known to be asymptotically normally distributed under quite general conditions [26]. Since the nonparametric estimator of the AUC is unbiased, the expectation of the AUC difference is 0 when the two AUCs are equal and this fact is also illustrated in the permutation space $\boldsymbol{\Omega}$ under the stricter assumption of exchangeability (see Appendix A). The exchangeability assumption also allows a simple calculation of the exact variance of the AUC difference in the permutation space as shown in Appendix A.

Hence, under the assumption of asymptotic normality of the U-statistic and the additional assumption of exchangeability of within subject rank-ratings:

$$\frac{\hat{A}^1 - \hat{A}^2}{\sqrt{Var_{\Omega}\left(\hat{A}^1 - \hat{A}^2\right)}} \xrightarrow{d} N(0,1).$$

Thus, a test of the hypothesis of equality of ROC curves that is sensitive to the differences in AUCs can be conducted using the statistic $\frac{\hat{A}_0^1 - \hat{A}_0^2}{\sqrt{Var_{\Omega}\left(\hat{A}^1 - \hat{A}^2\right)}}$, where the exact variance in the denominator is obtained as shown in Appendix A.

B. SIMULATION STUDY

We performed extensive computer simulations to investigate the type I error and the statistical power of the asymptotic procedure for different underlying AUCs, correlations between subject ratings across modalities and different sample sizes. In our simulations we assume equal correlation across modalities for the ratings of normal and abnormal subjects rated on both continuous and discrete scales and consider scenarios with non-crossing as well as crossing ROC curves.

The general protocol of simulations follows the approach described in Chapter II, Section A. In addition to simulations of continuous datasets, we also investigated the rejection rate of the proposed procedure in the discrete case. The discrete ratings were simulated by grouping the

binormal data into 5 categories. The parameters of each pair of binormal distributions ($\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{\rho}$) were selected to produce predetermined parameters in the resultant discrete distributions.

Table IV.1 compares the type I error and the statistical power of the exact permutation test to its normal approximation. Note that the rejection rate formally corresponds to the Type I error of the proposed procedure in cases of equal ROC curves (non-crossing ROC curves with 0 AUC difference) and to the power in all other cases considered. Due to the relatively large computational time required for the implementation of the exact procedure the comparisons presented here are limited to small sample sizes. However, even with these small samples there is a good agreement between the exact and approximate test. The simulations in Table IV.1 show that even for six normal and six abnormal subjects the asymptotic test is adequate. (In general we found that it is feasible to conduct the exact test with the sample size as large as fifteen normal and fifteen abnormal subjects without using a large amount of computer time.) Thus, for the larger sample sizes as presented in subsequent tables we simulate only the operating characteristics of the asymptotic test since the results for the exact test should be essentially the same.

Table IV.1 **Exact procedure vs. its approximation: rejection rate**

Average AUC	AUC difference	Non-crossing ROC curves ($b_1=b_2=1$)						Crossing ROC curves ($b_1=1, b_2=1/2$)					
		ρ=0.0		ρ=0.4		ρ=0.6		ρ=0.0		ρ=0.4		ρ=0.6	
		Asymptotic	Exact	Asymptotic	Exact	Asymptotic	Exact	Asymptotic	Exact	Asymptotic	Exact	Asymptotic	Exact
0.70	**0.00**	0.045	0.047	0.044	0.048	0.045	0.050	0.047	0.051	0.049	0.054	0.047	0.053
	0.05	0.048	0.051	0.051	0.055	0.051	0.057	0.051	0.055	0.055	0.060	0.056	0.062
	0.10	0.064	0.066	0.069	0.076	0.080	0.086	0.068	0.069	0.074	0.081	0.081	0.092
	0.15	0.089	0.093	0.108	0.114	0.133	0.143	0.093	0.098	0.110	0.121	0.130	0.144
	0.20	0.123	0.128	0.161	0.169	0.205	0.222	0.130	0.133	0.165	0.176	0.202	0.220
0.75	**0.00**	0.037	0.040	0.039	0.044	0.039	0.043	0.039	0.043	0.039	0.046	0.039	0.046
	0.05	0.046	0.047	0.046	0.050	0.046	0.051	0.046	0.048	0.045	0.051	0.048	0.056
	0.10	0.060	0.063	0.065	0.070	0.074	0.082	0.063	0.066	0.069	0.077	0.076	0.085
	0.15	0.084	0.087	0.103	0.111	0.125	0.137	0.090	0.095	0.107	0.117	0.128	0.143
	0.20	0.121	0.127	0.160	0.172	0.202	0.224	0.130	0.139	0.165	0.181	0.201	0.228
0.80	**0.00**	0.031	0.033	0.030	0.035	0.031	0.034	0.033	0.037	0.032	0.039	0.033	0.041
	0.05	0.035	0.037	0.037	0.042	0.038	0.045	0.038	0.041	0.038	0.046	0.043	0.049
	0.10	0.051	0.054	0.057	0.065	0.066	0.077	0.055	0.060	0.061	0.072	0.070	0.081
	0.15	0.076	0.081	0.097	0.109	0.124	0.140	0.083	0.088	0.101	0.114	0.123	0.142
	0.20	0.116	0.124	0.159	0.178	0.208	0.233	0.122	0.130	0.164	0.184	0.204	0.236
0.85	**0.00**	0.020	0.023	0.024	0.027	0.022	0.025	0.022	0.027	0.023	0.029	0.025	0.029
	0.05	0.025	0.027	0.030	0.034	0.031	0.034	0.027	0.029	0.031	0.037	0.035	0.039
	0.10	0.039	0.041	0.051	0.057	0.063	0.070	0.042	0.045	0.054	0.062	0.063	0.074
	0.15	0.064	0.068	0.091	0.102	0.117	0.135	0.069	0.074	0.093	0.107	0.117	0.141
	0.20	0.109	0.116	0.155	0.176	0.203	0.237	0.113	0.123	0.160	0.182	0.203	0.238
0.90	**0.00**	0.010	0.012	0.012	0.013	0.014	0.015	0.010	0.012	0.015	0.017	0.016	0.018
	0.05	0.012	0.015	0.020	0.021	0.023	0.024	0.015	0.017	0.021	0.023	0.026	0.028
	0.10	0.027	0.030	0.044	0.047	0.059	0.063	0.029	0.033	0.047	0.052	0.061	0.067
	0.15	0.055	0.060	0.087	0.098	0.118	0.137	0.057	0.062	0.092	0.102	0.121	0.138
0.95	**0.00**	0.002	0.002	0.004	0.003	0.007	0.005	0.003	0.002	0.004	0.004	0.007	0.005
	0.05	0.005	0.005	0.013	0.011	0.017	0.015	0.005	0.005	0.014	0.012	0.018	0.016

Simulated samples consist of 6 normal and 6 abnormal subjects

We compared the rejection rate of the proposed asymptotic test to that of the conventional nonparametric procedure developed by DeLong *et al.* [19]. The estimates are presented in Table IV.2 for continuous data and Table IV.3 for discrete data. Note that in these tables the rejection rate provides the estimates of the type I error of the conventional procedure for all combinations of parameters we considered. However, since the null hypothesis of the proposed procedure is formally the equality of ROC curves subject to exchangeability, in situations of crossing ROC curves the rejection rate is the statistical power. For moderate sample sizes and for the scenario where non-crossing ROC curves have equal and large AUC that are at least moderately correlated between modalities, the proposed permutation test demonstrates a type I error that is less conservative than the conventional test. This effect is especially evident with smaller sample sizes.

For equal AUCs arising from crossing ROC curves the rejection rate of the permutation test (power) is very close to that of the conventional nonparametric area test (type I error). The practical relevance of this finding is that the proposed procedure should not be used to detect crossing ROC curves with the same AUCs. However, the closeness of the rejection rate to the nominal significance level suggests that even though the proposed procedure is formally a test for equality of ROC curves it provides an approximate test of equality of AUCs. As such, it is useful to compare the power of the proposed procedure to that of the conventional method of DeLong *et al.* [19].

For non-crossing ROC curves with a correlation $\rho \geq 0.4$ and an average AUC $A \geq 0.80$ the power of the proposed test is greater than that of the conventional procedure (Table IV.4). This power increase is expected because the proposed test is less conservative in this range of parameters. For lower correlations and smaller average AUCs, DeLong *et al.*'s procedure has slightly greater power. However, this is a region where the type I error of the conventional test is slightly elevated. With increasing sample size the operating characteristics of the two procedures approach each other. For crossing ROC curves (Table IV.5), the pattern is similar. Specifically, for higher correlations and higher average areas the rejection rate for the proposed test is higher.

Table IV.2 **Permutation vs. conventional test: rejection rate (continuous data)**

Correlation	Average AUC	N=20 normal and M=20 abnormal subjects				N=40 normal and M=40 abnormal subjects				N=60 normal and M=60 abnormal subjects			
		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)	
		D	A	D	A	D	A	D	A	D	A	D	A
ρ=0.0	**0.70**	0.057	0.050	0.058	0.055	0.053	0.050	0.056	0.055	0.052	0.051	0.054	0.054
	0.75	0.053	0.049	0.056	0.053	0.051	0.049	0.055	0.054	0.052	0.050	0.054	0.053
	0.80	0.051	0.047	0.054	0.052	0.051	0.049	0.054	0.054	0.052	0.050	0.053	0.054
	0.85	0.047	0.045	0.051	0.049	0.049	0.047	0.053	0.053	0.051	0.050	0.053	0.053
	0.90	0.040	0.041	0.043	0.044	0.045	0.045	0.052	0.052	0.049	0.048	0.050	0.051
	0.95	0.020	0.027	0.021	0.027	0.037	0.041	0.043	0.045	0.044	0.046	0.046	0.048
ρ=0.2	**0.70**	0.054	0.050	0.055	0.053	0.053	0.051	0.055	0.056	0.054	0.053	0.053	0.055
	0.75	0.053	0.050	0.055	0.054	0.052	0.051	0.054	0.055	0.051	0.049	0.054	0.055
	0.80	0.049	0.048	0.052	0.052	0.050	0.049	0.052	0.054	0.051	0.051	0.052	0.053
	0.85	0.045	0.046	0.049	0.051	0.048	0.048	0.052	0.053	0.051	0.051	0.051	0.054
	0.90	0.036	0.042	0.041	0.045	0.045	0.046	0.050	0.052	0.048	0.049	0.049	0.051
	0.95	0.017	0.028	0.017	0.029	0.036	0.041	0.040	0.047	0.044	0.046	0.047	0.051
ρ=0.4	**0.70**	0.052	0.051	0.054	0.057	0.051	0.051	0.053	0.056	0.052	0.051	0.053	0.056
	0.75	0.049	0.049	0.053	0.055	0.051	0.051	0.052	0.056	0.050	0.051	0.052	0.058
	0.80	0.046	0.048	0.050	0.054	0.050	0.051	0.053	0.056	0.051	0.052	0.051	0.055
	0.85	0.042	0.047	0.045	0.050	0.048	0.050	0.049	0.053	0.051	0.052	0.050	0.054
	0.90	0.032	0.042	0.037	0.045	0.043	0.048	0.049	0.054	0.046	0.048	0.048	0.052
	0.95	0.014	0.028	0.015	0.030	0.032	0.042	0.040	0.048	0.040	0.046	0.043	0.050
ρ=0.6	**0.70**	0.047	0.052	0.049	0.058	0.048	0.050	0.050	0.060	0.050	0.050	0.051	0.062
	0.75	0.046	0.050	0.047	0.057	0.047	0.051	0.049	0.059	0.049	0.051	0.050	0.060
	0.80	0.042	0.048	0.045	0.056	0.048	0.052	0.049	0.058	0.049	0.051	0.050	0.057
	0.85	0.037	0.047	0.040	0.051	0.049	0.054	0.050	0.058	0.048	0.051	0.048	0.055
	0.90	0.026	0.042	0.031	0.047	0.042	0.048	0.048	0.058	0.044	0.049	0.046	0.053
	0.95	0.010	0.029	0.011	0.031	0.030	0.042	0.036	0.050	0.040	0.047	0.042	0.051

D- conventional procedure (DeLong et al.); A-approximation to permutation test
AUCs of two modalities are the same ($\Delta=0$)

Table IV.3 **Permutation vs. conventional test: rejection rate (discrete data)**

Correlation	AUC	N=20 normal and M=20 abnormal subjects				N=40 normal and M=40 abnormal subjects				N=60 normal and M=60 abnormal subjects			
		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1, b_2=1/2$)	
		D	**A**	**D**	**A**	**D**	**A**	**D**	**A**	**D**	**A**	**D**	**A**
ρ=0.0	**0.70**	0.057	0.049	0.059	0.049	0.054	0.050	0.056	0.053	0.054	0.051	0.054	0.052
	0.75	0.055	0.047	0.055	0.048	0.053	0.049	0.054	0.050	0.054	0.052	0.054	0.050
	0.80	0.054	0.048	0.054	0.047	0.052	0.049	0.054	0.051	0.055	0.053	0.054	0.049
	0.85	0.052	0.047	0.054	0.045	0.052	0.049	0.054	0.048	0.052	0.050	0.054	0.049
	0.90	0.046	0.045	0.051	0.044	0.049	0.046	0.054	0.047	0.048	0.049	0.050	0.045
	0.95	0.023	0.033	0.032	0.041	0.044	0.046	0.050	0.045	0.047	0.047	0.050	0.043
ρ=0.2	**0.70**	0.056	0.048	0.057	0.049	0.052	0.049	0.056	0.051	0.056	0.053	0.055	0.053
	0.75	0.055	0.049	0.056	0.047	0.052	0.049	0.053	0.048	0.053	0.050	0.053	0.049
	0.80	0.053	0.048	0.052	0.043	0.054	0.051	0.051	0.047	0.053	0.052	0.053	0.049
	0.85	0.049	0.046	0.050	0.041	0.052	0.050	0.053	0.046	0.053	0.051	0.051	0.045
	0.90	0.042	0.044	0.047	0.041	0.047	0.046	0.052	0.043	0.049	0.048	0.050	0.044
	0.95	0.016	0.028	0.028	0.039	0.040	0.043	0.049	0.043	0.048	0.048	0.046	0.039
ρ=0.4	**0.70**	0.055	0.049	0.055	0.048	0.052	0.049	0.055	0.051	0.054	0.052	0.055	0.052
	0.75	0.055	0.049	0.052	0.045	0.054	0.053	0.055	0.050	0.054	0.053	0.054	0.052
	0.80	0.052	0.048	0.048	0.039	0.053	0.052	0.051	0.046	0.052	0.049	0.054	0.048
	0.85	0.047	0.046	0.048	0.040	0.053	0.051	0.050	0.044	0.052	0.052	0.051	0.044
	0.90	0.037	0.040	0.042	0.036	0.047	0.048	0.050	0.042	0.048	0.048	0.045	0.039
	0.95	0.014	0.028	0.024	0.042	0.037	0.040	0.049	0.043	0.047	0.048	0.048	0.037
ρ=0.6	**0.70**	0.053	0.049	0.053	0.047	0.054	0.052	0.053	0.051	0.052	0.050	0.054	0.051
	0.75	0.051	0.049	0.050	0.046	0.053	0.052	0.056	0.049	0.052	0.052	0.052	0.049
	0.80	0.045	0.045	0.046	0.040	0.054	0.054	0.053	0.046	0.051	0.050	0.053	0.048
	0.85	0.044	0.047	0.047	0.039	0.050	0.050	0.049	0.041	0.051	0.049	0.049	0.039
	0.90	0.034	0.040	0.036	0.036	0.045	0.046	0.048	0.037	0.047	0.047	0.045	0.033
	0.95	0.011	0.025	0.018	0.044	0.031	0.037	0.044	0.039	0.043	0.043	0.047	0.034

D- conventional procedure (DeLong et al.); A-approximation to permutation test
AUCs of two modalities are the same ($\Delta=0$)

Table IV.4 **Permutation vs. conventional test: statistical power (non-crossing ROCs)**

Average AUC	AUC difference	N=20 normal and M=20 abnormal subjects						N=40 normal and M=40 abnormal subjects						N=60 normal and M=60 abnormal subjects					
		ρ=0.0		ρ=0.4		ρ=0.6		ρ=0.0		ρ=0.4		ρ=0.6		ρ=0.0		ρ=0.4		ρ=0.6	
		D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A
0.70	**0.05**	0.077	0.069	0.082	0.082	0.089	0.094	0.101	0.096	0.122	0.121	0.151	0.155	0.118	0.114	0.155	0.153	0.200	0.202
	0.10	0.143	0.130	0.185	0.181	0.240	0.245	0.237	0.230	0.334	0.332	0.446	0.450	0.321	0.315	0.466	0.464	0.613	0.614
	0.15	0.262	0.246	0.361	0.353	0.470	0.476	0.453	0.443	0.632	0.628	0.783	0.784	0.617	0.610	0.806	0.805	0.921	0.922
	0.20	0.415	0.393	0.569	0.563	0.712	0.716	0.698	0.688	0.864	0.863	0.956	0.956	0.858	0.855	0.967	0.966	0.996	0.996
0.75	**0.05**	0.077	0.072	0.082	0.083	0.092	0.097	0.105	0.100	0.132	0.131	0.162	0.166	0.127	0.124	0.167	0.167	0.220	0.223
	0.10	0.154	0.140	0.200	0.199	0.257	0.264	0.266	0.260	0.370	0.367	0.492	0.497	0.360	0.354	0.518	0.517	0.667	0.669
	0.15	0.289	0.271	0.394	0.390	0.511	0.520	0.510	0.501	0.691	0.688	0.827	0.831	0.679	0.673	0.857	0.855	0.953	0.953
	0.20	0.467	0.447	0.620	0.617	0.768	0.772	0.757	0.750	0.906	0.904	0.977	0.978	0.909	0.906	0.983	0.983	0.998	0.998
0.80	**0.05**	0.078	0.071	0.085	0.090	0.097	0.108	0.117	0.112	0.143	0.144	0.179	0.186	0.143	0.140	0.192	0.193	0.253	0.257
	0.10	0.170	0.160	0.227	0.231	0.285	0.301	0.309	0.302	0.429	0.429	0.558	0.567	0.427	0.422	0.594	0.595	0.746	0.749
	0.15	0.331	0.316	0.451	0.451	0.573	0.590	0.592	0.583	0.767	0.767	0.888	0.890	0.769	0.765	0.915	0.915	0.979	0.980
	0.20	0.547	0.528	0.703	0.703	0.829	0.840	0.843	0.837	0.955	0.956	0.992	0.993	0.959	0.958	0.995	0.996	0.999	0.999
0.85	**0.05**	0.083	0.079	0.092	0.098	0.103	0.121	0.132	0.130	0.167	0.171	0.215	0.225	0.174	0.171	0.231	0.234	0.310	0.317
	0.10	0.206	0.196	0.264	0.273	0.337	0.365	0.391	0.384	0.527	0.532	0.661	0.673	0.543	0.537	0.713	0.716	0.846	0.850
	0.15	0.422	0.409	0.548	0.557	0.667	0.693	0.726	0.720	0.872	0.874	0.955	0.957	0.891	0.889	0.971	0.972	0.997	0.997
	0.20	0.687	0.673	0.821	0.827	0.907	0.920	0.948	0.946	0.991	0.992	0.999	0.999	0.995	0.994	1.000	1.000	1.000	1.000
0.90	**0.05**	0.092	0.091	0.105	0.120	0.118	0.150	0.176	0.176	0.225	0.235	0.286	0.303	0.243	0.242	0.319	0.327	0.420	0.435
	0.10	0.278	0.273	0.345	0.374	0.427	0.488	0.559	0.556	0.704	0.714	0.827	0.843	0.749	0.749	0.884	0.888	0.956	0.960
	0.15	0.611	0.610	0.718	0.752	0.799	0.847	0.929	0.928	0.981	0.984	0.997	0.997	0.991	0.991	0.999	0.999	1.000	1.000
0.95	**0.05**	0.104	0.117	0.114	0.164	0.130	0.218	0.325	0.331	0.401	0.435	0.493	0.545	0.482	0.488	0.609	0.629	0.732	0.757

D- conventional procedure (DeLong et al.)
A-approximation to permutation test

Table IV.5 **Permutation vs. conventional test: statistical power (crossing ROCs)**

Average AUC	AUC difference	N=20 normal and M=20 abnormal subjects						N=40 normal and M=40 abnormal subjects						N=60 normal and M=60 abnormal subjects					
		ρ=0.0		ρ=0.4		ρ=0.6		ρ=0.0		ρ=0.4		ρ=0.6		ρ=0.0		ρ=0.4		ρ=0.6	
		D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A
0.70	**0.05**	0.077	0.072	0.083	0.087	0.091	0.103	0.100	0.098	0.119	0.126	0.139	0.155	0.116	0.116	0.147	0.156	0.181	0.201
	0.10	0.143	0.136	0.180	0.184	0.215	0.237	0.230	0.227	0.313	0.322	0.396	0.422	0.310	0.309	0.431	0.445	0.545	0.576
	0.15	0.252	0.241	0.333	0.340	0.423	0.450	0.439	0.436	0.588	0.599	0.721	0.742	0.597	0.596	0.765	0.775	0.880	0.895
	0.20	0.403	0.388	0.536	0.542	0.657	0.682	0.669	0.666	0.831	0.839	0.925	0.936	0.841	0.841	0.952	0.954	0.988	0.990
0.75	**0.05**	0.080	0.077	0.088	0.091	0.093	0.107	0.106	0.105	0.128	0.137	0.151	0.168	0.124	0.123	0.159	0.168	0.198	0.216
	0.10	0.155	0.146	0.195	0.202	0.234	0.258	0.255	0.253	0.346	0.360	0.438	0.468	0.351	0.349	0.476	0.491	0.594	0.625
	0.15	0.278	0.268	0.367	0.376	0.460	0.491	0.489	0.486	0.648	0.658	0.769	0.789	0.656	0.656	0.819	0.829	0.916	0.928
	0.20	0.453	0.438	0.588	0.596	0.707	0.731	0.736	0.733	0.875	0.881	0.951	0.958	0.893	0.893	0.970	0.972	0.994	0.995
0.80	**0.05**	0.085	0.081	0.094	0.101	0.102	0.119	0.115	0.115	0.141	0.148	0.166	0.185	0.140	0.140	0.182	0.191	0.226	0.247
	0.10	0.175	0.168	0.220	0.228	0.266	0.295	0.301	0.298	0.397	0.410	0.501	0.531	0.410	0.410	0.546	0.559	0.669	0.693
	0.15	0.323	0.314	0.424	0.437	0.523	0.554	0.572	0.571	0.726	0.735	0.835	0.852	0.747	0.748	0.888	0.896	0.954	0.961
	0.20	0.533	0.519	0.668	0.679	0.775	0.800	0.825	0.823	0.931	0.935	0.979	0.983	0.947	0.948	0.989	0.991	0.999	0.999
0.85	**0.05**	0.092	0.089	0.103	0.114	0.107	0.133	0.135	0.134	0.166	0.174	0.195	0.217	0.175	0.176	0.222	0.234	0.275	0.299
	0.10	0.206	0.201	0.260	0.275	0.323	0.361	0.377	0.378	0.488	0.503	0.596	0.630	0.516	0.518	0.661	0.675	0.778	0.800
	0.15	0.412	0.402	0.521	0.542	0.615	0.659	0.699	0.698	0.832	0.841	0.910	0.925	0.866	0.867	0.952	0.956	0.986	0.989
	0.20	0.672	0.664	0.794	0.808	0.868	0.892	0.930	0.931	0.982	0.984	0.995	0.997	0.990	0.990	0.999	0.999	1.000	1.000
0.90	**0.05**	0.106	0.105	0.116	0.136	0.126	0.163	0.179	0.180	0.222	0.235	0.268	0.296	0.239	0.242	0.305	0.320	0.374	0.406
	0.10	0.286	0.288	0.351	0.387	0.411	0.482	0.538	0.541	0.661	0.679	0.763	0.790	0.718	0.721	0.836	0.847	0.915	0.927
	0.15	0.612	0.614	0.697	0.737	0.765	0.823	0.905	0.906	0.961	0.967	0.986	0.990	0.983	0.984	0.997	0.998	1.000	1.000
0.95	**0.05**	0.119	0.138	0.127	0.187	0.137	0.233	0.340	0.351	0.401	0.439	0.471	0.528	0.478	0.488	0.569	0.596	0.666	0.701

D- conventional procedure (DeLong et al.)
A-approximation to permutation test

In summary, our simulations demonstrate close agreement of the type I error of the proposed permutation test and the nominal value with reasonably small sample sizes. Furthermore, for moderate correlation between modalities, large average AUC and small sample sizes the test possesses better operating characteristics than the conventional nonparametric AUC test developed by DeLong *et al*. Finally, within the considered range of parameters, the power of the proposed test to detect crossing ROC curves with equal AUCs is close to the nominal significance level suggesting that a rejection of the null hypothesis is unlikely to occur unless there is a difference in the AUCs of the two curves.

C. SUMMARY AND DISCUSSION

The proposed procedure offers a useful supplement to existing methods for comparing performances of diagnostic systems in a paired design setting. It provides the ability to conduct the exact test and allows for an easy-to-implement approximation when the sample size is large. This test has enhanced power against the alternatives of a difference in AUCs and its null hypothesis is equality of ROC curves under the additional assumption of exchangeability of the within subject's rank-ratings for modalities with equal ROC curves. In experiments with small to moderate sample sizes ($\leq$ 60 normal and 60 abnormal subjects) when the average of two correlated AUCs is at least moderate ($\geq$0.80) and correlation within subject's ratings is not low ($\geq$0.4) the presented test possesses more appropriate type I error and a greater statistical power as compared to the conventional nonparametric test by DeLong *et al.* [19]. Despite the fact that the conventional test has greater statistical power than the permutation test for small average AUC or low correlation between modalities, these situations are less likely to be encountered when evaluating diagnostic imaging technologies or practices. Furthermore, part of the observed superiority of the conventional procedure for low AUC might be attributed to its elevated type I error. For larger sample sizes the proposed test and the method of DeLong produce similar type I error and statistical power.

The simulations performed by Venkatraman and Begg [24] showed that for ROC curves that do not cross their procedure for the nonparametric comparison has lower power than the

conventional nonparametric test of DeLong *et al.* This is expected because the procedure is designed to detect differences in ROC curves rather than detecting differences in AUCs only, as does the conventional nonparametric AUC test. The procedure presented here, although formally a test of difference in ROC curves, is constructed to detect differences in AUCs. Our investigations show that it has comparable power to the conventional nonparametric AUC test and for some ranges of the parameters of practical interest has superior operating characteristics. Alternatively, if the primary interest of the investigator is to detect differences in ROC curves at every operating point, even if these have similar AUCs, then the method of Venkatraman and Begg should be used.

The derived formula for the exact variance of the difference between correlated AUC estimators in the permutation space ($\boldsymbol{\Omega}$) enables one to construct a normal approximation to the exact procedure that is precise even for small samples. The availability of an asymptotic procedure that provides a simple and precise approximation to the permutation test is a desirable property since with increasing sample size the exact permutation tests quickly become very demanding computationally. Also, the approach demonstrated in the Appendix A can be relatively easily adapted to different permutation schemes. For example, following the steps described in the Appendix A, one can derive the exact variance of the difference in nonparametric AUC estimators in the permutation space where ties between the permuted rank-ratings are uniformly broken, or alternatively in the permutation space where the rank-ratings are permuted within the groups of normal and abnormal subjects. The latter permutation scheme can be used to develop a procedure for an unpaired design [25].

V. BOOTSTRAP-VARIANCE AND ASYMPTOTIC TEST

The bootstrap is a powerful nonparametric approach [41] and the ideas of exploiting the bootstrap procedure in ROC analysis have been previously proposed [43,39,37]. Unfortunately the intensity of the computations required to create all bootstrap-samples or an additional error associated with incomplete sampling of the bootstrap-space reduce the attractiveness of the approach.

The conventional procedure for comparing correlated AUCs developed by DeLong *et al.* [19] is equivalent to the two-sample jackknife procedure [22]. Since the bootstrap approach is usually considered to be superior to the jackknife, it is reasonable to investigate the properties of the asymptotic bootstrap test compared to the conventional test. For a specific statistic such as the nonparametric estimator of the AUC, the closed-form bootstrap-variance can be derived allowing one to construct an easy-to-compute asymptotic test. We compare the properties of the variance estimators and the corresponding asymptotic procedures based on jackknife and bootstrap approaches using computer simulations.

A. EXACT VARIANCE

The essence of the bootstrap approach is to construct a space of equally-probable bootstrap-samples created from a single random sample observed originally. Each bootstrap-sample has the same size as the original sample and each data point in the bootstrap-sample is one of the original data points. (In other words the bootstrap-sample is a random sample of predetermined size that is drawn with replacement from the originally observed data.) The values of the primary statistic calculated from each bootstrap-sample constitute the bootstrap-distribution of that statistic and can be used for inferential purposes. We are interested only in one parameter of such

a bootstrap-distribution, namely in its variance. Since the nonparametric estimator of the AUC (or AUC difference) has a relatively simple form its variance is straightforward to express (II.B.2.2) and its bootstrap-variance can be computed exactly without creating all possible samples.

In the specific problem that we consider, the data is assumed to be based on a random sample of subjects; hence the subjects are appropriate units for bootstrap re-sampling. The sample of subjects is composed from the two independent samples of normal and abnormal subjects; therefore we resample within corresponding sub-samples (normal subjects separately from abnormal). Under the nonparametric bootstrap approach [41] that we adopt, 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.

As defined previously (Chapter II Section A), let $\left\{\left(x_i^1, x_i^2\right)\right\}_{i=1}^{N}$ be normal subjects' ratings and $\left\{\left(y_j^1, y_j^2\right)\right\}_{j=1}^{M}$ be abnormal subjects' ratings. Then a normal (abnormal) subject from a bootstrap-sample of subjects can, with equal probability, have one of the pairs of ratings observed in original data for normal (abnormal) subjects i.e. the pair of ratings in a bootstrap-sample is uniformly distributed over the discrete set of pairs of ratings present in the original dataset. We denote this as:

$$\left(X^1, X^2\right) \sim Uniform\left[\left\{\left(x_i^1, x_i^2\right)\right\}_{i=1}^{N}\right] \text{ and } \quad \left(Y^1, Y^2\right) \sim Uniform\left[\left\{\left(y_j^1, y_j^2\right)\right\}_{j=1}^{M}\right]$$

Every bootstrap-sample is taken with replacement from the original sample, therefore ratings of the subjects in a bootstrap-sample can be viewed as simultaneous realizations of identically and independently distributed (i.i.d.) random ratings, namely:

$$\left\{\left(X_{i`}^1, X_{i`}^2\right)\right\}_{i`=1}^{N} \overset{i.i.d.}{\sim} Uniform\left[\left\{\left(x_i^1, x_i^2\right)\right\}_{i=1}^{N}\right] \text{ and } \left\{\left(Y_{j`}^1, Y_{j`}^2\right)\right\}_{j`=1}^{M} \overset{i.i.d.}{\sim} Uniform\left[\left\{\left(y_j^1, y_j^2\right)\right\}_{j=1}^{M}\right]$$

After a bootstrap-sample is drawn it is used to compute the value of the primary statistic - nonparametric estimator of the AUC difference. This statistic depends on the ratings via the joint order indicators denoted by $\boldsymbol{w}$ and defined in II.A.3. The $\boldsymbol{w}_{ij}$ provides information on the difference in relative orders assigned to the pair of i^{th} normal and j^{th} abnormal subjects by two

modalities. The value of w_{ij} in a bootstrap-sample is uniformly distributed over all values of joint order indicators observed in the original data, i.e.:

$$W_{i`j`} \sim Uniform\left[\{w_{ij}\}_{i=1,j=1}^{N,M}\right]$$

In contrast to the random pairs of ratings, two random joint-order-indicators are not independent unless based on different subjects. However, covariances of two $\boldsymbol{W}$'s can be easily computed from the initially observed $\boldsymbol{N}$x$\boldsymbol{M}$ values (see derivations in Appendix B). Since the variance of the AUC difference can be expressed in terms of the covariances between two random joint-order-indicators (II.A.1) its exact variance in the bootstrap-space can be easily computed (Appendix B) resulting in the following formula:

$$V_B = \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{N^2 M^2}$$

The asymptotic bootstrap procedure for testing the difference between two AUCs in a paired design setting can be performed using the $\boldsymbol{Z}$- statistic:

$$Z = \frac{\hat{A}^1 - \hat{A}^2}{\sqrt{V_B(\hat{A}^1 - \hat{A}^2)}}$$

Its approximate normality (with mean 0 and variance 1) follows from the asymptotic normality of the nonparametric AUC estimator and the consistency of the bootstrap-variance.

B. SIMULATION STUDY

Using the derived formula for the bootstrap-variance we compare it to other estimators of the variance of nonparametric AUC difference. While some relationships between the various variance estimators are apparent from the formulas (Appendix C), the comparison between the bootstrap and jackknife variance estimators has to be done numerically. We performed simulations to investigate the properties of the bootstrap-variance and corresponding asymptotic test. The estimators of the variance compared include the two-sample jackknife ($\boldsymbol{V_{J2}}$) which is

equivalent to that proposed by DeLong *et al.* [19], the one-sample jackknife ($\boldsymbol{V_{J1}}$) which ignores the distinction between normal and abnormal subjects; and biased ($\boldsymbol{V_{Wb}}$) and unbiased ($\boldsymbol{V_W}$) estimators suggested by Wieand *et al.* [18]. The simulations follow the general approach described in Chapter III Section A. All figures illustrate the estimates computed for samples of size of 40 normal and 40 abnormal subjects, correlation between ratings ($\boldsymbol{\rho}$) of 0.4, shape parameter ($\boldsymbol{b}$) of 1 in both modalities. In addition, for Figure V.3.b the AUC of each modality is set equal to 0.85.

Figure V.1 illustrates the average variance estimates and their relative biases (percent of deviation from the empirical variance). The graph in Figure V.1.a indicates a strong decreasing relationship between the variance and average AUC

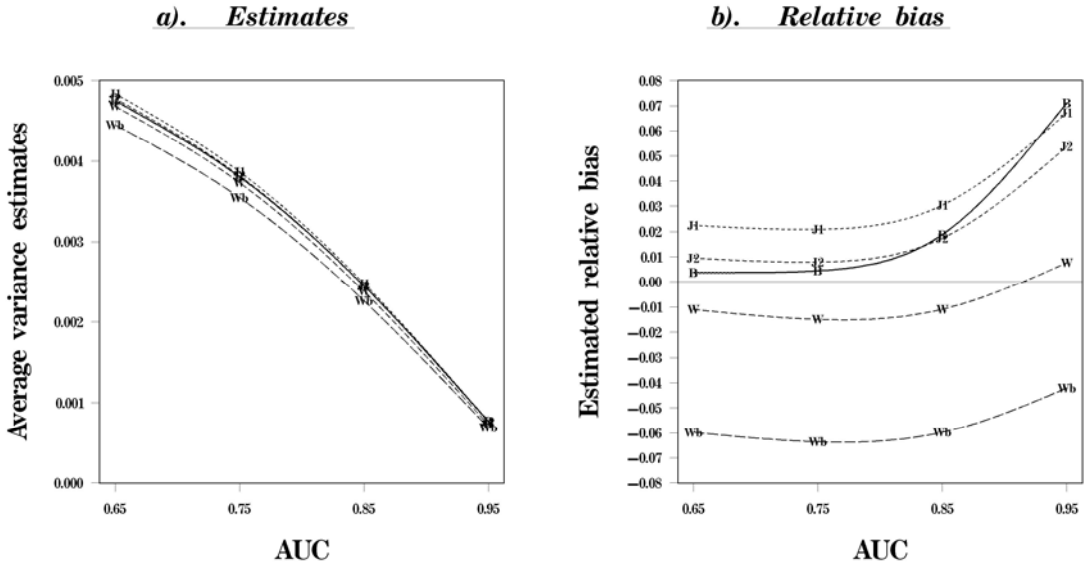

Figure V.1 **Expectations of the variance estimators**
Types of the variance estimators: Wb- Wieand (biased); W-Wieand (unbiased); J2-two-sample jackknife; J1-one-sample jackknife; B-bootstrap. Graph a): Average estimates of the variance; Graph b): Estimated relative bias of the estimates (percent of deviation from the empirical variance)

Figure V.1.b indicates that the bootstrap-variance ($\boldsymbol{V_B}$) has an upward bias that increases with increasing underlying AUC. The commonly used two-sample-jackknife-variance ($\boldsymbol{V_{J2}}$) demonstrates similar properties and the trend in upward bias is less sharp than the trend for the

bootstrap-variance. On average, however, the bootstrap-variance is much closer to the conventional estimator than to any other.

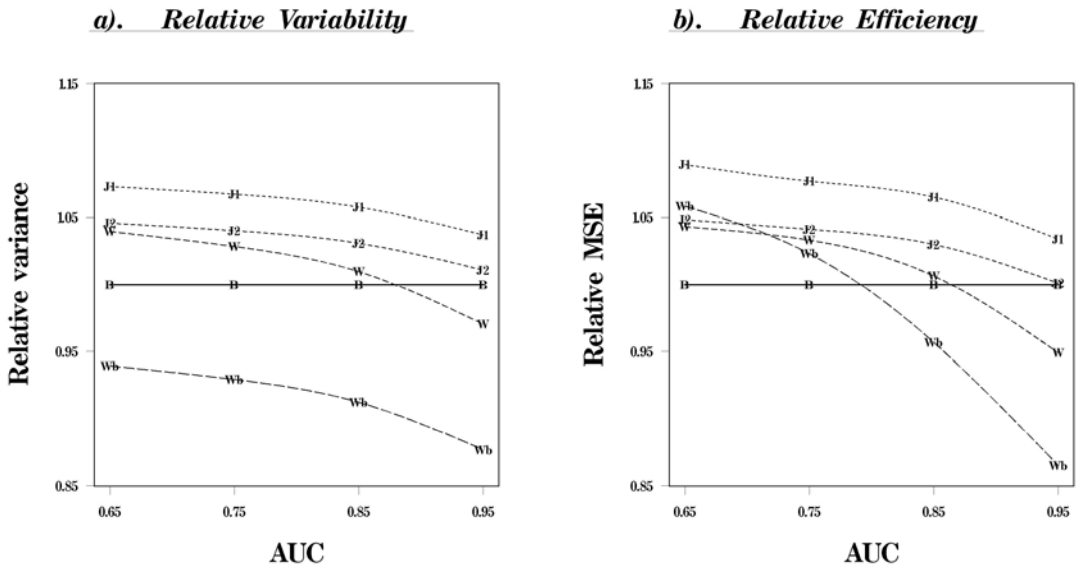

Figure V.2 **Efficiency of the variance estimators**
Types of the variance estimators: Wb- Wieand (biased); W- Wieand (unbiased); J2-two-sample jackknife; J1-one-sample jackknife; B-bootstrap. Graph a): Relative variability of the estimates (relative to the bootstrap); Graph b): Relative efficiency of the estimates (relative to bootstrap)

Figure V.2.a illustrates how variance estimators differ with respect to their variability. From this graph it can be seen that the variability of the bootstrap-variance ($\boldsymbol{V_B}$) is quite small and uniformly superior to both jackknife estimators. The biased estimator ($\boldsymbol{V_{Wb}}$) proposed by Wieand *et al.* has uniformly lower variance than the bootstrap estimator and the unbiased estimator ($\boldsymbol{V_W}$) has lower variance when AUC>85.

Since four out of five variance estimators demonstrate bias for some values of AUC we compare their efficiencies by considering the ratio of the "mean squared errors" (MSEs). Figure V.2.b demonstrates efficiencies of the estimators relative to that produced by the bootstrap approach. The bootstrap-variance ($\boldsymbol{V_B}$) has the mean squared error that is lower than that of the unbiased estimator ($\boldsymbol{V_W}$) when AUC is less than 0.85 and lower than that of the biased estimator ($\boldsymbol{V_{Wb}}$) proposed by Wieand *et al.* when AUC is less than 0.8. The efficiency of the bootstrap-variance is consistently better than that of the conventional variance estimator ($\boldsymbol{V_{J2}}$).

The results presented in Figure V.1 and Figure V.2 indicate an average superiority of the bootstrap-variance over the conventional two-sample jackknife estimator in terms of their proximity to the truth. We now directly compare the rejection rates of those statistical tests. Figure V.3, Table V.1 and Table V.2 illustrates the results of this part of the investigation. Graph a) and Table V.1 illustrate the relationship between the estimates of the type I error of different procedures and Graph b) and Table V.2 depict the statistical power. There appears to be little practical difference in the rejection rate of the asymptotic bootstrap and conventional tests, with discrepancies being consistent with those observed for variance estimators.

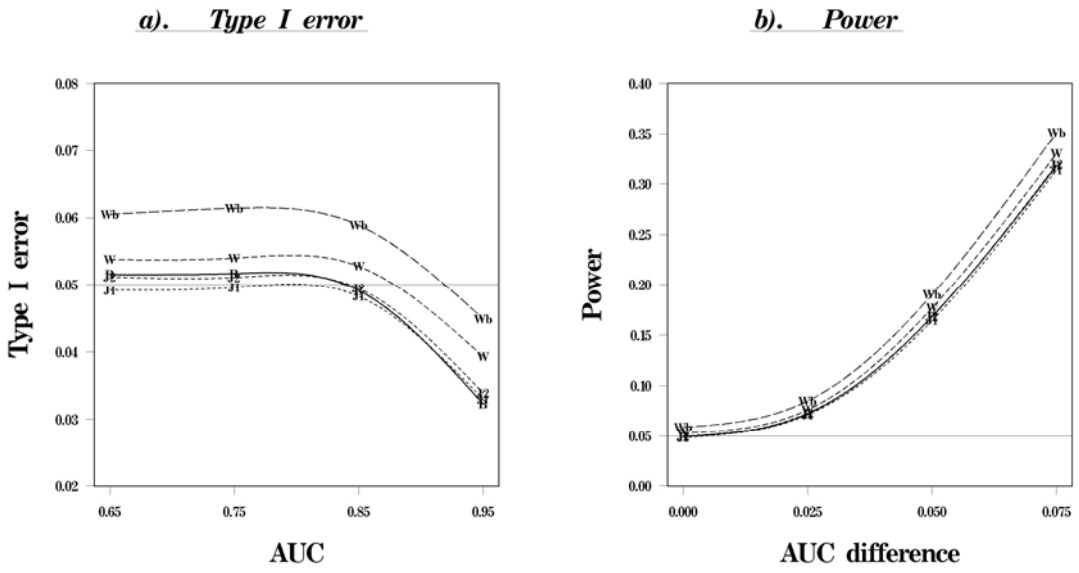

Figure V.3 **Rejection rates of asymptotic tests**

Types of the variance estimators: Wb- Wieand (biased); W-Wieand (unbiased); J2-two-sample jackknife; J1-one-sample jackknife; B-bootstrap.

Table V.1 **Bootstrap asymptotic test: type I error**

Correlation	AUC	Sample size: N=20 normal M=20 abnormal subjects				N=40 normal M=40 abnormal subjects				N=60 normal M=60 abnormal subjects			
		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1,b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1,b_2=1/2$)		The same ROC ($b_1=b_2=1$)		Crossing ROCs ($b_1=1,b_2=1/2$)	
		J2	**B**	**J2**	**B**	**J2**	**B**	**J2**	**B**	**J2**	**B**	**J2**	**B**
ρ=0.0	**0.65**	0.056	0.059	0.056	0.059	0.053	0.054	0.053	0.054	0.054	0.054	0.053	0.054
	0.75	0.052	0.054	0.057	0.062	0.051	0.052	0.052	0.053	0.051	0.052	0.051	0.052
	0.85	0.048	0.050	0.051	0.052	0.047	0.047	0.050	0.051	0.049	0.050	0.048	0.049
	0.95	0.018	0.019	0.020	0.020	0.038	0.037	0.043	0.043	0.044	0.043	0.046	0.046
ρ=0.4	**0.65**	0.054	0.055	0.054	0.056	0.051	0.052	0.053	0.055	0.054	0.054	0.053	0.054
	0.75	0.048	0.049	0.052	0.053	0.051	0.052	0.049	0.050	0.051	0.051	0.050	0.051
	0.85	0.038	0.039	0.044	0.045	0.050	0.049	0.050	0.050	0.053	0.053	0.045	0.046
	0.95	0.015	0.014	0.017	0.016	0.034	0.032	0.035	0.034	0.045	0.043	0.044	0.043
ρ=0.6	**0.65**	0.046	0.046	0.051	0.051	0.045	0.045	0.051	0.050	0.049	0.049	0.053	0.053
	0.75	0.046	0.046	0.047	0.047	0.053	0.053	0.049	0.049	0.054	0.054	0.045	0.045
	0.85	0.042	0.040	0.041	0.040	0.044	0.042	0.049	0.049	0.050	0.049	0.054	0.054
	0.95	0.011	0.009	0.015	0.014	0.030	0.027	0.036	0.034	0.037	0.035	0.048	0.046

Types of the variance-estimators: J2-two-sample jackknife; B-bootstrap.

Table V.2 **Bootstrap asymptotic test: statistical power**

Correlation	Average AUC	AUC difference	Sample size: N=20 normal and M=20 abnormal subjects		N=40 normal and M=40 abnormal subjects		N=60 normal and M=60 abnormal subjects	
			J2	B	J2	B	J2	B
ρ=0.0	0.65	0.025	0.064	0.066	0.061	0.062	0.066	0.067
		0.050	0.068	0.071	0.086	0.088	0.113	0.114
		0.075	0.102	0.106	0.145	0.147	0.196	0.198
	0.75	0.025	0.059	0.062	0.064	0.065	0.070	0.070
		0.050	0.077	0.081	0.104	0.106	0.124	0.125
		0.075	0.113	0.116	0.161	0.163	0.225	0.227
	0.85	0.025	0.056	0.058	0.070	0.071	0.079	0.080
		0.050	0.086	0.089	0.128	0.129	0.165	0.166
		0.075	0.135	0.140	0.239	0.241	0.336	0.339
	0.95	0.025	0.038	0.038	0.100	0.099	0.150	0.147
		0.050	0.106	0.106	0.316	0.314	0.486	0.484
		0.075	0.244	0.244	0.690	0.686	0.887	0.885
ρ=0.4	0.65	0.025	0.055	0.057	0.065	0.065	0.076	0.077
		0.050	0.083	0.086	0.115	0.116	0.147	0.148
		0.075	0.119	0.122	0.192	0.194	0.269	0.270
	0.75	0.025	0.058	0.058	0.069	0.069	0.080	0.080
		0.050	0.081	0.083	0.131	0.131	0.173	0.173
		0.075	0.133	0.135	0.227	0.227	0.321	0.322
	0.85	0.025	0.057	0.057	0.073	0.072	0.092	0.092
		0.050	0.094	0.094	0.168	0.168	0.237	0.236
		0.075	0.161	0.162	0.319	0.318	0.460	0.459
	0.95	0.025	0.033	0.033	0.115	0.111	0.181	0.176
		0.050	0.112	0.111	0.399	0.393	0.614	0.609
		0.075	0.280	0.278	0.789	0.784	0.950	0.949
ρ=0.6	0.65	0.025	0.060	0.059	0.073	0.073	0.086	0.085
		0.050	0.090	0.091	0.143	0.142	0.188	0.188
		0.075	0.139	0.139	0.251	0.250	0.362	0.361
	0.75	0.025	0.056	0.055	0.077	0.076	0.090	0.089
		0.050	0.096	0.094	0.156	0.155	0.214	0.213
		0.075	0.162	0.161	0.307	0.306	0.438	0.437
	0.85	0.025	0.055	0.054	0.080	0.079	0.108	0.107
		0.050	0.101	0.099	0.213	0.209	0.307	0.303
		0.075	0.203	0.199	0.431	0.426	0.601	0.598
	0.95	0.025	0.031	0.028	0.138	0.132	0.223	0.216
		0.050	0.120	0.113	0.495	0.481	0.735	0.728
		0.075	0.324	0.317	0.870	0.863	0.983	0.982

Types of the variance-estimators: J2-two-sample jackknife; B-bootstrap.

C. SUMMARY AND DISCUSSION

We have derived a closed-form solution for the bootstrap-variance of the nonparametric estimator of AUC difference. Availability of such an estimator allows for construction of an easy-to-implement asymptotic bootstrap test as well as alleviating the computational burden.

The results of our simulation study indicate that the bootstrap-variance provides a good estimate of the true variability. Among the estimators we considered it is the most efficient for AUCs lower than 0.8 but less efficient than both estimators proposed by Wieand *et al.* [18] for larger AUCs. The bootstrap estimator also has an upward bias which increases with increasing average AUC. Compared to the conventional two-sample-jackknife [19] the bootstrap estimator of the variance is more efficient but has greater bias for large AUCs (>0.85). Both estimators proposed by Wieand *et al.* [18] perform well. The biased estimator has lower mean squared error (MSE) than that of the bootstrap-variance for AUC>0.80 and the unbiased version of the estimator has lower MSE for AUC>0.85 (the low MSE of the biased estimator is perhaps due to its low variance).

For small AUCs the asymptotic bootstrap test, compared to the conventional procedure, has even more elevated type I error and, perhaps because of that, is slightly more powerful. For large AUCs the bootstrap-based test is more conservative than the conventional test implying even greater potential for the loss of the statistical power. Thus, although the bootstrap might offer a better way to estimate the variability than the conventional two-sample jackknife approach it leads to an asymptotic test with slightly inferior small-sample properties. However, our simulations indicate no practical difference between bootstrap and conventional test.

VI. CONDITIONAL TEST

In this chapter we develop a novel approach for statistical comparison of the overall performance of the two modalities in a paired design setting. The motivation for this approach was dependent on two factors. First, the AUC can be viewed as a simple function of the relative orderings of all pairs of normal and abnormal subjects. Secondly, the difference in the AUCs for two modalities in a paired design depends only on those pairs where the relative orderings of normal and abnormal cases differ. The corresponding statistical test is similar in spirit to McNemar's procedure [44] which conducts the analysis only on discordant pairs. Simulations are conducted to verify the small-samples properties of the conditional test. This part of the research is published in Academic Radiology [33].

A. CONDITIONAL APPROACH

In general, the ratings assigned to a randomly selected pair of normal and abnormal cases can only be in one of three possible orderings: $X < Y$, $X = Y$, $X > Y$ (i.e. normal case having score lower than, equal to, or higher than the score for the abnormal case). Each of these possibilities represents different degrees to which a modality (including the observer) can distinguish between a given set of actually positive and actually negative findings, namely:

$X < Y$ - the modality correctly discriminates between given cases;

$X = Y$ - the modality does not discriminate between given cases;

$X > Y$ - the modality incorrectly discriminates the between given cases.

In a paired design when the same cases are evaluated by two modalities the nine possible joint orderings between the ratings of the normal and abnormal cases can be classified as follows:

(VI.A.1)

	$X^2 < Y^2$	$X^2 = Y^2$	$X^2 > Y^2$
$X^1 < Y^1$	0	+	+
$X^1 = Y^1$	−	0	+
$X^1 > Y^1$	−	−	0

In the table above, the '+' indicates a combination of orderings implying the 1st modality is superior to the 2nd, '-' indicates a combination of orderings implying the 2nd modality is superior to the 1st modality, and "0" indicates a combination of orderings suggesting equivalence between the two modalities with respect to their ability to correctly discriminate between cases with and without the abnormality. In this work the orderings ('+' or '-') that contribute to the determination of which modality is superior are naturally termed "discordant" orderings while the others ('0') are termed "concordant".

The overall ability of the modality to identify the abnormality can be viewed as the AUC. The difference between the overall performance levels of two modalities can therefore be summarized as the difference between two AUCs:

(VI.A.2) $$A^1 - A^2 = P\left(X^1 < Y^1\right) + \tfrac{1}{2}P\left(X^1 = Y^1\right) - P\left(X^2 < Y^2\right) - \tfrac{1}{2}P\left(X^2 = Y^2\right)$$

The estimator of the difference above is written in terms of probabilities of "marginal" orderings (orderings that corresponds to rows and columns of table VI.A.1). Alternatively, in a paired design setting we can express this difference in terms of probabilities of discordant joint orderings. Namely, by replacing each probability of the "marginal" ordering in VI.A.2 with the sum of joint probabilities (probabilities of corresponding cells in table VI.A.1) and canceling common terms, one obtains the following expression directly:

(VI.A.3) $$\begin{aligned} A^1 - A^2 = {} & P\left(X^1 < Y^1, X^2 > Y^2\right) + \tfrac{1}{2}P\left(X^1 < Y^1, X^2 = Y^2\right) + \tfrac{1}{2}P\left(X^1 = Y^1, X^2 > Y^2\right) \\ & - P\left(X^2 < Y^2, X^1 > Y^1\right) - \tfrac{1}{2}P\left(X^2 < Y^2, X^1 = Y^1\right) - \tfrac{1}{2}P\left(X^2 = Y^2, X^1 > Y^1\right) \end{aligned}$$

In any given dataset, the probabilities of the joint discordant orderings contain all the needed information in order to quantify the differences in the area under the two ROC curves orderings. Motivated by this observation we construct a statistical test conditional on the pairs of cases with observed discordant orderings.

Note that for truly continuous ratings (i.e. when no ties are possible) the difference between two AUC in a paired design can be equivalently written as:

$$A^1 - A^2 = P\left(X^1 < Y^1, X^2 > Y^2\right) - P\left(X^2 < Y^2, X^1 > Y^1\right)$$

Since, in the continuous case:

$$D = \left\{(X,Y) : (X^1 < Y^1, X^2 > Y^2) \; or \; (X^1 > Y^1, X^2 < Y^2)\right\}$$

then:

$$A^1 - A^2 = P\left(X^1 < Y^1 \mid (X,Y) \in D\right) - P\left(X^2 < Y^2 \mid (X,Y) \in D\right) =$$
$$= 2P\left(X^1 < Y^1 \mid (X,Y) \in D\right) - 1$$

and the hypothesis of equality between two AUCs is equivalent to the hypothesis:

$$P\left(X^1 < Y^1 \mid (X,Y) \in D\right) = \tfrac{1}{2}.$$

B. CONDITIONAL PERMUTATION TEST

The test we propose makes use of a nonparametric estimator of the AUC difference and is based on the concept of estimating the variability of the sum of discordant order indicators (structural elements of the nonparametric AUC difference estimator). Namely, in the underlying sample space the initial discordant orderings may correspond to a different "degree" (level) of superiority of one of the modalities. In this sample space we estimate the variability of the sum of discordant ordering indicators:

(VI.B.1) $$W = \sum_{i,j:w_{ij} \neq 0} W_{ij}$$

Note that the quantity $\boldsymbol{W_{ij}}$ as defined in (II.A.3) differentiates the possible orderings for the pair of $\boldsymbol{i}^{\text{th}}$ normal and $\boldsymbol{j}^{\text{th}}$ abnormal cases (VI.A.1) in the following manner:

(VI.B.2)

W_{ij}	$X_i^2 < Y_j^2$	$X_i^2 = Y_j^2$	$X_i^2 > Y_j^2$
$X_i^1 < Y_j^1$	0	½	1
$X_i^1 = Y_j^1$	−½	0	½
$X_i^1 > Y_j^1$	−1	−½	0

We propose conditioning on the set of discordant orderings, $\boldsymbol{D}$. (non-zero $\boldsymbol{W_{ij}}$ in the data remain non-zero for all permutations). The variance of $\boldsymbol{W}$ can be expressed through the variances of each of the $\boldsymbol{W_{ij}}$ and the covariances between any pair of these. Thus, to compute the variance of the $\boldsymbol{W}$ we need moments of the joint distribution of any pair of $\{W_{ij}\}$. To obtain these moments we model the marginal joint distribution of any pair of $\{W_{ij}\}$.

Although, in general the joint distribution of any pair of $\{W_{ij}\}$ is difficult to derive, the use of the permutation test provides a direct way to estimate the needed distributions [32]. Following the approach used in [24] the rank-ratings (ranks of the rating) are used to generalize the inferences to a situation where the two modalities may have different underlying rating scales. If the two modalities are assumed to have the same underlying scales the actual ratings (not the rank ratings) should be used. The basic permutation space $\boldsymbol{\Omega}$ is created by permuting the case-specific rank-ratings between the two modalities within the structure of given pairs.

Thus, if for the pair consisting of the $\boldsymbol{i}^{\text{th}}$ normal case and the $\boldsymbol{j}^{\text{th}}$ abnormal case we observe rank-ratings of $\boldsymbol{x_i^1}$ and $\boldsymbol{y_j^1}$ for the first modality and $\boldsymbol{x_i^2}$ $\boldsymbol{y_j^2}$ for the second modality then the four permutations within this paired structure are:

$$\begin{aligned}
&(x_i^1, y_j^1) \quad , \quad (x_i^2, y_j^2) \\
&(x_i^2, y_j^1) \quad , \quad (x_i^1, y_j^2) \\
&(x_i^1, y_j^2) \quad , \quad (x_i^2, y_j^1) \\
&(x_i^2, y_j^2) \quad , \quad (x_i^1, y_j^1)
\end{aligned}$$

where the first pair is assumed to be observed for the first modality and the second pair is assumed to be observed for the second modality. We are assuming that under the null hypothesis the ranks for normal (abnormal) cases satisfy the statistical assumption of exchangeability implying that each of these four situations is equally likely. Let $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ be the score as assigned in II.A.3 for each of the four possible permutations associated with the $\boldsymbol{i^{th}}$ normal and $\boldsymbol{j^{th}}$ abnormal pair of cases. The value w_{ij}^{11} represents the value of $\boldsymbol{W_{ij}}$ associated with the observed data. A superscript of $p=2$ ($q=2$) corresponds to a permutation where the normal (abnormal) rank-ratings of the two modalities are interchanged.

Assuming equal probability of each permutation under the null hypothesis, each $\boldsymbol{W_{ij}}$ is uniformly distributed over the values $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ which belong to the set $\{-1, -½, 0, ½, 1\}$ and possess the anti-symmetric property: $w_{ij}^{11} = -w_{ij}^{22}$, $w_{ij}^{12} = -w_{ij}^{21}$.

The pairs of normal and abnormal cases with initially observed discordant order of ratings are distinguished by the condition $w_{ij}^{11} \neq 0$ and the random quantity in VI.B.1 can be written as:

$$W = \sum_{i,j:w_{ij}^{11}\neq 0} W_{ij} \ .$$

Constraining the discordant orderings to remain discordant (i.e. conditioning on $\boldsymbol{D}$) results in $\boldsymbol{W_{ij}}$ that are uniformly distributed over non-zero values from $\left\{w_{ij}^{p,q}\right\}_{p=1,q=1}^{2,2}$.

The set of values $\left\{w_{ij}^{p,q}\right\}_{i=1,j=1\,p=1,q=1}^{N,M,2,2}$ can be obtained by determining the relative orderings between rank-ratings for every normal and abnormal case and then taking the differences in a manner consistent with II.A.3. The moments of the $\left\{W_{ij}\right\}_{i,j:w_{ij}^{11}\neq 0}$ can be computed from the set of values $\left\{w_{ij}^{p,q}\right\}_{i=1,j=1\,p=1,q=1}^{N,M,2,2}$ using the formulae shown in the Appendix D. Finally, to implement the conditional test of equality of AUC between two modalities in a paired design we propose comparing the statistic $\dfrac{W}{\sqrt{Var_{\Omega}(W \mid D)}}$ to the pre-specified normal percentile.

C. SIMULATION STUDY

To verify the validity of the proposed test we performed simulations to investigate its type I error for different underlying AUCs, correlations between case ratings across modalities and different sample sizes. In our simulations we assume equal correlation across modalities for the ratings of normal and abnormal cases and consider scenarios with non-crossing as well as crossing ROC curves.

The general protocol of simulations follows the general approach described in Chapter II, Section A. In addition to simulations of conventional datasets, we conducted simulations where the samples from the binormal distribution ("typical" cases) were enriched with "easy" and "difficult" cases. Such enrichment has the practical effect of increasing the number of concordant pairs (i.e. pairs where there is an agreement for both modalities). When the concordance level is high the method by DeLong *et al*. has been shown to have a below nominal level (0.05) type I error [Chapter III,31]. Hence in these situations the conventional nonparametric test may have unnecessarily low statistical power.

The ratings of “easy” and “difficult” cases were defined relative to the ratings of cases in the other groups. Namely, “easy” normal cases are rated by both modalities below that of any “typical” abnormal cases but might be higher than some of the “difficult” abnormal cases; the “difficult” normal cases are those rated by both modalities higher than the “typical” abnormal cases but might be rated lower than “easy” abnormal cases; and the “typical” normal and abnormal cases have generally overlapping range of ratings. The fixed number of cases of each type was simulated from normal distributions in a manner that easy normal/abnormal cases had the same distributions as difficult abnormal/normal cases (e.g. completely missed abnormal cases were considered to be rated similar to easy normal cases). The distributions of easy and difficult cases were sufficiently different than the distribution for typical cases to prevent overlapping.

Table VI.1 illustrates that the rejection rate of the proposed procedure is generally close to the nominal level of 0.05 while being conservative for small sample sizes, large AUCs, and high correlation between ratings of the same case.

Table VI.1 **Conditional test: rejection rate**

	AUC	N=20 normal and M=20 abnormal subjects				N=40 normal and M=40 abnormal subjects				N=60 normal and M=60 abnormal subjects			
		ρ=0.0	**ρ=0.2**	**ρ=0.4**	**ρ=0.6**	**ρ=0.0**	**ρ=0.2**	**ρ=0.4**	**ρ=0.6**	**ρ=0.0**	**ρ=0.2**	**ρ=0.4**	**ρ=0.6**
The same ROC ($b_1=b_2=1$)	**0.60**	0.042	0.041	0.041	0.038	0.045	0.043	0.046	0.045	0.049	0.050	0.049	0.048
	0.65	0.043	0.040	0.040	0.039	0.046	0.046	0.046	0.044	0.050	0.050	0.049	0.049
	0.70	0.042	0.041	0.040	0.038	0.046	0.046	0.046	0.043	0.047	0.049	0.048	0.048
	0.75	0.041	0.038	0.038	0.036	0.046	0.046	0.045	0.043	0.047	0.046	0.046	0.046
	0.80	0.040	0.038	0.036	0.034	0.046	0.046	0.045	0.044	0.048	0.048	0.047	0.044
	0.85	0.036	0.036	0.034	0.030	0.043	0.043	0.043	0.045	0.047	0.047	0.048	0.045
	0.90	0.033	0.031	0.029	0.025	0.042	0.041	0.040	0.040	0.046	0.045	0.045	0.043
	0.95	0.018	0.017	0.015	0.015	0.036	0.036	0.033	0.033	0.043	0.043	0.040	0.040
Crossing ROC ($b_1=1$, $b_2=1/2$)	**0.60**	0.047	0.047	0.045	0.043	0.051	0.051	0.050	0.050	0.052	0.050	0.051	0.052
	0.65	0.047	0.044	0.043	0.043	0.050	0.049	0.049	0.048	0.051	0.051	0.052	0.053
	0.70	0.047	0.044	0.042	0.041	0.051	0.051	0.050	0.049	0.052	0.051	0.052	0.052
	0.75	0.046	0.043	0.044	0.039	0.051	0.049	0.049	0.048	0.052	0.052	0.052	0.051
	0.80	0.044	0.041	0.040	0.038	0.050	0.048	0.050	0.049	0.052	0.051	0.049	0.050
	0.85	0.041	0.039	0.037	0.035	0.050	0.048	0.048	0.049	0.051	0.050	0.049	0.049
	0.90	0.035	0.034	0.030	0.028	0.048	0.048	0.047	0.048	0.049	0.049	0.047	0.046
	0.95	0.018	0.017	0.016	0.015	0.040	0.040	0.040	0.040	0.045	0.046	0.044	0.043

We also compared the power of the proposed conditional procedure to that of the conventional nonparametric procedure developed by DeLong *et al.* [19]. For the binormal ("typical" cases only) datasets the conventional AUC test is somewhat more powerful than the conditional procedure. For example, for 20 normal and 20 abnormal cases that form non-crossing ROC curves with AUCs of 0.75 and 0.85, and a correlation of 0.4 between ratings on the same cases, the conventional AUC test has power and type I error correspondingly of 0.223 and 0.046 as compared with 0.191 and 0.036 for the conditional procedure proposed here. However, the presence of small number of "easy" and "difficult" cases may result in an advantage for the proposed conditional test in datasets with 20 "typical" normal and 20 "typical" abnormal cases. The estimates of the power for the two procedures in "enriched" datasets are presented in Table VI.2.

Table VI.2 **Conditional test: statistical power in the "enriched" datasets**

Average AUC	AUC difference	ρ=0.0		ρ=0.2		ρ=0.4		ρ=0.6	
		DeLong	Conditional	DeLong	Conditional	DeLong	Conditional	DeLong	Conditional
0.75	**0.1**	0.118	0.122	0.129	0.136	0.151	0.162	0.191	0.209
	0.2	0.380	0.394	0.436	0.454	0.526	0.545	0.658	0.685
0.8	**0.1**	0.128	0.135	0.143	0.153	0.167	0.181	0.208	0.233
	0.2	0.441	0.463	0.503	0.527	0.583	0.611	0.712	0.742
0.85	**0.1**	0.148	0.161	0.165	0.181	0.189	0.211	0.244	0.275
	0.2	0.537	0.576	0.597	0.634	0.671	0.713	0.773	0.816

"Enriched" datasets include: 20" typical" normal + 20 "typical" abnormal, 10 "easy" normal + 10 "easy" abnormal and 3 "difficult" normal +3"difficult" abnormal subjects.

D. SUMMARY AND DISCUSSION

The proposed procedure illustrates a conceptually new approach to testing the equality of overall diagnostic performances between two modalities in a paired design setting. Using the nature of a paired design and relative orderings we introduced the idea of concordances and discordances in the task of comparing overall performances of diagnostic systems. Conditioning on the discordant order indicators resulted in a test similar in spirit to McNemar's test. However the

complex correlation structure of the discordant order indicators prevents construction of an exact procedure and greatly complicates the process of developing an asymptotic test. The estimator of the variance used in our method is derived using the assumption of exchangeability of the case-specific rank-ratings under the null hypothesis. The condition of exchangeability is stricter than the condition of equality of AUCs and implies an equality of the two ROC curves. However, our computer simulations indicate that the rejection rate of the proposed test remains close to nominal significance level even in cases of substantially crossing ROC curves ($\boldsymbol{b_1}$=1, $\boldsymbol{b_2}$=1/2), at least for moderate sample sizes**,** hence rejection of the null hypothesis is unlikely to occur unless there is a difference in the AUCs.

A substantial number of concordances may occur in a screening population where there may be a substantial number of "easy" or "difficult" cases or in laboratory experiments where the method of selection of cases could result in a higher level of concordance. In datasets with ratings that can be monotonically transformed to a binormal distribution the number of "concordant" orderings increases with increasing AUC and correlation between ratings in different modalities. This may explain in part the conservative behavior of the conventional nonparametric AUC test [19]. It should be noted however that the impact of "concordant" orderings on the efficiency of the conditional procedure is most evident for reasonably small sample sizes (less than 60 normal and 60 abnormal).

In conclusion, we presented a conceptually new approach to the assessment of differences between two diagnostic modalities in a paired design. This method, which is conditioned on discordances in discrimination between normal and abnormal cases in the two modalities, may provide advantages in relatively small studies where the selection of cases results in a high level of concordance.

VII. CONCLUSIONS AND DISCUSSION

In this work we investigated the effects of various parameters on the small-sample properties of the conventional nonparametric procedure for comparing correlated AUCs and developed three novel nonparametric approaches for comparing two diagnostic modalities in a paired design setting. The conducted research provides important information and methods that can be useful for study design and choice of an appropriate statistical method for analysis. Also the proposed statistical approaches create a solid foundation for further development of nonparametric methods and the results of simulation studies may offer guidelines for more complex scenarios.

In our study of the properties of the conventional procedure for nonparametric comparison of the correlated AUCs we attempted to characterize the effect of various parameters on the statistical hypothesis testing with small samples. For each parameter we described the direction and relative magnitude of the effect on the type I error and power of a statistical test. The parameters we identified as having an effect are (in decreasing order of influence): average AUC ($\boldsymbol{A}$), correlation between modalities ($\boldsymbol{\rho}$), and the prevalence of abnormal subjects in the selected sample ($\boldsymbol{p}$).

The proposed permutation procedure for comparing two diagnostic systems provides the ability to perform the exact test for small samples and the asymptotic test for larger ones. The easy-to-implement asymptotic test offers an excellent approximation of the exact procedure even for sample sizes as low as 6 normal and 6 abnormal subjects. The quality of approximation can be attributed in part to the exact nature of the variance-estimator used in the construction of the asymptotic test and to the symmetry of the permutation distribution of the nonparametric estimator of AUC difference. The developed procedure is based on all permutations of the subject specific rank ratings and is formally a test for equality of ROC curves that is sensitive to the alternatives of AUC difference. For small samples and for underlying parameters that are common in experimental studies in the field of diagnostic test evaluation (AUC of more than

0.75, correlation of more than 0.4) the permutation test possesses good operating characteristics and is more powerful than the conventional nonparametric procedure for AUC comparisons.

Exploiting the properties of the nonparametric estimator of AUC difference we derived a closed-form solution for the bootstrap-variance and constructed an easy-to-implement asymptotic test. The results of our simulation study indicate that the bootstrap-variance is uniformly more efficient than the conventional two-sample-jackknife estimator; however it has a higher bias for large AUCs. Also for small AUCs the bootstrap-variance was shown to have a relatively small bias and the best efficiency among considered estimators. It is worth noting that we measure the efficiency by the "mean squared error" (MSE) what induces a specific type of tradeoff between bias and variability; hence it is possible that under a different measure (for instance absolute instead of squared distance) the relative efficiencies of the estimators will change. Despite its good properties the bootstrap variance leads to an asymptotic test with small-sample properties slightly inferior to that of the conventional procedure developed by DeLong *et al.* [19]. In conclusion, for the nonparametric estimator of the AUC difference, the bootstrap approach might offer a better estimator of the variability than the conventional two-sample jackknife procedure; however it does not produce a better asymptotic test.

Using the relationship of the AUC difference to the relative orderings of the ratings assigned to pairs of normal and abnormal subjects by two modalities we introduced the concept of concordances and discordances in the task of comparing overall performances of diagnostic systems with paired data. Conditioning on the discordant order indicators resulted in a test similar in spirit to McNemar's test. However the complex correlation structure of the discordant order indicators prevents construction of an exact procedure and greatly complicates the process of developing an asymptotic test. The problem of constructing the asymptotic test was solved with the help of the previously developed permutation test. The type I error of the procedure for small samples was verified using computer simulations. The conditional nonparametric test presented here is an alternative approach to existing unconditional procedures and may offer statistical advantages in the presence of highly concordant data.

In developing the permutation approach we have restricted our attention to comparing two diagnostic modalities with paired data where the primary summary statistic is the area under the ROC curve. As mentioned previously the permutation approach can be applied to the comparison

of two diagnostic systems evaluated on independent datasets as well. Furthermore, if the data is paired but incomplete, our test could be modified using an approach similar to the one proposed by Zhou and Gatsonis [29] for correcting the conventional approach of DeLong *et al*. [19]. The permutation approach can also be applied either when different methods of estimating the area under the ROC curve are employed or when different summary statistics are used [7,8,14,24]. Although the permutation test might be used with various test statistics the computation time for the exact test might increase for some of them. Furthermore it may be impossible to derive exact permutation moments as was done for the nonparametric estimator of the difference in AUCs and thus there may be no simple approximation to the exact test.

An alternative summary statistic that is often used is the partial area under the ROC curve [20,28]. In theory, the permutation approach could be applied to this summary statistic although several issues need to be addressed. Some of these issues are discussed in [28] where the authors attempted to compare the partial areas under two ROC curves by modifying the conventional approach of DeLong *et al*. [19].

The permutation approach we have used might be also applicable to a more general approach of comparing diagnostic systems than ROC curve analysis. Bunch *et al*. [45] proposed a Free-response Receiver Operating Characteristic (FROC) curve which describes the task of detection and localization of multiple abnormalities per image. Although some work has been done addressing the comparison of FROC curves [46,47], the statistical methodology has been much less developed than for ROC curves. The permutation approach could circumvent some of the problems encountered in FROC analysis such as potential correlation between the multiple observations per image.

Other directions of future research include extension of the proposed procedures to accommodate the "multiple-reader" setting – a commonly used design in which several readers evaluate selected cases using different modalities. The random effects models in the multiple-reader settings offer another area of possible development [34,35,36,37].

APPENDIX A

PERMUTATION TEST: EXACT VARIANCE

In the permutation sample space, **Ω**, the exact mean and variance of the distribution of the difference between two AUC can be found. To simplify the derivations, consider the distribution of the random variables $\{W_{ij}\}_{i=1,j=1}^{N.M}$ defined over the set of all permutations by definition II.A.3.

Assuming equal probability of all permutations the random variable W_{ij} is uniformly distributed over the four possible values $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ defined as:

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

In other words $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ are the scores assigned by II.A.3 for each of the four possible permutations associated with the $\boldsymbol{i}^{\text{th}}$ normal and $\boldsymbol{j}^{\text{th}}$ abnormal pair of subjects (as illustrated in IV.A.1). The value w_{ij}^{11} represents the value of $\boldsymbol{W_{ij}}$ associated with the observed data. A superscript of $p=2$ ($q=2$) corresponds to a permutation where the normal (abnormal) rank-ratings of the two modalities are interchanged.

For example if we observe rank-ratings 1, 2 for the $\boldsymbol{i}^{\text{th}}$ normal subject in the first and the second modalities and corresponding rank-ratings 2, 1 for the $\boldsymbol{j}^{\text{th}}$ abnormal subject, then the four possible values $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ are:

$$\begin{aligned}
w_{ij}^{11} &= \psi\left(x_i^1, y_j^1\right) - \psi\left(x_i^2, y_j^2\right) = \psi(1,2) - \psi(2,1) = 1 - 0 = 1 \\
w_{ij}^{21} &= \psi\left(x_i^2, y_j^1\right) - \psi\left(x_i^1, y_j^2\right) = \psi(2,2) - \psi(1,1) = \tfrac{1}{2} - \tfrac{1}{2} = 0 \\
w_{ij}^{12} &= \psi\left(x_i^1, y_j^2\right) - \psi\left(x_i^2, y_j^1\right) = \psi(1,1) - \psi(2,2) = \tfrac{1}{2} - \tfrac{1}{2} = 0 \\
w_{ij}^{22} &= \psi\left(x_i^2, y_j^2\right) - \psi\left(x_i^1, y_j^1\right) = \psi(2,1) - \psi(1,2) = 0 - 1 = -1
\end{aligned}$$

Note also that the set $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ naturally possesses certain anti-symmetric properties, namely:

$$\forall i=1,...,N,\ j=1,...,M \quad \forall p,q=1,2 \quad w_{ij}^{p,q} \in \{-1,-\tfrac{1}{2},0,\tfrac{1}{2},1\}, \quad w_{ij}^{11}=-w_{ij}^{22}, \quad w_{ij}^{12}=-w_{ij}^{21}.$$

The values $\{w_{ij}^{p,q}\}_{p=1,q=1}^{2,2}$ for each pair of normal-abnormal subjects $\{w_{ij}^{p,q}\}_{i=1,j=1\,p=1,q=1}^{N,M,2,2}$ can be obtained by comparing each available rank-rating for a normal subject with the rank-ratings of every abnormal one using II.A.3.

Thus, the distribution of W_{ij} can be summarized as:

$$\forall i=1,...,N;\ j=1,...,M \quad W_{ij} \in \{w_{ij}^{11},w_{ij}^{12},w_{ij}^{21},w_{ij}^{22}\} \quad P(W_{ij}=w_{ij}^{p,q})=\tfrac{1}{4} \quad \forall p,q=1,2.$$

The joint distribution of any two W_{ij} with the same subscripts follows naturally from II.A.3 and the permutation algorithm. For example, the joint distribution of Ws sharing the same normal subjects (W_{ij},W_{il}) can be summarized as follows:

$$\forall i=1,...,N;\ j=1,...,M;\ l\neq j \quad (W_{ij},W_{il}) \in \left\{\begin{matrix}(w_{ij}^{11},w_{il}^{11}),(w_{ij}^{11},w_{il}^{12}),(w_{ij}^{12},w_{il}^{11}),(w_{ij}^{12},w_{il}^{12}),\\(w_{ij}^{21},w_{il}^{21}),(w_{ij}^{21},w_{il}^{22}),(w_{ij}^{22},w_{il}^{21}),(w_{ij}^{22},w_{il}^{22})\end{matrix}\right\}$$

$$P[(W_{ij},W_{il})=(w_{ij}^{p,q},w_{il}^{p,o})]=\tfrac{1}{8} \quad \forall p,q,o=1,2$$

We can now derive the mean and covariance structure of the set of the random variables $\{W_{ij}\}_{i=1,j=1}^{N,M}$. These are as follows:

$$\forall i,k=1,...,N;\ j,l=1,...,M;\ k\neq i;\ l\neq j \quad ,$$

$$E_{\Omega}(W_{ij})=\sum_{p=1}^{2}\sum_{q=1}^{2}w_{ij}^{p,q}/4=\bar{w}_{ij}^{\bullet\bullet}=\frac{(w_{ij}^{11}+w_{ij}^{22})+(w_{ij}^{12}+w_{ij}^{21})}{4}=0,$$

$$Var_{\Omega}(W_{ij})=E_{\Omega}(W_{ij}^{2})-E_{\Omega}(W_{ij})^{2}=\sum_{p=1}^{2}\sum_{q=1}^{2}\frac{(w_{ij}^{p,q})^{2}}{4}=\frac{(w_{ij}^{11})^{2}+(w_{ij}^{12})^{2}}{2},$$

and

$$Cov_{\Omega}(W_{ij},W_{il})=E_{\Omega}(W_{ij}W_{il})-E_{\Omega}(W_{ij})E_{\Omega}(W_{il})=$$

$$= \sum_{p=1}^{2}\sum_{q=1}^{2} \frac{w_{ij}^{p,q} w_{il}^{p,q} + w_{ij}^{p,q} w_{il}^{p,3-q}}{8} = \frac{\left(w_{ij}^{11} + w_{ij}^{12}\right) \times \left(w_{il}^{11} + w_{il}^{12}\right)}{4} = \overline{w}_{ij}^{1\bullet} \times \overline{w}_{il}^{1\bullet}$$

$$Cov_{\Omega}(W_{ij}, W_{kj}) = E_{\Omega}(W_{ij} W_{kj}) - E_{\Omega}(W_{ij}) E_{\Omega}(W_{kj}) =$$

$$= \sum_{p=1}^{2}\sum_{q=1}^{2} \frac{w_{ij}^{p,q} w_{kj}^{p,q} + w_{ij}^{p,q} w_{kj}^{3-p,q}}{8} = \frac{\left(w_{ij}^{11} + w_{ij}^{21}\right) \times \left(w_{kj}^{11} + w_{kj}^{21}\right)}{4} = \overline{w}_{ij}^{\bullet 1} \times \overline{w}_{kj}^{\bullet 1}$$

$Cov_{\Omega}(W_{ij}, W_{kl}) = 0$.

Using the derived formulae for the moments of $\left\{W_{ij}\right\}_{i=1,j=1}^{N,M}$, the moments of the estimator of the difference between AUC in Ω are as follows:

$$E_{\Omega}(\hat{A}^1 - \hat{A}^2) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} E_{\Omega}(W_{ij})}{NM} = 0$$

$$Var_{\Omega}(\hat{A}^1 - \hat{A}^2) = \frac{Var_{\Omega}\left(\sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}\right)}{(NM)^2} = \frac{1}{(NM)^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M} Var_{\Omega}(W_{ij}) + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\substack{l=1,\\ l\neq j}}^{M} Cov_{\Omega}(W_{ij}, W_{il}) + \right.$$

$$\left. + \sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{\substack{k=1,\\ k\neq i}}^{N} Cov_{\Omega}(W_{ij}, W_{kj}) + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\substack{k=1,\\ k\neq i}}^{N}\sum_{\substack{l=1,\\ l\neq j}}^{M} Cov_{\Omega}(W_{ij}, W_{kl})\right) =$$

$$= \frac{1}{(NM)^2}\left(\sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\left(w_{ij}^{11}\right)^2 + \left(w_{ij}^{12}\right)^2}{2} + \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\substack{l=1,\\ l\neq j}}^{M}\left(\overline{w}_{ij}^{1\bullet} \times \overline{w}_{il}^{1\bullet}\right) + \sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{\substack{k=1,\\ k\neq i}}^{N}\left(\overline{w}_{ij}^{\bullet 1} \times \overline{w}_{kj}^{\bullet 1}\right)\right) =$$

$$= \frac{1}{(NM)^2}\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\left(w_{ij}^{11}\right)^2 + \left(w_{ij}^{12}\right)^2\right\} + \left\{M^2\sum_{i=1}^{N}\left(\overline{w}_{i\bullet}^{1\bullet}\right)^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\overline{w}_{ij}^{1\bullet}\right)^2\right\} + \left\{N^2\sum_{j=1}^{M}\left(\overline{w}_{\bullet j}^{\bullet 1}\right)^2 - \sum_{j=1}^{M}\sum_{i=1}^{N}\left(\overline{w}_{ij}^{\bullet 1}\right)^2\right\}\right)$$

$$= \frac{1}{(NM)^2}\left(M^2\sum_{i=1}^{N}\left(\overline{w}_{i\bullet}^{1\bullet}\right)^2 + N^2\sum_{j=1}^{M}\left(\overline{w}_{\bullet j}^{\bullet 1}\right)^2\right) = \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet}^{1\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j}^{\bullet 1}\right)^2}{M^2}.$$

APPENDIX B

EXACT BOOTSTRAP-VARIANCE

The nonparametric estimator of AUC difference:

$$\hat{\Delta} = \frac{1}{NM}\sum_{i'=1}^{N}\sum_{j'=1}^{M} W_{i'j'}$$

Distribution of relative orderings in the bootstrap-space $\boldsymbol{B}$:

$$W_{i'j'} \sim Uniform\left[\{w_{ij}\}_{i=1,j=1}^{N,M}\right]$$

$$j' \neq l' \quad W_{i'j'} \times W_{i'l'} \sim Uniform\left[\{w_{ij} \times \{w_{il}\}_{l=1}^{M}\}_{i=1,j=1}^{N,M}\right]$$

$$i' \neq k' \quad W_{i'j'} \times W_{k'j'} \sim Uniform\left[\{w_{ij} \times \{w_{kj}\}_{k=1}^{N}\}_{i=1,j=1}^{N,M}\right]$$

Statistical moments of the relative orderings in $\boldsymbol{B}$:

$$E_B(W_{i'j'}) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij} = \overline{w}_{\bullet\bullet} \qquad E_B({W_{i'j'}}^2) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} {w_{ij}}^2$$

$$Var_B(W_{i'j'}) = E_B({W_{i'j'}}^2) - E_B(W_{i'j'})^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{\bullet\bullet})^2$$

$$E_B(W_{i'j'} \times W_{i'l'}) = \frac{1}{NM^2}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{M} w_{ij}w_{il} = \frac{1}{N}\sum_{i=1}^{N}{\overline{w}_{i\bullet}}^2 \qquad j' \neq l'$$

$$Cov_B(W_{i'j'}, W_{i'l'}) = E_B(W_{i'j'} \times W_{i'l'}) - E_B(W_{i'j'}) \times E_B(W_{i'l'}) = \frac{1}{N}\sum_{i=1}^{N}{\overline{w}_{i\bullet}}^2 - {\overline{w}_{\bullet\bullet}}^2 = \frac{\sum_{i=1}^{N}(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet})^2}{N}$$

$$E_B(W_{i'j'} \times W_{k'j'}) = \frac{1}{N^2M}\sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{k=1}^{N} w_{ij}w_{kj} = \frac{1}{M}\sum_{j=1}^{M}{\overline{w}_{\bullet j}}^2 \qquad i' \neq k'$$

$$Cov_B(W_{i'j'},W_{k'j'}) = E_B(W_{i'j'} \times W_{k'j'}) - E_B(W_{i'j'}) \times E_B(W_{k'j'}) = \frac{\sum_{j=1}^{M} \overline{w}_{\bullet j}^{\,2}}{M} - \overline{w}_{\bullet\bullet}^{\,2} = \frac{\sum_{j=1}^{M} (\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet})^2}{M}$$

Using the derived moments the bootstrap-variance of the nonparametric estimator of AUC difference can be computed in closed form:

$$V_B(\hat{\Delta}) = \frac{1}{N^2M^2}\left\{\sum_{i'=1}^{N}\sum_{j'=1}^{M} Var_B(W_{i'j'}) + \sum_{i'=1}^{N}\sum_{j'=1}^{M}\sum_{l'\neq j'} Cov_B(W_{i'j'},W_{i'l'}) + \sum_{j'=1}^{M}\sum_{i'=1}^{N}\sum_{k'\neq i'} Cov_B(W_{i'j'},W_{k'j'})\right\} =$$

$$= \frac{1}{N^2M^2}\left\{ NM\,\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{\bullet\bullet})^2}{NM} + \frac{NM(M-1)}{N}\sum_{i=1}^{N}(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet})^2 + \frac{MN(N-1)}{M}\sum_{j=1}^{M}(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet})^2 \right\} =$$

$$= \frac{1}{N^2M^2}\sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{\bullet\bullet})^2 + \frac{(M-1)}{N^2M}\sum_{i=1}^{N}(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet})^2 + \frac{(N-1)}{NM^2}\sum_{j=1}^{M}(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet})^2$$

Note that

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{\bullet\bullet})^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet})^2 + M\sum_{i=1}^{N}(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet})^2 + N\sum_{j=1}^{M}(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet})^2$$

Finally

$$V_B(\hat{\Delta}) = \frac{\sum_{i=1}^{N}(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet})^2}{N^2} + \frac{\sum_{j=1}^{M}(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet})^2}{M^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet})^2}{N^2M^2}$$

APPENDIX C

VARIANCE ESTIMATORS OF THE AUC DIFFERENCE

Bootstrap-variance:

$$V_B = \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M^2} + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{N^2 M^2}$$

Biased variance-estimator proposed by Wieand *et al.*

$$V_{Wb} = \frac{M-1}{M}\frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N^2} + \frac{N-1}{N}\frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M^2} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{N^2 M^2}$$

Unbiased variance-estimator proposed by Wieand *et al.*

$$V_W = \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N(N-1)} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M(M-1)} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{NM(N-1)(M-1)}$$

Two-sample jackknife (DeLong *et al.*)

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

One-sample jackknife

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

Certain deterministic relationships exist between considered variance-estimators:

$$V_{Wb} \le V_W \le V_{J2} \le V_{J1} \qquad V_{Wb} \le V_B \tag{C.1}$$

These inequalities can be proved by the following observations:

$$V_{Wb} \times \frac{NM}{(N-1)(M-1)} = V_W, \qquad \text{hence} \qquad V_{Wb} \le V_W$$

$$V_W + \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{NM(N-1)(M-1)} = V_{J2} \qquad \text{hence} \qquad V_W \le V_{J2}$$

To establish relationship $V_{J2} \le V_{J1}$ Note that:

$$V_{J1} = \frac{N+M-1}{N+M} \times \left[\frac{N}{N-1} \frac{\sum_{k=1}^{N}(\overline{w}_{k\bullet} - \overline{w}_{\bullet\bullet})^2}{N(N-1)} + \frac{M}{M-1} \frac{\sum_{l=1}^{M}(\overline{w}_{\bullet l} - \overline{w}_{\bullet\bullet})^2}{M(M-1)} \right] =$$

$$= \frac{(N+M-1)N}{(N+M)(N-1)} \left[\frac{\sum_{k=1}^{N}(\overline{w}_{k\bullet} - \overline{w}_{\bullet\bullet})^2}{N(N-1)} \right] + \frac{(N+M-1)M}{(N+M)(M-1)} \left[\frac{\sum_{l=1}^{M}(\overline{w}_{\bullet l} - \overline{w}_{\bullet\bullet})^2}{M(M-1)} \right]$$

Since $\frac{(N+M-1)N}{(N+M)(N-1)} = \frac{N^2+NM-N}{N^2+NM-N-M} > 1$ and $\frac{(N+M-1)M}{(N+M)(M-1)} = \frac{M^2+NM-M}{M^2+NM-M-N} > 1$

then the following relationship is straightforward: $V_{J2} \le V_{J1}$

Finally the inequality between the bootstrap-variance ($\boldsymbol{V_B}$) and biased estimator developed by Wieand *et al.* ($\boldsymbol{V_{WB}}$) can be established in a following way:

$$V_{Wb} = \frac{M-1}{M} \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N^2} + \frac{N-1}{N} \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M^2} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{N^2M^2} \le$$

$$\le \frac{\sum_{i=1}^{N}\left(\overline{w}_{i\bullet} - \overline{w}_{\bullet\bullet}\right)^2}{N^2} + \frac{\sum_{j=1}^{M}\left(\overline{w}_{\bullet j} - \overline{w}_{\bullet\bullet}\right)^2}{M^2} - \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(w_{ij} - \overline{w}_{i\bullet} - \overline{w}_{\bullet j} + \overline{w}_{\bullet\bullet}\right)^2}{N^2M^2} \le V_B$$

The estimator $\boldsymbol{V_W}$ proposed by Wieand *et al.* [18] is unbiased, allowing determination of the direction of the biases for other estimators. Namely, the relationship (C.1) indicates that two- and one- sample jackknife variance estimators ($\boldsymbol{V_{J2}}$ and $\boldsymbol{V_{J1}}$) are biased upwards while biased estimator ($\boldsymbol{V_{Wb}}$) is biased downwards.

ONE-SAMPLE JACKKNIFE (variance derivation)

The nonparametric estimator of AUC difference (later referred to as estimator):

$$\hat{\Delta} = \overline{w}_{\bullet\bullet} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}$$

The estimator computed on "reduced" sample:

$$\hat{\Delta}^{k,\bullet}_{(N-1)M} = \frac{1}{(N-1)M}\sum_{i\neq k}\sum_{j=1}^{M} w_{ij} = \frac{1}{(N-1)M}\left\{w_{\bullet\bullet} - w_{k\bullet}\right\} = \frac{N}{N-1}\overline{w}_{\bullet\bullet} - \frac{1}{N-1}\overline{w}_{k\bullet}$$

$$\hat{\Delta}^{\bullet,l}_{N(M-1)} = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j\neq l} w_{ij} = \frac{1}{N(M-1)}\left\{w_{\bullet\bullet} - w_{\bullet l}\right\} = \frac{M}{M-1}\overline{w}_{\bullet\bullet} - \frac{1}{M-1}\overline{w}_{\bullet l}$$

The pseudovalues:

$$\hat{\Delta}_{k,\bullet} = (N+M)\hat{\Delta} - (N+M-1)\hat{\Delta}^{k,\bullet}_{(N-1)M} = -\frac{M}{N-1}\overline{w}_{\bullet\bullet} + \frac{N+M-1}{N-1}\overline{w}_{k\bullet}$$

$$\hat{\Delta}_{\bullet,l} = (N+M)\hat{\Delta} - (N+M-1)\hat{\Delta}^{\bullet,l}_{N(M-1)} = -\frac{N}{M-1}\overline{w}_{\bullet\bullet} + \frac{N+M-1}{M-1}\overline{w}_{\bullet l}$$

Note that:

$$\sum_{k=1}^{N}\hat{\Delta}_{k,\bullet} = -\frac{NM}{N-1}\overline{w}_{\bullet\bullet} + \frac{(N+M-1)N}{N-1}\overline{w}_{\bullet\bullet} = \frac{N(N-1)}{N-1}\overline{w}_{\bullet\bullet} = N\overline{w}_{\bullet\bullet}$$

$$\sum_{l=1}^{M}\hat{\Delta}_{\bullet,l} = -\frac{NM}{M-1}\overline{w}_{\bullet\bullet} + \frac{(N+M-1)M}{M-1}\overline{w}_{\bullet\bullet} = \frac{M(M-1)}{M-1}\overline{w}_{\bullet\bullet} = M\overline{w}_{\bullet\bullet}$$

The one-sample jackknife-estimator (average of the pseudovalues):

$$\hat{\hat{\Delta}} = \frac{1}{N+M}\left\{\sum_{k=1}^{N}\hat{\Delta}_{k,\bullet} + \sum_{l=1}^{M}\hat{\Delta}_{\bullet,l}\right\} = \frac{1}{N+M}\left\{N\overline{w}_{\bullet\bullet} + M\overline{w}_{\bullet\bullet}\right\} = \overline{w}_{\bullet\bullet}$$

The one-sample jackknife-variance (a sample variance of the pseudo-values):

$$V_{J1} = \frac{1}{(N+M)(N+M-1)}\left[\sum_{k=1}^{N}(\hat{\Delta}_{k,\bullet} - \hat{\hat{\Delta}})^2 + \sum_{l=1}^{M}(\hat{\Delta}_{\bullet,l} - \hat{\hat{\Delta}})^2\right] = \frac{1}{(N+M)(N+M-1)}\times$$

$$\times\left[\sum_{k=1}^{N}\left\{-\frac{N+M-1}{N-1}\overline{w}_{\bullet\bullet} + \frac{N+M-1}{N-1}\overline{w}_{k\bullet}\right\}^2 + \sum_{l=1}^{M}\left\{-\frac{N+M-1}{M-1}\overline{w}_{\bullet\bullet} + \frac{N+M-1}{M-1}\overline{w}_{\bullet l}\right\}^2\right] =$$

$$= \frac{N+M-1}{N+M}\times\left[\frac{\sum_{k=1}^{N}(\overline{w}_{k\bullet} - \overline{w}_{\bullet\bullet})^2}{(N-1)^2} + \frac{\sum_{l=1}^{M}(\overline{w}_{\bullet l} - \overline{w}_{\bullet\bullet})^2}{(M-1)^2}\right]$$

APPENDIX D

CONDITIONAL TEST: VARIANCE ESTIMATOR

In the permutation space **Ω**, the moments of $\{W_{ij}\}_{w_{ij}^{11}\neq 0}$ constrained on the set of discordant pairs ***D***, can be expressed in a following manner:

$$\forall i,k=1,...,N;\ j,l=1,...,M;\ k\neq i; l\neq j: w_{ij}^{11}\neq 0, w_{il}^{11}\neq 0, w_{kj}^{11}\neq 0, w_{kl}^{11}\neq 0$$

$$E_{\Omega}(W_{ij}\mid D)=\frac{\sum_{p=1}^{2}\sum_{q=1}^{2}w_{ij}^{p,q}}{\sum_{p=1}^{2}\sum_{q=1}^{2}I(w_{ij}^{p,q}\neq 0)}=\frac{\left(w_{ij}^{11}+w_{ij}^{22}\right)+\left(w_{ij}^{12}+w_{ij}^{21}\right)}{\sum_{p=1}^{2}\sum_{q=1}^{2}I(w_{ij}^{p,q}\neq 0)}=0$$

$$Var_{\Omega}(W_{ij}\mid D)=E_{\Omega}(W_{ij}^{2}\mid D)-E_{\Omega}(W_{ij}\mid D)^{2}=\frac{\sum_{p=1}^{2}\sum_{q=1}^{2}\left(w_{ij}^{p,q}\right)^{2}}{\sum_{p=1}^{2}\sum_{q=1}^{2}I(w_{ij}^{p,q}\neq 0)}=2\frac{\left(w_{ij}^{11}\right)^{2}+\left(w_{ij}^{12}\right)^{2}}{\sum_{p=1}^{2}\sum_{q=1}^{2}I(w_{ij}^{p,q}\neq 0)}\ \text{and}$$

$$Cov_{\Omega}(W_{ij},W_{il}\mid D)=E_{\Omega}(W_{ij}W_{il}\mid D)-E_{\Omega}(W_{ij}\mid D)E_{\Omega}(W_{il}\mid D)=$$

$$=\frac{2\times\sum_{p=1}^{2}\sum_{q=1}^{2}\left(w_{ij}^{p,q}w_{il}^{p,q}+w_{ij}^{p,q}w_{il}^{p,3-q}\right)}{\left(\sum_{p=1}^{2}\sum_{q=1}^{2}I\left(w_{ij}^{p,q}\neq 0\right)\right)\times\left(\sum_{p=1}^{2}\sum_{q=1}^{2}I\left(w_{il}^{p,q}\neq 0\right)\right)}=\frac{4\times\left(w_{ij}^{11}+w_{ij}^{12}\right)\times\left(w_{il}^{11}+w_{il}^{12}\right)}{\left(\sum_{p=1}^{2}\sum_{q=1}^{2}I\left(w_{ij}^{p,q}\neq 0\right)\right)\times\left(\sum_{p=1}^{2}\sum_{q=1}^{2}I\left(w_{il}^{p,q}\neq 0\right)\right)}$$

$$Cov_{\Omega}(W_{ij},W_{kj}\mid D)=E_{\Omega}(W_{ij}W_{kj}\mid D)-E_{\Omega}(W_{ij}\mid D)E_{\Omega}(W_{kj}\mid D)=$$

$$= \frac{2 \times \sum_{p=1}^{2} \sum_{q=1}^{2} \left(w_{ij}^{p,q} w_{kj}^{p,q} + w_{ij}^{p,q} w_{kj}^{3-p,q} \right)}{\left(\sum_{p=1}^{2} \sum_{q=1}^{2} I\left(w_{ij}^{p,q} \neq 0 \right) \right) \times \left(\sum_{p=1}^{2} \sum_{q=1}^{2} I\left(w_{kj}^{p,q} \neq 0 \right) \right)} = \frac{4 \times \left(w_{ij}^{11} + w_{ij}^{21} \right) \times \left(w_{kj}^{11} + w_{kj}^{21} \right)}{\left(\sum_{p=1}^{2} \sum_{q=1}^{2} I\left(w_{ij}^{p,q} \neq 0 \right) \right) \times \left(\sum_{p=1}^{2} \sum_{q=1}^{2} I\left(w_{kj}^{p,q} \neq 0 \right) \right)}$$

$Cov_{\Omega}(W_{ij}, W_{kl} \mid D) = 0$.

The moments of the W follow:

$$E_{\Omega}(W \mid D) = \sum_{i,j: w_{ij}^{11} \neq 0} E\left(W_{ij} \mid D\right) = 0$$

$$Var_{\Omega}(W \mid D) = Var_{\Omega}\left(\sum_{\substack{i,j: \\ w_{ij}^{11} \neq 0}} W_{ij} \middle| D \right) =$$

$$= \sum_{\substack{i,j: \\ w_{ij}^{11} \neq 0}} Var_{\Omega}\left(W_{ij} \mid D\right) + \sum_{\substack{i,j,l \neq j: \\ w_{ij}^{11} \neq 0 \text{ and } w_{il}^{11} \neq 0}} Cov_{\Omega}\left(W_{ij}, W_{il} \mid D\right) + \sum_{\substack{i,j,k \neq i: \\ w_{ij}^{11} \neq 0 \text{ and } w_{kj}^{11} \neq 0}} Cov_{\Omega}\left(W_{ij}, W_{kj} \mid D\right).$$

In the above equations I(x) designates the indicator function.

BIBLIOGRAPHY

1. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. *Wiley & Sons Inc.*: New York, 2002.
2. Swets JA, Picket RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. *Academic Press*: New York, 1982.
3. Greenhouse SW, Mantel N. The evaluation of diagnostic tests. *Biometrics* 1950; **6**(4): 399-412.
4. Metz CE. Basic Principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; **8**(4): 283-298.
5. Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology* 1986; **21**(9): 720-733.
6. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994; **13**: 499-508.
7. Lee WC, Hsiao CK. Alternative summary indices for the receiver operating characteristic curve. *Epidemiology* 1996; **7**: 605-611.
8. Hilden J. The area under the ROC curve and its competitors. *Medical Decision Making* 1991; **11**: 95-101.
9. Hanley JA. The Robustness of the 'Binormal' Assumption Used in Fitting ROC Curves. *Medical Decision Making* 1988; **8**(3): 197-203.
10. Dorfman DD, Alf JrE. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology* 1969; **6**: 487-496.
11. Metz CE, Herman BA, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* 1998; **17**: 1033-1053.
12. Zou KH, Hall WJ, Shapiro DE. Smooth nonparametric receiver operating characteristics (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; **16**(9): 2143-2156.
13. Hanley JA, McNeil BJ. The meaning and use of the Area under Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982; **143**: 29-36.
14. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine* 2002; **21**: 3093-3106.

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

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

17. Hanley JA, McNeil BJ. A method of comparing the area under two ROC curves derived from the same cases. *Radiology* 1983; **148**: 839-843.

18. Wieand HS, Gail MM, Hanley JA. A nonparametric procedure for comparing diagnostic tests with paired or unpaired data. *I.M.S. Bulletin* 1983; **12**: 213-214.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Area under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988; **44**(3): 837-845.

20. Wieand HS, Gail M, James B, James K. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**: 585-592.

21. Metz CE, Wang P-L, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. *Information Processing in Medical Imaging VIII, F. Deconick (ed.)* 1984; 432-445. The Hangue: Martinus Nijhof.

22. Arvesen JN. Jackknifing U-statistics. *Annals of Mathematical Statistics* 1969; **40**(6): 2076-2100.

23 Beam CA, Wieand SH. A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. *Biometrics* 1991; **47**(3): 907-919.

24. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 1996; **83**(4): 835-848.

25. Venkatraman ES. A permutation test to compare receiver operating characteristic curves. *Biometrics* 2000; **56**: 1134-1136.

26. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* 1948; **19**(3): 293-325.

27. Rockette HE, Campbell WL, Britton CA, Holbert JM, King JL, Gur D. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Academic Radiology* 1999; **6**: 723-729.

28. Zhang DD, Zhou XH, Freeman DH, Freeman JL. A nonparametric method for the comparison of partial areas under ROC curves and its application to large health case data sets. *Statistics in Medicine* 2002; **21**: 701-715.

29. Zhou XH, Gatsonis CA. A simple method for comparing correlated ROC curves using incomplete data. *Statistics in Medicine* 1996; **15**: 1687-1693.

30. Metz C., Herman BA, Roe CA. Statistical comparison of two ROC estimates obtained from partially paired datasets. *Medical Decision Making* 1998; **18**: 110-121.

31. Bandos AI, Rockette HE, Gur D. Small sample size properties of the nonparametric comparison of the area under two ROC curves. *Medical Image Perception Society Conference X,* September 2003, Durham, NC.

32. Bandos AI, Rockette HE, Gur D. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine* 2005; scheduled for **24**(19).

33. Bandos AI, Rockette HE, Gur D. A conditional nonparametric test for comparing areas under two ROC curves from a paired design. *Academic Radiology* 2005; **12**: 291-297.

34 Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an ANOVA approach with dependent observations. *Communications in Statistics: Simulation and Computation* 1995; **24**(2): 285-308.

35. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* 1992; **27**: 723-731.

36. Roe CA, Metz CE. Variance-components modeling in the analysis of receiver operating characteristic index estimates. *Academic Radiology* 1997; **4**: 587-600.

37. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects receiver operating characteristic analysis. *Academic Radiology* 2000; **7**: 341-349.

38. Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997: **53**(1): 370-382.

39. Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology* 1998; **5**: 561-571.

40. Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Academic Radiology* 2002; **9**: 1278-1285.

41. Efron B, Tibshirani RJ. An introduction to the bootstrap. *Chapman & Hall*: New York, NY, 1993.

42. Efron B. Bootstrap Methods: Another look at the jackknife. *The Annals of Statistics* 1979; **7**: 1-26.

43. Mossman D. Resampling techniques in the analysis of non-binormal ROC data. *Medical decision making* 1995; **15**: 358-366.

44. McNemar Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* 1947; **12**: 153-157.

45. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A fee-response approach to the measurement and characterization of radiographic-observer performance. *Journal of Applied Photography and Engineering* 1978; **4**: 166-171.

46. Chakraborty DP. Maximum-likelihood analysis of free-response receiver operating characteristic (FROC) data. *Medical Physics* 1989; **16**: 561-568

47. Chakraborty DP, Winter LHL. Free-response methodology: Alternative analysis and a new observer-performance experiment. *Radiology* 1990; **174**: 873-881.