2011

Nonparametric Simultaneous Confidence Intervals for Multiple Comparisons of Correlated Areas Under the ROC Curves

Li Yue

Nonparametric Simultaneous Confidence Intervals for Multiple Comparisons of Correlated Areas Under the ROC Curves

(Spine title: Nonparametric SCIs for Multiple Comparisons of Correlated Areas Under the ROC Curves)

(Thesis format: Monograph)

by

Li Yue

Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

The performance of a medical diagnostic test yielding quantitative or ordinal measurements is often assessed in terms of its AUC, area under the receiver operating characteristic curve. As new tests constantly being developed, an essential task is to compare multiple AUCs, commonly derived from the same group of subjects. For this purpose, previous research usually uses an omnibus chi-square test that is non-informative and lacks power. In this study, we propose new methods of constructing simultaneous confidence intervals based on theory of nonparametric U-statistics. To improve the small sample properties, we adapt the method of variance estimates recovery by obtaining confidence limits for each AUC based on logit and inverse sinh transformation. A large simulation study demonstrates the good performance of our new method.

Key Words: Diagnostic Test, U-statistic, Coverage Probability, Hypothesis Testing, All Pairwise Comparisons, Comparisons with a Control

ACKNOWLEDGMENTS

I would like to express my deep appreciation to all people who have contributed to my progress during my study in the University of Western Ontario.

I am deeply grateful to my supervisors Dr. Guangyong Zou and Dr. Yun-Hee Choi for their patient guidance, kind support and encouragement throughout my study. Their persistence and enthusiasm for research inspire me to work harder.

I would also like to thank all professors and faculties in the Department of Epidemiology and Biostatistics at the University of Western Ontario who taught us very interesting courses and provided a perfect environment for our study.

I am indebted to Shun-Fu Chen, my classmate and also my good friend. She gave me useful suggestions and discussions while I encountered problems working on my thesis. Her help and encouragement made my thesis writing easier and stress free.

At last, I want to give my special thank to my parents Jinhai Yue and Baoju Xu, and my beloved husband Zhu Lan. Without their support and love, this work would not have been a success.

Contents

List of Tables

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Receiver Operating Characteristic curve

Accurate diagnosis of disease is the first step for appropriate treatment. The performance of a diagnostic test depends on its ability to distinguish individuals with a disease condition from those absent of the condition. A variety of approaches can be used to quantify the accuracy of a binary diagnostic test, including sensitivity/specificity, likelihood ratios, etc. When test outcomes are ordinal or continuous, the receiver operating characteristic (ROC) curve has been the most popular tool, especially after Hanly and McNeil (1982) introduced the method to the medical field.

Assume two groups of subjects, one with disease and the other without disease. For a diagnostic test measurement T, suppose that a subject with $T > c_0$ be classified as test positive, otherwise as test negative. For a given cutpoint, sensitivity is defined as the probability that a subject with condition is correctly diagnosed as positive (i.e. true positive rate (TPR)), and specificity is the probability that a subject absent of disease condition is classified as negative (i.e. true negative rate (TNR)). A ROC curve can be constructed by varying values of the cutpoint and then plotting the sensitivitiesagainst one minus specificities, or false positive rates (FPR).

A test with perfect discrimination would have a ROC curve that passes through the point $(0, 1)$ on the unit grid, while a test without discrimination would have a 45° diagonal line from the lower left corner to the upper right corner. Usually, a ROC curve is between these two extreme plots; the closer the plot is to the point $(0, 1)$, the higher the diagnostic

accuracy of the test. Three cases of diagnostic accuracy are depicted in Figure 1.1.

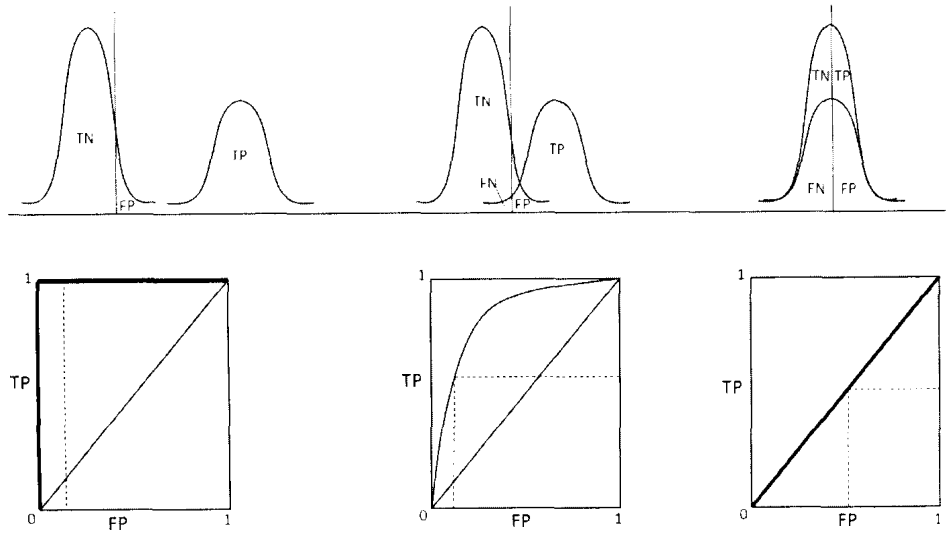

Figure 1.1: Three cases of diagnosis measurement distribution and ROCs. 'TN' = true negative rate, 'FN' = false negative rate, 'TP' = true positive rate and 'FP' = false positive rate.

1.2 The Area Under the ROC Curve

Test performance is usually summarized by the area under the ROC curve (AUC), which is closely related to Mann-Whiteney U statistic (Bamber, 1975). To see this, let X be the diagnostic test measurement on a randomly selected individual with disease and Y be the value of the same test result on a randomly selected individual without disease. Following

convention, we assume lower values of measurement are associated with normal subjects. As shown by Bamber (1975), the area under the ROC curve is given by:

$$\text{AUC} = \Pr(Y < X) + \frac{1}{2}\Pr(X = Y). \tag{1.1}$$

For continuous test results, $\Pr(X = Y) = 0$. Thus, AUC can be explained as a probability that a random individual from normal group will have a lower test value than that of a random individual from disease group.

Under the normality assumption for measurements, AUC can be defined as (Li *et al.*, 2010):

$$\begin{aligned}
\text{AUC} &= \Pr(Y < X) \\
&= \Pr\left(\frac{(Y - X) - (\mu_Y - \mu_X)}{\sqrt{\sigma_X^2 + \sigma_Y^2}} < \frac{-(\mu_Y - \mu_X)}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \right) \\
&= \Phi\Big(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \Big),
\end{aligned}$$

where μ_X and μ_Y are population means, σ_X^2 and σ_Y^2 are variances, and Φ is the standard normal cumulative distribution function. In this case, the estimated AUC can be easily obtained by substituting the sample means and variances for their unknown population parameters.

By convention, the area under the ROC curve is always ≥ 0.5 (if it is less than 0.5, we can reverse the decision role for 'tested positivity' from 'greater than' to 'less than' or visa versa), values of AUC range between 0.5, for no apparent difference between the test outcomes among two groups, resulting in a diagonal line in the ROC curve, and 1 for perfect separation of the test values of the two groups, resulting in a ROC curve that rises steeply along the left axis to the point (FPR=0, TPR=1).

1.3 Confidence intervals for the AUC

In present context, statistical inference for AUC usually focuses on confidence interval estimation which in turn reduces to the problem of variance estimation. Under the normal-

ity assumption for measurement of two groups, parametric maximum likelihood methods and delta method can provide direct estimates of the variance of AUC (Dorfman and Alf, 1969; Wieand *et al.*, 1989). Bamber (1975) derived the variance estimate of AUC by using theory of U-statistic. Hanley and McNeil (1982) provided an alternative approach for estimating the variance of AUC based on the traditional form of the nonnull variance for the Wilcoxon statistic. A more general approach was provided by DeLong *et al.* (1988). In addition, semi-parametric methods for variance estimate of AUC provide an intermediate strategy between the parametric and non-parametric approaches (Metz *et al.*, 1998; Wan and Zhang, 2007).

1.4 Multiple comparisons of AUCs

It is often of interest to compare two (or more) tests in terms of AUCs. An efficient design for this purpose is to apply each test to the same group of subjects. One strategy for comparing multiple diagnostic tests is by the use of simultaneous confidence intervals for multiple AUCs. Several approaches have been proposed for the construction of simultaneous confidence intervals for contrasts of AUCs in the literature (Hanley and McNeil, 1983; DeLong *et al.*, 1988; Hsu *et al.*, 2004). Among these approaches, the nonparametric method proposed by DeLong *et al.* (1988) is the most popular one. The method has now been implemented in SAS 9.2 PROC LOGISTIC. However, the DeLong's method relies on the large sample theory for all linear contrasts of AUCs and generates symmetric simultaneous confidence intervals that may not reflect sampling distribution of AUC estimates in finite sample sizes. Its performance is questionable for small to medium sample size (Hsu *et al.*, 2004).

1.5 Objective of the thesis

The objective of this thesis is to construct and evaluate the simultaneous confidence intervals for multiple comparisons of AUCs using the method of variance estimates recovery (MOVER) proposed by Zou and Donner (2008) and Donner and Zou (2010). The strength of this method is that it can reflect the asymmetry of simultaneous confidence intervals , and thus may improve the performance of simultaneous confidence intervals for multiple AUCs.

1.6 Organization of the thesis

Chapter 2 presents the literature review regarding the development of ROC curve and multiple comparisons of AUCs. Chapter 3 details the deviation of DeLong's method (DeLong *et al.*, 1988) and our method based on MOVER for constructing simultaneous CIs for multiple AUCs using these two approaches. Chapter 4 reports on a simulation study comparing the performances of DeLong's and our approach. The thesis concludes in Chapter 5 with general discussions and possible future works. The related SAS Macro is presented in the appendix.

Chapter 2

LITERATURE REVIEW

2.1 A brief history of the ROC curve

The ROC curve was first developed in 1950s in engineering for detecting radar signals (Peterson *et al.*, 1954). In the 1960s, ROC curve has found applications in psychology and psychophysics to assess weak signals in humans and occasionally non-human animals (Green and Swets, 1966). With the development of science and technology, ROC curve has now been applied to a variety of fields including engineering, quality control (e.g., materials testing), and weather forecasting. Swets and Pickett (1982) marked the beginning of the widespread use of this technique outside of psychophysics, where ROC analysis is often called the ROC Accuracy Ratio as a common technique for judging the accuracy of default probability models. ROC curve analysis has also been proven useful for the evaluation of machine learning techniques, as the first application by Spackman (1989) in comparing and evaluating different classification algorithms.

Application of ROC analysis in medicine to assess diagnostic test performance was first described by Lusted (1971) in the context of evaluating the performance of criteria for radiologists' assistants and radiologic systems. Erdreich and Lee (1981) also applied the method of ROC analysis to epidemiologic problems. Hanley and McNeil (1982, 1983, 1984) contributed greatly to the methodology improvement on the evaluation and comparisons of ROC curves with a series of related articles. DeLong *et al.* (1988) presented variance formulae that have become popular.

Methodological reviews on ROC curves can now be found in a variety of sources, e.g.

Zweig and Campbell (1993) and Pepe (2003).

2.2 Summary measures for ROC curve

A ROC curve depicts a graphical summary of discriminatory accuracy, but often a one-number summary index of discriminatory accuracy for ROC curve is desired. Several indices are present in literature and have been used in various applications (Shapiro, 1999; Greiner *et al.*, 2000). Here we provide a brief summary of the most widely used indices.

2.2.1 Youden index

Youden index is frequently used in practice to summarize diagnostic accuracy in terms of both sensitivity and specificity (Aoki *et al.*, 1997; Grmec and Gasparovic, 2001). It is defined as $J = \max_c\{Se(c) + Sp(c) - 1\}$ and ranges between 0 and 1. Complete separation of the distributions of the diagnostic test outcomes for the non-diseased and diseased groups results in $J = 1$ whereas complete overlap gives $J = 0$. Youden Index provides a criterion for choosing the optimal threshold value which maximizes the difference between the TPR and FPR (Greiner *et al.*, 2000). Graphically, Youden index is the maximum vertical distance between the ROC curve and the diagonal line.

2.2.2 Likelihood ratios

Similar to Youden index, likelihood ratios (LR) also summarize information about a diagnostic test by combining sensitivity and specificity.

Positive likelihood ratio (LR_+) has the expression that

$$LR_+ = \frac{\text{True Positive Rate}}{\text{False Postive Rate}} = \frac{\text{Sensitivity}}{1 - \text{Specificity}},$$

and negative likelihood ratio (LR_-) is

$$LR_- = \frac{\text{False Negative Rate}}{\text{True Negative Rate}} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}.$$

As compared to these two types of likelihood ratios, positive likelihood ratio is much more useful and is often called Likelihood ratio for short (Deeks and Altman, 2004). In clinical practice, LR_+ indicates how likely a positive result will be found in a person with the disease compared to a person without the disease. Graphically, it represents the slope of a ROC curve at a fixed cutoff point.

2.2.3 The area under the ROC curve

The area under the ROC curve (AUC) is the most commonly used summary measure of diagnostic accuracy. It represents the average sensitivity over all values of FPR and can be expressed as $\text{AUC} = \Pr(Y < X) + \frac{1}{2}\Pr(X = Y)$ (Bamber, 1975; Hanley and McNeil, 1982), where X and Y denote test responses from the diseased and non-diseased populations respectively. For continuous test results where $\text{P}(X = Y) = 0$, Bamber (1975) pointed out that AUC means the probability that a random subject from normal group will have a lower test value than that of a random subject in diseased group. He also made the connection to the Mann-Whitney U-statistic for comparing two independent groups.

2.2.4 The partial area under the ROC curve

The partial area under a ROC curve (PAUC) represents the area under the ROC curve over a range of FPR that is relevant to a particular setting. This index is useful when only relatively small false-positive rates (FPRs) are of interest. For example, The ROC_{50} index—the area under the lower portion of the ROC curve up to 0.5 FPR—has been applied to genomic research and clinical context (Gribskov and Robinson, 1996; Saigo *et al.*, 2004). Shapiro (1999) pointed out that if only a particular range of specificity or sensitivity values is relevant in study, PAUC may provide more detailed information and more appropriate accuracy than AUC. Few statistical analysis for PAUC are introduced in the literature (McClish, 1989; He and Escobar, 2008).

Other indices such as the projected length of the ROC curve (PLC) and the area swept

out by the ROC curve (ASC) have been introduced as alternatives to the AUC for continuous diagnostic tests (Lee and Hsiao, 1996). In this thesis, we will focus on the statistical inference for AUC because of its most popularity and attractive properties.

2.3 Statistical inference for a single AUC

Diagnostic outcomes can be obtained in dichotomous, e.g. positive or negative, ordinal, as in the case of confidence rating for presence of disease - definitely, probably, possibly, probably not, definitely not, and continuous, e.g. density serum cholesterol. Depending on assumptions made about the different outcome distributions, at least three approaches can be taken to perform statistical inference for AUC.

2.3.1 Parametric approach for normal data

One can assume a parametric distribution for the test outcomes of the diseased and non-diseased individuals such as normal, lognormal, negative exponential, beta distribution (Hanley, 1996; Goddard and Hinberg, 1990). Among these distributions, bivariate normal model is the most widely used one which assumes that the continuous test results from diseased and non-diseased individuals are independently normal distributed. Let μ_X and σ_X^2 represent the mean and variance of the diagnostic result X for the diseased population while μ_Y and σ_Y^2 denote these for the non-diseased population Y. Then

$$\mathrm{AUC} = \Pr(Y < X) = \Phi\Big(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\Big),$$

and the estimator $\widehat{\mathrm{AUC}}$ can be obtained by substituting the sample mean and variance for the unknown parameters. The variance estimate of $\widehat{\mathrm{AUC}}$ can be derived using delta method (Wieand *et al.*, 1989).

For ordinal rating data, Tsimikas *et al.* (2002) suggested a profile-likelihood inference. This method is particularly useful for high values of AUC. Simulation results showed that

the coverage rates of CIs derived from this method are close to the nominal level, even for extremely small or unbalanced sample sizes. However, this approach is not applicable to continuous measurements because the likelihood functions are constructed based on multinomial distribution.

A major advantage of parametric method is that it yields a smooth ROC curve for continuous test outcomes, and it can simplify the statistical inference for the summary indices such as AUC (Shapiro, 1999). However, substantial lack-of-fit may occur if the distributional assumptions are violated. Goddard and Hinberg (1990) pointed out that if the distribution of raw data from a quantitative test is far from normal distribution, the estimated AUC and corresponding standard error derived from a directly fitted binormal model can be seriously distorted. One way to avoid the possible distortion is to use semi-parametric approach.

2.3.2 Semi-parametric approach

Semi-parametric approach is an intermediate strategy between parametric and non-parametric methods. As Zou *et al.* (1997) pointed out, the transformation in semi-parametric method is non-parametric, but after transformation the model is parametric. Metz *et al.* (1998) introduced a semi-parametric method, with the assumptions that the underlying distributions of the test results for non-diseased and diseased groups can be transformed to approximately normal distributions by an unspecified monotone transformation, maximum likelihood algorithm can be used for the transformed data to estimate the unknown parameters (Dorfman and Alf, 1969). Bi-gamma (Dorfman *et al.*, 1997) and bi-beta (Zou *et al.*, 2004) models are also introduced when bi-normality is not satisfied after transformation.

Recently, several new methods have been proposed to produce semi-parametric ROC curves. Pepe and Cai (2004) regard the ROC curve as the distribution of placement values and then estimate AUC using pseudo-likelihood function. Erkanli *et al.* (2006) proposed a semi-parametric Bayesian approach for AUC estimate and concluded that their Bayesian estimation was very close to kernel density estimation (Green and Swets, 1988).

Semi-parametric approach can create a smooth ROC curve with an available program named LABROC4 (Metz *et al.*, 1998). However, its complicated way to group the data and the possible lack-of-fit are the main disadvantages.

2.3.3 Non-parametric approach

Non-parametric approach does not make distributional assumptions for diagnostic test results. It depends only on the ranks of the observations in the combined sample. The resulted empirical ROC curve is a series of horizontal and vertical steps, which can be jagged (Zweig and Campbell, 1993).

The commonly used non-parametric estimate for AUC is Mann-Whitney-Wilcoxon U-statistic with the expression given by:

$$\widehat{\mathrm{AUC}} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \Psi(X_i, Y_j),$$

where

$$\Psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases},$$

X_i, $i = 1, \cdots, m$ and Y_j, $j = 1, \cdots, n$, are the test outcomes measured from diseased and non-diseased groups, respectively (Bamber, 1975). The advantages of this statistic are that not only its computation is simple and straightforward, but also it provides an unbiased estimate for the AUC generated from discrete test outcomes (Zweig and Campbell, 1993; Hsu *et al.*, 2004). However, if the outcomes are continuous data, since the empirical cumulative curve is jagged and under the theoretical smooth ROC curve, the non-parametric area estimates may tend to underestimate the AUC in particular when the number of distinct values is small (DeLong *et al.*, 1988; Hajian-Tilaki *et al.*, 1997).

At least four approaches can be adopted in estimating the variance of $\widehat{\mathrm{AUC}}$. The first approach was proposed by Bamber (1975) using the trapezoidal rule and the Mann-Whitney

U-statistic theory. For any two Y values Y_j, Y_k, and any X value X_i, let

$$b_{yyx} = \Pr(Y_j, Y_k < X_i) + \Pr(X_i < Y_j, Y_k) - 2\Pr(Y_j < X_i < Y_k),$$

and for any two X values X_i, X_l, and any Y value Y_j,

$$b_{xxy} = \Pr(X_i, X_l < Y_j) + \Pr(Y_j < X_i, X_l) - 2\Pr(X_i < Y_j < X_l).$$

The variance estimate for $\widehat{\text{AUC}}$ is given by

$$\begin{aligned}\text{var}(\widehat{\text{AUC}}) &= \frac{1}{4}(m-1)(n-1))\{\Pr(X \neq Y) + (m-1)b_{xxy} \\ &\quad + (n-1)b_{yyx} - 4(m+n-1)(\widehat{\text{AUC}} - 0.5)^2\}.\end{aligned}$$

The second approach was suggested by Hanley and McNeil (1982) who derived the results by assuming negative exponential distribution for measurements.

Let $Q_1 = \Pr(Y_j < X_i, X_l)$, $Q_2 = \Pr(Y_j, Y_k < X_i)$, where X_i, X_l; Y_j, Y_k are randomly chosen subjects from diseased and non-diseased groups respectively. The variance for the AUC can then be estimated by

$$\text{var}(\widehat{\text{AUC}}) = \left[\widehat{\text{AUC}}(1 - \widehat{\text{AUC}}) + (m-1)(Q_1 - \widehat{\text{AUC}}^2) + (n-1)(Q_2 - \widehat{\text{AUC}}^2)\right] \Big/ mn.$$

However, because the underlying negative exponential distribution assumptions are made for this approach, it tends to underestimate the variance when the AUC is close to 0.5 and overestimate it when the AUC is near 1 (Hanley and Hajian-Tilaki, 1997).

Based on theory of U-statistic, DeLong *et al.* (1988) derived the variance estimates under two or more related ROC curves derived from same individuals. For what to follow, we refer to this method as DeLong's method. For each diseased subject $i, i = 1, \cdots, m$, we have

$$V_{10}(X_i) = \frac{1}{n}\sum_{j=1}^{n} \Psi(X_i, Y_j) \text{ and } S_{10} = \frac{1}{m-1}\sum_{i=1}^{m}(V_{10}(X_i) - \widehat{\text{AUC}})^2.$$

Similarly, for each non-diseased subject $j, j = 1, \cdots, n$, we define

$$V_{01}(Y_j) = \frac{1}{m}\sum_{i=1}^{m} \Psi(X_i, Y_j) \text{ and } S_{01} = \frac{1}{n-1}\sum_{j=1}^{n}(V_{01}(Y_j) - \widehat{\text{AUC}})^2.$$

Then, DeLong's variance of the estimated AUC is given by

$$\text{var}(\widehat{\text{AUC}}) = \frac{1}{m}S_{10} + \frac{1}{n}S_{01}. \tag{2.1}$$

Using the idea of jackknife, Hanley and McNeil (1984) proposed the fourth approach. Hanley and McNeil (1984) first introduced the AUC pseudo value (pAUC) corresponding to observation i as

$$\text{pAUC}_i = (m+n)\text{AUC} - (m+n-1)\text{AUC}_{(-i)},$$

where AUC is the area calculated with all $m+n$ observations and AUC_{-i} the area obtained from the $(m+n-1)$ observations, with observation i deleted. And the variance of the AUC is given by

$$\begin{aligned} \text{var}(\text{AUC}) &= \text{variance of mean of all } m+n \text{ pAUCs} \\ &= \frac{\text{variance of all pAUCs}}{m+n : \text{number of pAUCs}} \end{aligned} \tag{2.2}$$

Hanley and Hajian-Tilaki (1997) compared the above four methods and concluded that DeLong's method is the most accurate. Cleves (1999, 2002) numerically compared the first three approaches and came to a similar conclusion. However, DeLong's method may break down when sample sizes are small or the underlying AUC is high (Tsimikas *et al.*, 2002; Vergara *et al.*, 2008).

An alternative nonparametric approach for estimating AUC is kernel density estimation based on kernal smoothing techniques. Zou *et al.* (1997) pointed out that Kernel method can create a smooth ROC curve, and it also follows closely the details of the original data. However, this method may generate wide CIs that are too conservative for small sample size (Zou *et al.*, 1997).

Standard softwares can be used to perform the statistical inference on one or more AUCs based on DeLong's method, such as the program package in SAS, Stata and a free software called StAR in R (Vergara *et al.*, 2008). In this thesis, we attempt to improve DeLong's method by applying logit and inverse sinh transformation to AUC. Comparisons

between AUCs are then conducted using the MOVER approach (Zou and Donner, 2008), in conjunction with multiple comparisons theory (Nelson, 1989; Donner and Zou, 2010). A SAS macro implementing our method is also provided in the thesis.

2.4 Confidence interval construction for a single AUC

Confidence interval (CI) estimates are usually regarded as more informative than significance tests because they provide a range of parameter values that reflect the degree of uncertainty in the estimation procedure. Moreover, confidence interval estimation encompasses hypothesis testing (Altman, 2005). In fact, a CI may be regarded as performing significance tests for all values of a parameter, not just the single value corresponding to the null hypothesis (Cox and Hinkley, 1974, Section 7.2). Therefore, in this thesis, we focus mainly on the confidence interval construction for AUC rather than the hypothesis test.

Variance estimation in section 2.3 can be used for constructing CI for a single AUC. Most parametric and non-parametric methods result in a symmetrically Wald-type confidence interval using $\widehat{\text{AUC}} \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\text{AUC}})}$. Asymmetric CI may be provided by semi-parametric approach since it needs a monotone transformation from a Wald-type CI. As pointed out by Efron and Tibshirani (1993) that when the sample distribution is skewed, symmetric CI does not perform well. Moreover, in diagnostic research, the AUC is usually close to 1, a symmetric CI may produce upper limit that is greater than 1 and may lead to a confused interpretation.

In the case of proportions, Newcombe (2001) proposed two methods based on logit or inverse sinh transformation to improve the performance of the Wald method. Since the AUC by definition is a probability, we can apply the same transformations to improve the performance.

2.5 Multiple comparisons of multiple AUCs

It is often of interest to compare several diagnostic measures simultaneously. In order to compare k ($k \geq 2$) diagnostic tools, two or more biomarkers can be simultaneously measured on paired data, where all tests are applied to the same subjects; or unpaired data, where different tests are performed on the different groups. Of these two designs, Zweig and Campbell (1993) pointed that the paired study using the same individuals has more efficiency because it can better control the patient-to-patient variation.

A conventional approach for comparing several diagnostic tests is to compare the entire ROC curves by using a global measure such as AUC. Parametric procedures have been suggested to compare two or more AUCs from independent binormal ROC curves (Dorfman and Alf, 1969; Metz and Kronman, 1980). Metz *et al.* (1984) extended 'binormal' model to 'bivariate' model for comparing areas under two correlated ROCs, one approach used studentized range (SR) test and another method was based on jackknife theory. Simulation results (Metz et al., 1984) showed that the two methods are comparable when the number of subjects in both non-diseased and diseased groups are the same, but the jackknife methodology performs better than SR test for unequal subjects in two groups. Note that only the hypothesis test: $H_0 : \mathrm{AUC}_1 = \cdots = \mathrm{AUC}_k$ was provided in McClish (1987) using these two methods, but simultaneous confidence intervals were not available.

DeLong *et al.* (1988) proposed a non-parametric approach for comparing several correlated ROC curves based on Scheffe's method and provided asymptotically simultaneous confidence intervals for several correlated AUCs. This method is intended for any set of linear contrasts of areas under correlated ROC curves. Hsu et al. (2004) derived an asymptotic method for identifying the best among several groups. This method is commonly known as multiple comparison with the best, which depends on 'many-to-one' comparsion treating each test in turn as a control to compare diagnostic tools using AUCs. In this method, the point estimation of AUC and its variance estimate are based on DeLong's method, with critical value obtained using Dunnett's many-to-one comparisons (Hsu, 1996). Furthermore, approximately normal assumptions of the statistics for constructing simultaneous confi-

dence intervals are needed. Simulation results suggested that the simultaneous coverage probabilities are lower than the nominal level, due partly to the enforced symmetry.

Donner and Zou (2010) has proposed an approach to construct simultaneous confidence intervals for proportions. The key advantage of this method is that it can avoid the enforced symmetry of simultaneous confidence intervals . In this thesis, we adopt this approach in the construction of simultaneous confidence intervals for comparing AUCs.

Chapter 3

DEVELOPMENT OF THE METHOD

This chapter first reviews methods for constructing simultaneous confidence intervals using variance estimator proposed by DeLong *et al.* (1988). To improve the performance of these intervals, we apply the method of varance estimates recovery (Zou and Donner, 2008; Zou, 2008) to the context of simultaneous confidence intervals for AUC.

3.1 Multiple comparisons of AUCs based on U-statistic

3.1.1 Point estimation of a single AUC

Let Y denote test scores from a non-diseased subjects and X as scores from diseased population. Following the convention in the literature, assume that smaller scores are more likely related to non-diseased subjects. The area under the ROC curve (AUC) can be defined as:

$$\text{AUC} = \Pr(Y < X) + \frac{1}{2}\Pr(Y = X).$$

For continuous test scores, $\Pr(Y = X) = 0$.

Suppose $X_i, i = 1,2,\cdots,m$; $Y_j, j = 1,\cdots,n$, are scores from the diseased and non-diseased groups respectively. The AUC can be estimated with Mann-Whitney U-statistics as:

$$\widehat{\text{AUC}} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\Psi(X_i, Y_j),$$

where

$$\Psi(X_i,Y_j) = \begin{cases} 1 & Y_j < X_i \\ \frac{1}{2} & Y_j = X_i \\ 0 & Y_j > X_i \end{cases}$$

3.1.2 Estimation of variance covariance matrix for multiple AUCs

Assume k ($k \geq 2$) diagnostic tests are performed on the same subjects with $\{X_i^r\}, \{Y_j^r\}$ $i = 1,\cdots,m$; $j = 1,\cdots,n$; $1 \leq r \leq k$, representing the measured scores for diseased and non-diseased groups from the rth test. DeLong *et al.* (1988) adapted a method for non-parametric statistic (Sen, 1960) to estimate the variance-covariance matrix for the vector of parameter estimates $\widehat{\mathbf{AUC}} = (\widehat{\mathrm{AUC}}_1, \cdots, \widehat{\mathrm{AUC}}_k)$ as follows:

$$S = \frac{1}{m}S_{10} + \frac{1}{n}S_{01}, \tag{3.1}$$

where the (r,s)th element of $k \times k$ matrix S_{10} is

$$s_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^r(X_i) - \hat{\theta}^r][V_{10}^s(X_i) - \hat{\theta}^s],$$

in which $V_{10}^r(X_i) = \frac{1}{n}\sum_{j=1}^{n}\Psi(X_i^r, Y_j^r)$, $i = 1,2,\cdots,m$.

Similarly, S_{01} has the (r,s)th element:

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^r(Y_j) - \hat{\theta}^r][V_{01}^s(Y_j) - \hat{\theta}^s]$$

and $V_{01}^r(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\Psi(X_i^r, Y_j^r)$, $j = 1,2,\cdots,n$.

Hanley and Hajian-Tilaki (1997) provided an intuitive approach to obtain the point estimates of AUCs and the associated variance estimates. The key idea is to first transfer measurements X_i, Y_i into placement value according to $\Psi(X_i,Y_i)$, and then conduct analysis on the placement values. To illustrate Hanley and Hajian-Tilaki's method, here we use a subset of the data from DeLong *et al.* (1988) as an example.

3.1.3 Hanley and Hajian-Tilaki's simplification on the covariance matrix estimate for multiple AUCs

3.1.3.1 Data used for illustration

The full data set arose from the study that evaluates the discriminating abilities of 3 preoperative scoring measurements on the prognosis of surgical correction for patients with ovarian cancer who also get intestinal obstruction (Krebs and Goplerud, 1983). Patients who survived longer than 2 months after operation are considered surgical successes, otherwise, they are regarded as failed operation cases. In this study, 49 patients were observed at Duke University Medical Center after their surgery, among those patients, 12 survived more than 2 months postoperatively and were marked as 'successful cases'; the remaining 37 survived no more than 2 months and were labeled as 'failure cases'. Based on the 49 observations, 3 tests including Krebs-Goplerud (*K-G*) score and two measures of nutritional status: total protein (*TP*) and albumin (*ALB*) were evaluated and compared on their discriminant accuracy. The full data set has been used by SAS version 9 as an example in PROC LOGISTIC.

It has been shown that the increasing levels of both *ALB* and *TP* are related to better prognosis. In contrast, the higher the level of *K-G*, the poorer the prognosis. Thus, each value of *K-G* is subtracting from 12, the maximum possible value, so that all three indices can be discriminated in the same direction.

For illustration, we only use data on *ALB* and *TP* from the first 20 subjects (Table 3.1).

Table 3.1: Two measures of nutritional status–Albumin (*ALB*), Total Protein (*TP*) and post-operative result on 20 patients with ovarian cancer and intestinal obstruction

ALB	TP	Postoperative result
3.0	5.8	success
3.2	6.3	failure
3.9	6.8	failure
2.8	4.8	success
3.2	5.8	failure
0.9	4.0	success
2.5	5.7	success
1.6	5.6	failure
3.8	5.7	failure
3.7	6.7	failure
3.2	5.4	failure
3.8	6.6	failure
4.1	6.6	failure
3.6	5.7	failure
4.3	7.0	failure
3.6	6.7	success
2.3	4.4	failure
4.2	7.6	success
4.0	6.6	success
3.5	5.8	failure

3.1.3.2 Illustration of obtaining AUC point estimates and variance-covariance estimates using placement value

We start with estimates for a single diagnostic outcome. Table 3.1 shows the ALB scores of 7 success subjects $X_1, \cdots, X_7$, and those of 13 failure individuals $Y_1, \cdots, Y_{13}$. Thus, we form a 13×7 matrix with X values in the top row margin and Y values in the left column margin as displayed in Table 3.2. In each cell of the matrix, we assign value 1 if $Y_i < X_j$, 0.5 if $Y_i = X_j$, and 0 if $Y_i > X_j$, $i = 1, \cdots, 7,\ j = 1, \cdots, 13$. Then, the placement value V_X of a particular value X is defined as the average of the column entries corresponding to that X. For example, for X_1 in Table 3.2, its corresponding V_{X_1} is equal to the average of 13 entries in column 1, that is $V_{X_1} = 2/13 = 0.15$. Similarly, the replacement value V_Y for Y is the average of the row of the entries to that related Y. Calculation details are shown in the Table 3.2.

Once we get the replacement values V_X and V_Y, the variance of the AUC estimate is given by:

$$\widehat{\text{Var}}(\widehat{\text{AUC}}) = \text{Variance of mean for } V_X + \text{Variance of mean for } V_Y \tag{3.2}$$

The variance has two components because it is affected by the variability of both X and Y.

Table 3.2: Variance estimate for AUC of Albumin (*ALB*) from rating data for 7 success and 13 failure subjects*

Subjects	$X_1 = 3$	$X_2 = 2.8$	$X_3 = 0.9$	$X_4 = 2.5$	$X_5 = 3.6$	$X_6 = 4.2$	$X_7 = 4.0$	V_Y
Y_1 =3.2	0	0	0	0	1	1	1	3/7
Y_2 =3.9	0	0	0	0	0	1	1	2/7
Y_3 =3.2	0	0	0	0	1	1	1	3/7
Y_4 =1.6	1	1	0	1	1	1	1	6/7
Y_5 =3.8	0	0	0	0	0	1	1	2/7
Y_6 =3.7	0	0	0	0	0	1	1	2/7
Y_7 =3.2	0	0	0	0	1	1	1	3/7
Y_8 =3.8	0	0	0	0	0	1	1	2/7
Y_9 =4.1	0	0	0	0	0	1	0	1/7
Y_{10} =3.6	0	0	0	0	0.5	1	1	2.5/7
Y_{11} =4.3	0	0	0	0	0	0	0	0
Y_{12} =2.3	1	1	0	1	1	1	1	6/7
Y_{13} =3.5	0	0	0	0	1	1	1	3/7
V_X	2/13	2/13	0	2/13	6.5/13	12/13	11/13	0.39

*The entries in the table are the placement values of X with respect to Y, defined as 1 if $Y < X$, 0 if $Y > X$ and 0.5 if $Y = X$. The data in the margins of the Table are calculated with the averages of the related rows/columns and are defined as the placements corresponding to each X and each Y. Thus, AUC = average of $V_X's$= average of $V_Y's = 35.5 \Big/ (13 \times 7) = 0.39$. $\text{Var}(V_X) = \sum_{i=1}^{7}(V_{X_i} - \bar{V}_X)^2 \Big/ (7-1) = 0.137$, $\text{Var}(V_Y) = \sum_{j=1}^{13}(V_{Y_j} - \bar{V}_Y)^2 \Big/ (13-1) = 0.058$. $\widehat{\text{Var}}(\widehat{\text{AUC}}_{ALB}) = 0.137/7 + 0.058/13 = 0.024$.

The covariance for the related AUC_1 and AUC_2 is then obtained as:

$$\widehat{\mathrm{Cov}}(\widehat{\mathrm{AUC}}_1, \widehat{\mathrm{AUC}}_2) = \frac{\mathrm{Cov}(V_{X1}, V_{X2})}{n_X} + \frac{\mathrm{Cov}(V_{Y1}, V_{Y2})}{n_Y} \tag{3.3}$$

where V_{X_i} and V_{Y_i} are replacements values for AUC_i, $i = 1, 2$. For data in Table 3.1, the covariance of AUC_{ALB} and AUC_{TP}, using replacement V_X, V_Y for two nutritious scores ALB and TP is given in Table 3.3.

Table 3.3: Calculation of covariance of two AUCs using the covariance of Placement Values

Subj Group	Placements	
and No.	V_{ALB}	V_{TP}
Success Subjects		
1	2/13	6/13
4	2/13	1/13
6	0	0
7	2/13	4/13
16	6.5/13	10.5/13
18	12/13	13/13
19	11/13	9/13
Covariance	0.123	
Failure Subjects	V_{ALB}	V_{TP}
2	3/7	3/7
3	2/7	1/7
5	3/7	3.5/7
8	6/7	5/7
9	2/7	4.5/7
10	2/7	1.5/7
11	3/7	5/7
12	2/7	2.5/7
13	1/7	2.5/7
14	2.5/7	4.5/7
15	0	1/7
17	6/7	6/7
20	3/7	3.5/7
Covariance	0.042	

$\text{Cov}(\text{AUC}_{ALB}, \text{AUC}_{TP}) = 0.123/7 + 0.042/13 = 0.021.$

3.1.4 A Chi-square test for multiple contrasts of AUCs

Having obtained the variance-covariance matrix for the vector of parameter estimates $\widehat{\mathbf{AUC}} = (\widehat{\mathrm{AUC}}_1, \cdots, \widehat{\mathrm{AUC}}_k)'$, we can construct, for any linear comparison L' **AUC**,

$$\frac{L'\widehat{\mathbf{AUC}} - L'\mathbf{AUC}}{[L'SL]^{1/2}} \sim N(0,1),$$

where L is a column vector of coefficients, S is the variance-covariance matrix of $\widehat{\mathbf{AUC}}$. Then a Wald-type confidence interval for the linear comparison is easily obtained. DeLong *et al.* (1988) gave an example to compare the diagnostic accuracy of *K–G* to the average of *ALB* and *TP* based on the same data we described in former section. In this case, $L = (1, -0.5, -0.5)'$, and the two-sided confidence interval for this contrast is given by $(-0.223,\ 0.231)$. Since the CI covers 0, indicating that the test accuracy of *K–G* is not significantly improved as compared with the average of *ALB* and *TP*.

DeLong *et al.* (1988) also concluded that their results can be generalized to any set of linear comparisons L(**AUC**) which is similar to Scheffe's method (Nelson, 1989), here L is a comparison matrix. Similarly, we have

$$(\widehat{\mathbf{AUC}} - \mathbf{AUC})'L'[LSL']^{-1}L(\widehat{\mathbf{AUC}} - \mathbf{AUC})$$

is χ^2 distributed with degrees of freedom equal to the rank of LSL'. For example, DeLong *et al.* (1988) compared the *K-G* versus *ALB* and *K-G* versus *TP* simultaneously to test whether the *K-G* score is better than at least one of the other indices, *ALB* and *TP*, then the contrast matrix is defined as

$$L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}.$$

The relevant p-value is computed as 0.47 showing that the 2 degrees of freedom test that the *K-G* score is different from at least one other test is not significant at the 0.05 level.

This omnibus χ^2 test above has been implemented in PROC LOGISTIC in SAS V9.2. However, it does not indicate which tests are significantly different even if the overall test is significant. One alternative is to construct simultaneous confidence intervals for multiple

contrasts of AUCs. In addition, we improve the small sample properties by applying the method of variance estimates recovery (MOVER) as shown in Donner and Zou (2010) with the transformed CIs by logit and inverse sinh transformation for each AUC (Newcombe, 2001).

3.2 Our Approach

When setting simultaneous confidence intervals for multiple contrasts of AUCs, the first step should be constructing a valid CI for each AUC. DeLong *et al.* (1988) provided a Wald type CI based on U statistic. However, a Wald type symmetric CI only performs well when sampling distributions of parameter estimates are at least approximately normal distributed. Since AUC values between 0.5 to 1, and the point estimate of AUC is related to its variance, the sample distribution for $\widehat{\text{AUC}}$ is hardly symmetric especially when the underlying value is large. For large values of AUC, Wald-type CI can lead to an upper limit of greater than 1.

3.2.1 Improving CI for a single AUC with transformations

Newcombe (2001) introduced logit and inverse sinh transformations for binomial proportion (p) to improve the performance of confidence interval procedure when sample size is not large. It is shown that the former interval always contains the latter, which is the Wilson interval (Wilson, 1927). Specifically, he showed that the Wilson interval is symmetric on the logit scale, given by

$$\begin{aligned} & \text{logit}(\hat{p}) \pm 2\text{arcsinh}\left(\frac{z_{\alpha/2}}{2}\frac{1}{\sqrt{npq}}\right) \\ = \; & \text{logit}(\hat{p}) \pm 2\text{arcsinh}\left(\frac{z_{\alpha/2}}{2} s.e.(\text{logit}(\hat{p}))\right) \end{aligned}$$

Since by definition AUC is a probability, here we apply the same transformations to get the improved CI for a single AUC by replacing p for AUC. For logit transformation, we

first take a logit transformation for AUC:

$$\text{logit}(\text{AUC}) = \ln[\text{AUC}/(1-\text{AUC})],$$

and then apply the delta method to obtain the standard error, which is given by

$$\widehat{s.e.}(\text{logit}(\widehat{\text{AUC}})) = \frac{1}{\widehat{\text{AUC}}(1-\widehat{\text{AUC}})}\widehat{s.e.}(\widehat{\text{AUC}}).$$

Thus the 100(1-α)% CI for logit(AUC) is given by

$$(l,u) = \text{logit}(\widehat{\text{AUC}}) \pm z_{\alpha/2}\frac{\widehat{s.e.}(\widehat{\text{AUC}})}{\widehat{\text{AUC}}(1-\widehat{\text{AUC}})},$$

yielding the CI for AUC as

$$\left(\frac{e^l}{1+e^l}, \frac{e^u}{1+e^u}\right).$$

For method based on inverse sinh transformation, we take inverse sinh transformation of the margin of error for logit($\widehat{\text{AUC}}$), resulting in the modified CI for AUC as

$$\left(\frac{e^{l_n}}{1+e^{l_n}}, \frac{e^{u_n}}{1+e^{u_n}}\right),$$

where $(l_n, u_n) = \text{logit}(\widehat{\text{AUC}}) \pm 2\text{arcsinh}\left(\frac{z_{\alpha/2}}{2}\frac{\widehat{s.e.}(\widehat{\text{AUC}})}{\widehat{\text{AUC}}(1-\widehat{\text{AUC}})}\right)$.

Both transformations generate asymmetric CIs around AUC.

3.2.2 Confidence interval estimation for a linear combination of parameters

Once we get the CI for each AUC, the next step is to generate simultaneous confidence intervals for multiple contrasts of AUCs. The key point for simultaneous confidence intervals construction is to set the CI for a linear combination of parameters of interest when the CIs for each parameter components are available.

Zou and Donner (2008) proposed a method to construct confidence interval for a difference between two measures using confidence limits for each parameter. This method is

further generalized to estimate the CI for a linear combination of parameters (Zou, 2008; Zou *et al.*, 2009) and a ratio of two parameters (Li *et al*, 2010; Donner and Zou, 2010). The approach is termed as MOVER meaning a method of variance estimates recovery because the key idea is to recover variance estimates from confidence limits for single parameters. The MOVER only requires the efficient confidence limits for each parameter components and does not enforce symmetry around the final CI. Furthermore, the confident limit for function of parameters using MOVER can be easily calculated in a closed form, no interactive or re-sampling procedures are needed. Here we provide a summary of the MOVER approach.

We start with an approximate $100(1-\alpha)\%$ two-sided confidence interval construction (L,U) for $\theta_1+\theta_2$, where θ_1 and θ_2 are parameters of interest. Denote $\hat{\theta}_1$ and $\hat{\theta}_2$ are their point estimates separately and assumed to be independent. Using the central limit theorem, the lower limit L is given by

$$L = \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)}, \tag{3.4}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and $\text{var}(\hat{\theta}_i), i = 1,2$, are unknown, so they need to be estimated.

Now suppose that a $100(1-\alpha)\%$ confidence interval for θ_i is (l_i,u_i), $i=1,2$. Simple derivation can show that l_1+l_2 is closer to L than $\hat{\theta}_1+\hat{\theta}_2$, thus we can estimate $\text{var}(\hat{\theta}_i)$ at $\theta_i = l_i, i = 1,2$. Again, by the central limit theorem, we have

$$l_i = \hat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta}_i)},\ i = 1,2,$$

then the variance estimate for $\text{var}(\hat{\theta}_i)$ at $\theta_i = l_i$ is given by

$$\widehat{\text{var}}(\hat{\theta}_i) = (\hat{\theta}_i - l_i)^2/z^2_{\alpha/2}.$$

Substituting these variance estimates into equation (3.4), it gives the lower limit L for $\theta_1+\theta_2$ as

$$L = \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta}_1) + \widehat{\text{var}}(\hat{\theta}_2)} \tag{3.5}$$

$$= \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{(\hat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (\hat{\theta}_2 - l_2)^2/z^2_{\alpha/2}}$$
$$= \hat{\theta}_1 + \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (\hat{\theta}_2 - l_2)^2}.$$

Similarly, we estimate $\text{var}(\hat{\theta}_i)$ for upper limit U at $\theta_i = u_i, i = 1,2$, since $u_1 + u_2$ is close to U, then we obtain the upper limit U for $\theta_1 + \theta_2$ as

$$U = \hat{\theta}_1 + \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (u_2 - \hat{\theta}_2)^2}. \tag{3.6}$$

Noticing that the confidence interval for $-\theta_i$ is $(-u_i, -l_i)$, we rewrite $\theta_1 - \theta_2$ as $\theta_1 + (-\theta_2)$, then the confidence limits for $\theta_1 - \theta_2$ are:

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2} \tag{3.7}$$
$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2}. \tag{3.8}$$

The MOVER approach can also be extended to the cases where $\hat{\theta}_1$ and $\hat{\theta}_2$ are correlated with the correlation coefficient $\rho > 0$, in this condition, the confidence limits for $\theta_1 - \theta_2$ are given by adding covariance terms in equations (3.7–3.8), yielding expressions as

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2 - 2\hat{\rho}(\hat{\theta}_1 - l_1)(u_2 - \hat{\theta}_2)} \tag{3.9}$$
$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2 - 2\hat{\rho}(u_1 - \hat{\theta}_1)(\hat{\theta}_2 - l_2)}, \tag{3.10}$$

where $\hat{\rho}$ is the estimated correlation coefficient between $\hat{\theta}_1$ and $\hat{\theta}_2$.

The above steps can be generalized to obtain CI for a linear combination of K parameters $\mathbf{c}^T\theta = \sum_{i=1}^{K} c_i\theta_i$ as

$$L = \sum_{i=1}^{K} c_i\hat{\theta}_i - \sqrt{\sum_{i=1}^{K}[c_i\hat{\theta}_i - \min(c_i l_i, c_i u_i)]^2} \tag{3.11}$$
$$U = \sum_{i=1}^{K} c_i\hat{\theta}_i + \sqrt{\sum_{i=1}^{K}[c_i\hat{\theta}_i - \max(c_i l_i, c_i u_i)]^2}, \tag{3.12}$$

where (l_i, u_i) is the CI for θ_i, $i = 1, \cdots, K$.

3.2.3 Confidence interval for a difference of two correlated AUCs

For two correlated AUCs derived from the same cases, once we get the confidence limits for each AUC, it is easy to get the CI for their difference using MOVER. Let (l_i, u_i) be the confidence limits for AUC_i, $i = 1,2$, and $\widehat{AUC}_i$, $i = 1,2$ are the related estimates. Applying equations (3.9–3.10), the confidence limits for $AUC_1 - AUC_2$ are:

$$L = \widehat{AUC}_1 - \widehat{AUC}_2 - \sqrt{(\widehat{AUC}_1 - l_1)^2 + (u_2 - \widehat{AUC}_2)^2 - 2\hat{\rho}(\widehat{AUC}_1 - l_1)(u_2 - \widehat{AUC}_2)} \quad (3.13)$$

$$U = \widehat{AUC}_1 - \widehat{AUC}_2 + \sqrt{(u_1 - \widehat{AUC}_1)^2 + (\widehat{AUC}_2 - l_2)^2 - 2\hat{\rho}(u_1 - \widehat{AUC}_1)(\widehat{AUC}_2 - l_2)}, \quad (3.14)$$

where $\hat{\rho}$ denotes the estimated correlation coefficient between the estimated AUC_1 and AUC_2.

3.2.4 Critical values for multiple comparisons

When we construct a confidence interval for a difference between two AUCs, we simply apply the MOVER to the limits for single AUCs obtained with critical values from the standard normal distribution. However, when we construct simultaneous confidence intervals for multiple comparisons of more than 2 AUCs, the confidence level for each contrast should be higher than 100(1-α)%, which suggests that the critical value for simultaneous confidence intervals must be higher than that from the standard normal distribution to maintain the overall coverage. Donner and Zou (2010) gave the details of obtaining appropriate critical value for multiple comparisons. Here we present a summary.

Denote $\theta = (\theta_1, \theta_2, ..., \theta_K)^T$ as the parameters of interest from K comparable groups, with $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K)$ and variance-covariance matrix given by V. For any I sets of linear functions ${c_i}^T\theta$, $i = 1,2,...,I$, using the multivariate central limit theorem, asymptotically, we have ${c_i}^T\hat{\theta} \sim N(c_i^T\theta, \text{var}(c_i^T\hat{\theta}))$. By definition of simultaneous confidence intervals, we need to select critical value c_α which satisfies

$$\Pr\left[c_i^T\hat{\theta} - c_\alpha\sqrt{\text{var}(c_i^T\hat{\theta})} < c_i^T\theta < c_i^T\hat{\theta} + c_\alpha\sqrt{\text{var}(c_i^T\hat{\theta})}, \text{ for all } i\right] = 1 - \alpha.$$

The above equation is

$$\Pr\left(-c_\alpha \le \frac{c_i^T\hat{\theta} - c_i^T\theta}{\sqrt{\text{var}(c_i^T\hat{\theta})}} \le c_\alpha \text{ for all } i\right) = 1-\alpha \tag{3.15}$$

Denote $T = (T_1, T_2, ..., T_I)$, with element

$$T_i = \frac{c_i^T\hat{\theta} - c_i^T\theta}{\sqrt{\text{var}(c_i^T\hat{\theta})}},\ i = 1, ..., I.$$

In order to obtain the critical value c_α, we first need to get the dispersion matrix for T. Let D be a diagonal matrix with $\text{var}(c_i^T\theta)$ as the ith element, i.e.

$$D = \text{diag}(\text{var}(c_1^T\theta), \text{var}(c_2^T\theta), \cdots, \text{var}(c_I^T\theta)),$$

then we have

$$T = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_I \end{pmatrix} = \begin{pmatrix} c_1^T(\hat{\theta}-\theta)\text{var}(c_1^T\hat{\theta})^{-1/2} \\ c_2^T(\hat{\theta}-\theta)\text{var}(c_2^T\hat{\theta})^{-1/2} \\ \vdots \\ c_I^T(\hat{\theta}-\theta)\text{var}(c_I^T\hat{\theta})^{-1/2} \end{pmatrix} = D^{-1/2}\begin{pmatrix} c_1^T(\hat{\theta}-\theta) \\ c_2^T(\hat{\theta}-\theta) \\ \vdots \\ c_I^T(\hat{\theta}-\theta) \end{pmatrix}$$

$$= D^{-1/2}C^T(\hat{\theta}-\theta),$$

where $C = (c_1, c_2, ..., c_I)$. Then the dispersion matrix of vector T is

$$R = D^{-1/2}C^T\text{Var}(\hat{\theta})CD^{-1/2},$$

which can be estimated by substituting estimates for unknown parameters.

Since T is asymptotically multivariate normal, we can get the critical value c_α by calculating the $1-\alpha$ quantile of the distribution of $\max_i|T_i|$ with simulation approach (Westfall *et al.*, 1999) or the inversion algorithm of the multivariate normal distribution (Genz, 1992).

Alternatively, the the SAS IML function PROBMC can provide critical value c_α for the Tukey-Kramer all pairwise comparisons and the Dunnett's comparisons with a standard. In this SAS function, the estimated covariances in the variance-covariance matrix $\widehat{\text{Var}}(\hat{\theta})$

are substituted as 0. It has been shown that this substitution turns out a conservative critical value especially when the estimated covariance between parameter estimates are not 0 (Nelson, 1989; Hayter, 1984).

For other cases that the critical value c_α is not available in PROBMC, Donner and Zou (2010) provided a SAS IML function to calculate c_α based on data generated from I-dimensional normal distribution with mean vector $\mathbf{0}$ and estimated variance matrix $\hat{R}$.

3.2.5 Simultaneous confidence intervals for comparing multiple AUCs using MOVER

In general, 5 types of multiple comparisons are frequently discussed for different problems of interest in application (Hsu, 1996).

1. Scheffé's all contrast comparisons are used to test all possible linear contrasts of groups, and the number of comparison groups can be infinite.

2. Tukey's all pairwise comparisons are applied to compare all possible pair-wise groups.

3. Bonferroni's subgroup selection comparisons are to compare some pre-selected groups, the number of these groups can exceed the set of pairwise comparisons specified in the Tukey procedure.

4. Dunnett's multiple comparisons with a standard are designed for situations where all groups are to be compared with one reference group.

5. Multiple comparisons with the best are useful to identify the best among groups.

Among these 5 comparisons, all pairwise comparisons and multiple comparisons with a standard are most common, and thus will be our focus.

Similar to the simultaneous confidence intervals for proportions (Donner and Zou, 2010), we can get the simultaneous confidence intervals for AUCs with 4 steps:

1. Obtain critical value c_α using SAS IML function PROBMC;

2. Calculate the variance-covariance matrix for $\widehat{\mathbf{AUC}} = \left(\widehat{\mathrm{AUC}}_1, \cdots, \widehat{\mathrm{AUC}}_k\right)$ using the formulae provided in DeLong *et al.* (1988);

3. Construct modified CIs by logit or inverse sinh transformations based on the critical value obtained in step 1 and variance covariance estimates in step 2;

4. Obtain simultaneous confidence intervals for multiple comparisons with the MOVER.

Here we use the same data in DeLong *et al.* (1988) to show how to get the two-sided simultaneous confidence intervals with our approach. First, we compare both *TP* and *ALB* to *K-G* using Dunett's comparisons with a standard, and the logit transformation is chosen in calculating the improved CI for AUC.

1. We first get the appropriate critical value c_α for two-sided Dunnett's comparisons with a standard with 3 measurement groups using

$$\begin{aligned} c_\alpha &= \text{probmc}(\text{“Dunnett2”}, ., \text{confidence level}, ., \text{number of groups} - 1) \\ &= \text{probmc}(\text{“Dunnett2”}, ., 0.95, ., 2) \end{aligned}$$

   resulting in $c_\alpha = 2.2121$.

2. Estimates of 3 AUCs

$$\mathbf{AUC} = (\mathrm{AUC}_{TP},\ \mathrm{AUC}_{ALB},\ \mathrm{AUC}_{K-G}),$$

   are given by (0.6478, 0.7366, 0.7258), with the estimated variance-covariance matrix:

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{AUC}}) = \begin{pmatrix} 0.1^2 & & \\ 0.0076 & 0.0927^2 & \\ 0.0028 & 0.0033 & 0.1028^2 \end{pmatrix}.$$

3. Using the critical value in step 1 and the estimates in step 2, we get the confidence limits for logit(AUC_{K-G}) as

$$\begin{aligned}(l,\, u) &= \text{logit}(\widehat{\text{AUC}}_{K-G}) \pm c_\alpha \frac{\widehat{s.e.}(\widehat{\text{AUC}}_{K-G})}{\widehat{\text{AUC}}_{K-G}(1-\widehat{\text{AUC}}_{K-G})} \\ &= \text{logit}(0.7258) \pm 2.2121 \times \frac{0.1028}{0.7258(1-0.7258)} \\ &= 0.9734 \pm 1.1427 \\ &= (-0.1693,\, 2.1162). \end{aligned} \tag{3.16}$$

So the logit Wald-type CI for AUC_{K-G} is

$$\begin{aligned}(L,\, U)_{K-G} &= \left(\frac{e^l}{1+e^l},\, \frac{e^u}{1+e^u}\right) \\ &= \left(\frac{e^{-0.1693}}{1+e^{-0.1693}},\, \frac{e^{2.1162}}{1+e^{2.1162}}\right) \\ &= (0.5422,\, 0.8925). \end{aligned} \tag{3.17}$$

Similarly, the CIs for AUC_{ALB} and AUC_{TP} are given by

$$(L,\, U)_{ALB} = (0.5071,\, 0.8895),$$

$$(L,\, U)_{TP} = (0.4132,\, 0.8277).$$

4. Applying equations (3.9–3.10) to the former 3 CIs gives the simultaneous confidence intervals of Dunett's multiple comparisons with a standard for the 3 diagnostic tests treating K-G score as a control as follows

$$ALB\ vs\ K-G = (-0.2192,\, 0.206),$$

$$TP\ vs\ K-G = (-0.3264,\, 0.127).$$

For example, the lower limit for the difference of *ALB* and *K-G* is given by

$$\begin{aligned} & L_{ALB\ vsK-G} \\ = \; & 0.7366 - 0.7258 - \sqrt{0.2295^2 + 0.1667^2 - 2 \times 0.34 \times 0.2295 \times 0.1667} \\ = \; & -0.2192. \end{aligned} \tag{3.18}$$

Based on the simultaneous confidence intervals, we conclude that *K-G* score is not significantly different from either *ALB* score or *TP* test at the 5% overall significant level.

Similarly, we compare each two of the 3 diagnostic measurements using two sided Tukey's all pairwise comparison, and the inverse sinh transformation is used to calculate the improved CI for a single AUC.

1. We first get the appropriate critical value c_α for two-sided Dunnett's comparisons with a standard with 3 measurement groups using

$$\begin{aligned} c_\alpha &= \text{probmc}(\text{“range”}, ., \text{confidence level}, ., \text{number of groups})/\sqrt{2} \\ &= \text{probmc}(\text{“range”}, ., 0.95, ., 3)/\sqrt{2} \end{aligned}$$

resulting in $c_\alpha = 2.3437$.

2. Estimates of 3 AUCs are

$$\widehat{\mathbf{AUC}} = (\widehat{\text{AUC}}_{TP},\ \widehat{\text{AUC}}_{ALB},\ \widehat{\text{AUC}}_{K-G}) = (0.6478,\ 0.7366,\ 0.7258),$$

with the estimated variance-covariance matrix:

$$\widehat{\text{Var}}(\widehat{\mathbf{AUC}}) = \begin{pmatrix} 0.1^2 & & \\ 0.0076 & 0.09^2 & \\ 0.0028 & 0.0033 & 0.1028^2 \end{pmatrix}.$$

3. Using the critical value in step 1 and the estimates in step 2, we get the confidence limits for logit(AUC_{K-G}) based on the inverse sinh transformation as:

$$\begin{aligned} (l_n,\ u_n) &= \text{logit}(\widehat{\text{AUC}}_{K-G}) \pm 2\text{arcsinh}\left(\frac{c_\alpha}{2} \cdot \frac{\widehat{s.e.}(\widehat{\text{AUC}}_{K-G})}{\widehat{\text{AUC}}_{K-G}(1-\widehat{\text{AUC}}_{K-G})}\right) \\ &= \text{logit}(0.7258) \pm 2\text{arcsinh}\left(\frac{2.3437}{2} \times \frac{0.1028}{0.7258(1-0.7258)}\right) \\ &= 0.9734 \pm 2\text{arcsinh}(0.6053) \\ &= (-0.1733,\ 2.1201). \end{aligned} \tag{3.19}$$

So the CI for AUC_{K-G} based on inverse sinh transformation is

$$\begin{aligned}(L,\, U)_{K-G} &= \left(\frac{e^{l_n}}{1+e^{l_n}},\, \frac{e^{u_n}}{1+e^{u_n}}\right) \\ &= \left(\frac{e^{-0.1733}}{1+e^{-0.1733}},\, \frac{e^{2.1201}}{1+e^{2.1201}}\right) \\ &= (0.4568,\, 0.8928). \end{aligned} \tag{3.20}$$

Using the same procedure, the modified CIs for AUC_{ALB} and AUC_{TP} with inverse sinh transforamtion are given by

$$(L,\, U)_{ALB} = (0.4900,\, 0.8906),$$

$$(L,\, U)_{TP} = (0.4068,\, 0.8315).$$

4. Applying equations (3.9–3.10) to the former 3 CIs gives the simultaneous confidence intervals of Tukey's all pairwise comparisons for the 3 diagnostic tests as follows

$$ALB\ vs\ K-G = (-0.2356,\, 0.2714),$$

$$TP\ vs\ K-G = (-0.3314,\, 0.2038),$$

$$ALB\ vs\ TP = (-0.0536,\, 0.2335).$$

For example, the lower limit for the difference of *ALB* and *K-G* is given by

$$\begin{aligned} & L_{ALB\ vs\ K-G} \\ = \ & 0.7366 - 0.7258 - \sqrt{0.2466^2 + 0.167^2 - 2 \times 0.34 \times 0.2466 \times 0.167} \\ = \ & -0.2356. \end{aligned} \tag{3.21}$$

Based on the simultaneous confidence intervals, we can conclude that there is no evidence to suggest that these three diagnostic tests are different from each other at the 5% overall significant level. However, judging from the interval width, it is clear that the study had little power to detect meaningful difference.

Chapter 4

SIMULATION STUDY

The methodology presented in the previous chapter was developed using large sample theory. Therefore, its performance in finite samples must be evaluated before applying to practice. To investigate and compare the performance of these three CI methods for both single and multiple AUCs, we carried out a series of simulation studies. Our evaluation focused on the extent to which the empirical coverage of the confidence interval matched with the nominal level, while tail errors and confidence interval width were regarded as secondary criteria.

4.1 Study design

4.1.1 Selection of parameter values

The parameters examined were the number of diagnostic tests (k), sample size n_X for diseased group X and n_Y for non-diseased group Y, the correlation coefficient ρ between tests, the mean test outcomes for diseased subjects X, the variances of test outcomes for the diseased and non-diseased subjects (σ_X^2, σ_Y^2), and the area under the ith ROC curve (AUC_i, $i = 1, \cdots, k$).

In these examined parameters, the values of k were selected as 1, 2, 4 and 7. For $k = 1$, we assessed the performance of 3 CI procedures for a single AUC. For $k = 2$, we compare the corresponding methods using the MOVER for a difference between two AUCs. And for k= 4 and 7, we compare 3 simultaneous confidence intervals approaches for 4 and 7 AUCs, respectively. The sample size combinations (n_Y, n_X) were chosen as small balanced

(25, 25), moderate balanced (50, 50), large balanced (100, 100), and unbalanced (25, 50), (25, 75), (50, 100). The strength of the correlation ρ between the test groups were set as weak (0.2), medium (0.5) and strong (0.8). The mean and variance of test values for the 'non-diseased' subjects were chosen as $\mu_Y = 0$ and $\sigma_Y^2 = 1$. While the mean of test values for the 'diseased' subjects μ_X was given by $\mu_X = [\Phi^{-1}(\text{AUC})\sqrt{\sigma_X^2 + \sigma_Y^2}] + \mu_Y$, where Φ^{-1} is the inverse cumulative density function of standard normal distribution. The variance for the 'diseased' subjects σ_X^2 was chosen as 1, 2 and 3 so that it was equal, moderately close and the least close to σ_Y^2.

When $k = 1$, since values of AUC can range between 0.5 to 1, we chose true AUC value from 0.6 to 0.9 in step of 0.1 to imitate a wide range of test accuracy in practice from low to high.

For multiple comparisons k = 2, 4 and 7, a specified group of values for AUC may not represent the whole possible situation, we randomly generated values of AUC from uniform distribution $U(0.5, 1)$. To have a panoptic assessment of the performance, we considered 1000 sets of random values for AUCs and interpret their results with basic statistics (quartiles, mean, extrema).

Tukey's all pair-wise and Dunett's comparisons with a standard were considered to construct the related simultaneous confidence intervals for $k > 2$. For Dunett's comparisons, both one sided and two sided simultaneous confidence intervals were discussed. The nominal levels were set at $\alpha = 0.10$ and 0.05.

4.1.2 Methods compared

The methods considered include the Wald method, and two methods based on logit and inverse sinh transformation of AUCs. Variance estimate for the Wald method was obtained using results of DeLong *et al.* (1988), and that for the transformed AUC were obtained using the delta method. The performances of the three procedures were evaluated in terms of coverage probability, the symmetry of tail errors (i.e. non-coverage probabilities), and average interval width.

For confidence intervals for a difference between two AUCs, we applied the MOVER to each of the three methods for a single AUC.

Three simultaneous confidence intervals approaches for multiple AUCs were constructed using the MOVER in combination of appropriate critical values for multiple comparisons.

The coverage probability for a single AUC and a difference of two AUCs is defined as the percentage of CI that covers the true AUC value, the coverage probability for a difference of two AUCs represents the percentage of CI that covers the difference of two true AUC values, and the coverage probability for multiple AUCs means the percentage of simultaneous confidence intervals that cover the true values of linear comparisons of AUCs simultaneously. Recognizing that no hard rules exist, we used empirical coverage probability to be in the range of 94.57–95.43%, which is $0.95 \pm 1.96\sqrt{(0.95 \times 0.05)/10000}$ in the evaluation when the number of simulation runs are 10000 and for the 95% CI with 1000 runs, we used the target range of 93.65–96.35%.

The average interval width for a single AUC or a difference between two AUCs is the average width of CIs based on the simulation runs, and the average interval width for multiple AUCs is defined as the average interval width for each contrast in multiple comparisons. For a certain coverage probability, we prefer confidence interval with the least average width as it represents higher precision.

The left and right tail errors were used to measure the symmetry of tail errors. The left tail error was defined as the proportion that the true value is less than the lower limit, and the right tail error was the proportion that the true value of the parameter is greater than the upper limit. For a $100(1-\alpha)\%$ confidence interval $(l,\ u)$ of a parameter θ, it should have $Pr(l \geq \theta) = Pr(u \leq \theta) = \alpha/2$ ensuring that only extreme values are excluded from the interval. As a result, the symmetry of tail errors are desirable for confidence interval construction. We did not compute tail errors for simultaneous confidence intervals because both left and right tail errors may occur simultaneously in one set of simultaneous confidence intervals.

4.2 Data generation

4.2.1 Data generation for a single AUC

Without loss of generality, the data for non-diseased subjects were generated from standard normal distribution, and the outcomes for diseased subjects were generated from $N(\mu_X, \sigma_X^2)$, where σ_X^2 was chosen as 1, 2 and 3 which are equal, closely equal and far close to the related variance in non-diseased group. For a given value of AUC, the value for μ_X was given by $\mu_X = \Phi^{-1}(\text{AUC})\sqrt{\sigma_X^2 + \sigma_Y^2}$, where Φ^{-1} is the inverse standard normal cumulative distribution function.

For each parameter combination, a total of 10,000 replicates were conducted. Both 90% and 95% confidence intervals were constructed, using three approaches discussed in Chapter 3.

4.2.2 Data generation for multiple AUCs

For comparing k diagnostic tests derived from the same subjects, the data of all subjects from non-diseased and diseased groups were both generated from multiple normal distribution. The k test results for 'non-diseased' subjects Y were obtained from k dimensional normal distribution with mean vector $\mu_Y = (0, \cdots, 0)_{1\times k}$ and variance-covariance matrix $V_Y = (1-\rho)I_k + \rho J_k$, where I_k and J_k are k-dimensional identity matrix and $k \times k$ unity matrix. The k test results for 'diseased' subjects X were generated from k dimensional normal distribution with dispersion matrix $V_Y = (i-\rho)I_k + \rho J_k$, where ρ was chosen as 0.2, 0.5 and 0.8 showing weak, moderate and strong correlations between tests and i was taken as 1, 2 and 3 indicating the different values of variance of Y for each diagnostic test. The mean vector for diseased subjects can be calculated by the given values of AUCs for k tests and the known information for test outcomes of X and Y. For example, the j-th mean of test outcomes for diseased subjects was calculated as: $\mu_X{}^j = \Phi^{-1}(\text{AUC}_j)\sqrt{V_X(j,j) + V_Y(j,j)}$, where $V_X(j,j)$ and $V_Y(j,j)$ are the j-th element of the diagonal of the related dispersion

matrix.

To have a broad view of the performance, we randomly sampled 1000 sets of true values for AUCs. For each set of the true values for AUCs, 1000 datasets were generated and simultaneous confidence intervals were constructed to determine coverage rate and average interval width. We interpreted the results for the 1000 sets of true values for AUCs using coverage probability and average width with summary statistics such as quartiles, mean, minimum and maximum values.

4.3 Results for a single AUC

Simulation results for comparing the three confidence interval procedures for a single AUC is shown in Table 4.1, in which only results for $\alpha = 0.05$ and $\sigma_X^2 = 2$ are presented. The results for $\alpha = 0.1$ are similar to those for $\alpha = 0.05$, and that for $\sigma_X^2 = 2$ and 3 are similar to those for $\sigma_X^2 = 1$.

Results show that the Wald-type procedure tends to be anti-conservative. The deficiency is more profound for higher values of AUC. For example, when sample sizes for diseased subjects X and non-diseased subjects Y are small $n_X = n_Y = 25$, the coverage probability can be as low as 90.35% when the true AUC value is 0.9. Even when sample sizes are quite large $n_Y = n_X = 100$, the coverage rate of Wald-type CI is lower than the target range of 94.57% to 95.43%.

Procedures based on logit and inverse sinh transformations perform better than Wald approach in terms of coverage probability. For large or balanced sample sizes, the coverage rates generated from these two approaches are both very close to the nominal level, and compared to these two modified CIs, inverse sinh transformed CI performs better with coverage rate closer to the nominal level and slightly shorter interval width as well. However, when sample sizes are small and the numbers of subjects in non-diseased group and diseased group are unequal, the inverse sinh method turns to be liberal. For example, when $n_Y = 25$, $n_X = 75$ and AUC = 0.9, the inverse sinh method gives the coverage probability

as 93.7%, below the target range of 94.57% to 95.43%.

The Wald procedure is noticeably unbalanced in tail errors, and the difference of tail errors increase as the true AUC value increases for all sample sizes, the two transformed CIs improve the symmetry of tail errors and inverse sinh type CI is seen to be slightly more balanced. For instance, when $n_Y = 25$, $n_X = 25$ and AUC $= 0.8$, the asymmetry of tail errors is serious with 6.04% left tail error rate and 0.94% right rate, after inverse sinh transformation, the tail errors are fairly balanced with 2.97% left error rate and 2.72% right error rate.

The three methods provide similar interval width for the same parameter combination, and all three methods give shorter interval width when the sample sizes are larger or the AUC value is bigger. Although the width of inverse sinh type CI is shorter than the logit type CI in many cases, the difference is small.

4.4 Results for a difference of two AUCs

We only report the results for $\alpha = 0.05$, $\sigma_X^2 = 1$ and sample sizes $n_X = n_Y = 25, 50, 100$, and $n_X = 50, n_Y = 25$. Only the left tail error rates are shown in the figures for brevity. Since the results for $\alpha = 0.1$ are similar to those for $\alpha = 0.05$, the results for $\sigma_X^2 = 2$, 3 are similar to those for $\sigma_X^2 = 1$, and the results for $n_X = 75, n_Y = 25$, and $n_X = 100, n_Y = 50$ are similar to the results for $n_X = 50, n_Y = 25$. The results for right tail error rates are similar to the results for left tail error rates.

4.4.1 For $n_X = n_Y = 25$

Figure 4.1 shows the box plots of coverage probability and left tail error rate as well as average interval width using the three CI methods and the MOVER for CI of a difference of two AUCs when sample sizes are $n_X = n_Y = 25$. Figures 4.1 (a)–(c) suggest that for the three correlation coefficients considered among measurements by two diagnostic tests, the Wald method generates CI with at least 25% of coverage probabilities outside the target

range, especially when $\rho = 0.8$, some coverage probabilities are even below 92%. For the two methods based on transformation, when ρ = 0.2 and 0.5, the coverage rates are close to the target range. The inverse sinh approach performs better in terms of coverage probability, in that most of the coverage probabilities fall inside the range, and those fell outside the range are still close to the target range. Furthermore, the median coverage rate obtained from inverse sinh transformation is virtually identical to the nominal level. When $\rho = 0.8$, two modified methods produce the most of coverage rates that are above the lower bound of the target level of 93.65%. The procedure based on the transformation provides conservative coverage percentages.

Figures 4.1 (d)–(f) indicate that Wald method provides unbalanced tail errors, especially when $\rho = 0.8$, some left error rates are even higher than 6%. Two transformed methods give similar range of tail rates, and the inverse sinh transformation gives the left tail error rates more symmetric around 2.5% than logit method.

Figures 4.1 (g)–(i) show that all three methods provide shorter interval width as ρ increases. For example, the maximum interval width for logit method decreases from more than 0.4 when $\rho = 0.2$ to 0.3 when $\rho = 0.8$. For a fixed ρ, the method based on inverse sinh transformation has slightly shorter interval width than logit method.

4.4.2 For $n_X = 50$, $n_Y = 25$

Figure 4.2 presents the boxplots of coverage probability, left tail error rate and average interval width for CI construction of a difference of two AUCs when the sample size are unbalanced $n_X = 50, n_Y = 25$. Figures 4.2 (a)–(c) suggest that, similar to the results for $n_X = n_Y = 25$, when ρ = 0.2 and 0.5, the inverse sinh method has the best performance among these three methods in terms of coverage rate, when $\rho = 0.8$, two methods based on transformation have all the coverage rates above the lower bound of target range, but the two methods are conservative especially for logit method.

The left tail errors are shown in Figures 4.2 (d)–(f), Wald method has the largest range of tail errors for all three correlation coefficients. When $\rho = 0.8$, the left tail error rate for

Wald method can be as high as 6%. Two transformed methods give similar range of error rates.

Figures 4.2 (g)–(i) display that Wald method has the slightly wider range of interval width compared to the two modified methods, and two transformed methods give similar interval width.

4.4.3 For $n_X = n_Y = 50$

Figure 4.3 gives the boxplots of coverage probability, left tail error rate and average interval width for CI construction of a difference of two AUCs when the sample sizes are moderate and balanced $n_X = n_Y = 50$. Figures 4.3 (a)–(c) show that when $\rho = 0.2$ and 0.5, the inverse sinh method has the best performance with the most of the coverage rates inside the target region and those fell outside the range are still close to the target range. When $\rho = 0.8$, the Wald method provides coverage close to the nominal level, and two transformed methods tend to be conservative especially for logit method.

Figures 4.3 (d)–(f) show that similar to the results for small sample size combination, Wald method has the widest range of left tail error rate and two transformed methods have the similar performance.

Figures 4.3 (g)–(i) indicate that three methods have similar interval width, and the width decreases as ρ increases.

4.4.4 For $n_X = n_Y = 100$

Figure 4.4 shows the boxplots of coverage probability, left tail error rate and average interval width for CI construction of a difference of two AUCs when the sample sizes are large and balanced $n_X = n_Y = 100$. Figures 4.4 (a)–(c) show that when $\rho = 0.2$ and 0.5, the three methods have similar performance that the coverage rates are in the target range in most cases and those fell outside the range are still close to the target range, furthermore, the median coverage probabilities are all close to 95%. When $\rho = 0.8$, the coverage rates from

Wald method is closest to the nominal level, and the two methods based on transformation are slightly conservative.

Figures 4.4 (e)–(f) show that the Wald method has the widest range of left tail error rate especially for large ρ, two transformed methods have similar performance for tail error rates.

Figures 4.4 (g)–(i) indicate that three methods have similar interval width, and the width decreases as ρ increases.

4.5 Results for multiple comparisons of AUCs

For brevity, we only present the results for two sided simultaneous confidence intervals of Tukey's all pairwise comparisons (Figures 4.5, 4.7, 4.9 and 4.11) and Dunett's multiple comparisons with a standard (Figures 4.6, 4.8, 4.10 and 4.12) when $\alpha = 0.05$, $\sigma_X^2 = 1$, $k = 4$, and sample sizes $n_X = n_Y = 25, 50, 100$, and $n_X = 50, n_Y = 25$. Since the results for $\alpha = 0.1$ are similar to those for $\alpha = 0.05$, the results for $\sigma_X^2 = 2$, 3 are similar to those for $\sigma_X^2 = 1$, the results for $k = 7$ are similar to those for $k = 4$, and the results for $n_X = 75, n_Y = 25$, and $n_X = 100, n_Y = 50$ are similar to the results for $n_X = 50, n_Y = 25$, and the results for one sided simultaneous confidence intervals are similar to those for two sided simultaneous confidence intervals .

4.5.1 For $n_X = n_Y = 25$

Figure 4.5 shows the box plots of coverage probability and average interval width using the three SCI approaches for all pairwise comparisons when sample sizes are $n_X = n_Y = 25$. Figures 4.5 (a)–(c) indicate that for all three correlation coefficients among measurements by two diagnostic tests, the Wald method results in half of coverage probabilities below the target level, especially when $\rho = 0.8$, some coverage rates for Wald simultaneous confidence intervals are even below 88%. For the two methods based on transformation of AUCs, when $\rho = 0.2$ and 0.5, their coverage rates are closer to the target region than those

from Wald method, and the inverse sinh transformation performs better in terms of coverage rate, in that most of the coverage probabilities fall inside the range, and those fell outside the range are still close to the target range. Furthermore, the median coverage rate obtained from inverse sinh transformation is virtually identical to the nominal level. When $\rho = 0.8$, two transformed methods provide coverage rates that are above the lower bound of the target region (93.65%), but the simultaneous confidence intervals are conservative especially for logit method.

Figures 4.5 (d)–(f) suggest that all three methods provide shorter average interval width as ρ increases. For example, the maximum interval width for Wald method keeps decreasing from larger than 0.55 when $\rho = 0.2$ to less than 0.45 when $\rho = 0.5$ and it decreases below 0.35 when ρ is 0.8. For a given correlation, two transformed methods can provide tighter intervals than those from the Wald method, and compared with the two transformed simultaneous confidence intervals , the estimated interval width from inverse sinh transformation is slightly shorter than that from logit transformation.

Figure 4.6 illustrates the box plots of coverage probability and average interval width derived from the three SCI approaches for multiple comparisons with a standard when sample sizes are $n_X = n_Y = 25$. The plots suggest that logit transformation has the best performance in terms of coverage probabilities and interval width when ρ = 0.2 and 0.5. When $\rho = 0.8$, two transformed methods partly improve the Wald method with the minimum coverage rates closer to the nominal level, but two modified simultaneous confidence intervals are quite conservative in many cases. Figures 4.6 (d)–(f) indicate that the interval width also decreases as ρ increases.

4.5.2 For $n_X = 50$, $n_Y = 25$

Figures 4.7 (a)–(b) show that the box plots of coverage probabilities for the three SCI approaches for all pairwise comparisons when sample sizes are $n_X = 50$, $n_Y = 25$ and ρ = 0.2, 0.5 and 0.8. The plots show that the Wald method tends to provide confidence intervals having coverage probabilities lower than the nominal level in at least half of the

cases, especially, when $\rho = 0.5$, where some coverage rates are even lower than 90%. For the two methods based on transformation, when ρ = 0.2 and 0.5, the simultaneous confidence intervals result in coverage rates closer to the nominal level than Wald method, where two modified methods give the simultaneous confidence intervals with most of the coverage probabilities falling inside the range, and those fell outside the range are close to the target range. When $\rho = 0.8$, two methods based on transformation produce simultaneous confidence intervals with most of the coverage rates above the lower bound of target region (93.6), but the simultaneous confidence intervals are conservative particularly for logit method.

Figures 4.7 (d)–(f) show consistently that the interval widths from all three methods decrease as ρ increases. As shown in the graphs, when $\rho = 0.2$, the maximum interval width for Logit method is around 0.48, and it reduces to 0.37 when $\rho = 0.5$, and when $\rho = 0.8$, the maximum interval width is only 0.27. For a fixed ρ, two methods based on transformation provide tighter interval width than that of Wald method, and compared to the two transformed simultaneous confidence intervals , the estimated interval width from inverse sinh transformation is slightly shorter than that from logit transformation.

Figure 4.8 displays box plots for coverage probabilities and average interval width obtained from the three SCI approaches for multiple comparisons with a standard when sample sizes are $n_X = 50$, $n_Y = 25$. The results are similar to those of all pairwise comparisons. The figure indicates that two transformations perform better than Wald method, but when $\rho = 0.8$, two modified simultaneous confidence intervals give conservative overall coverage, especially for Logit method. Figures 4.8 (d)–(f) suggest very similar tendency of the interval width to those for the all pairwise comparisons at the same sample size as shown in Figure 4.7.

4.5.3 For $n_X = n_Y = 50$

Figure 4.9 gives box plots of coverage probabilities and average interval width for the three SCI approaches for all pairwise comparisons at balanced medium sample sizes $n_X = n_Y =$

50. Figures 4.9 (a)–(b) show that when ρ = 0.1 and 0.5, the proportion of coverage probability that is outside the target range of coverage probability, (93.6%, 96.4%), is higher for the Wald method than for the other two transformed approaches. The inverse sinh method performs best among these three methods in which most coverage rates are included in the target range and those outside the range are still close to the target range. When $\rho = 0.8$, Figure 4.9 (c) indicates that more than half of the coverage rates of Wald method are below the nominal level and some coverage rates are even less than 91%, two transformed methods give the simultaneous confidence intervals with the most of the coverage probabilities larger than the lower bound of the target range, but some are too conservative especially for logit transformation.

Regarding the average interval width, Figures 4.9 (d)–(f) show the similar results to those from other parameter combinations. All three methods give narrower average interval width when ρ is larger. For instance, when $\rho = 0.2$, the maximum interval width for Wald method is around 0.4 and it keeps decreasing to around 0.22 when ρ reaches 0.8. For a given correlation ρ, two transformed methods provide more intent interval width than that of Wald method, and compared with the two transformed simultaneous confidence intervals , the estimated interval width from inverse sinh transformation is slightly shorter than that from logit transformation.

Figure 4.10 presents the box plots of coverage probability and average interval width derived from the three SCI approaches for multiple comparisons with a standard when sample sizes are $n_X = n_Y = 50$. The graphs give the similar results to the all pairwise comparisons for the same sample sizes. Figures 4.10 (a)-(b) suggest that logit transformation has the best performance in terms of coverage probabilities when ρ = 0.2 and 0.5. Figure 4.10 (c) shows that When $\rho = 0.8$, two transformed methods partly improve the Wald method with the minimum coverage rates closer to the nominal level, but two modified simultaneous confidence intervals are too conservative in some cases especially for logit method. Figures 4.10 (d)–(f) indicate the very similar tendency of the interval width to those for the all pairwise comparisons at the same sample size as shown in Figure 4.9.

4.5.4 For $n_X = n_Y = 100$

Figure 4.11 gives the box plots of coverage probability and average interval width derived from the three SCI approaches for all pairwise comparisons when sample sizes are $n_X = n_Y = 25$. Figures 4.11 (a)–(c) show that when $\rho = 0.2$ and 0.5, the inverse sinh method has the best performance among the three method with most of the coverage rates are inside the target range, and those outside the range are still close to the target range. When $\rho = 0.8$, the Wald method has the closest coverage rates to the target range, and both two transformed methods give conservative results.

For the average interval width, Figures 4.11 (d)–(f) suggest that all three methods give shorter average interval width when ρ is larger. For example, the maximum interval width for Wald method is around 0.3 when $\rho = 0.2$ and it keeps decreasing to lower than 0.2 when $\rho = 0.8$. For a fixed ρ, all three approaches give very similar interval width.

Figure 4.12 illustrates the box plots of coverage probability and average interval width derived from the three SCI approaches for comparisons with a standard when sample sizes $n_X = n_Y = 100$. As shown in the graph, it gives the similar results to the all pairwise comparisons for the same sample size. The figure suggests that logit transformation has the best performance in terms of coverage probabilities and interval width when $\rho = 0.2, 0.5$. When $\rho = 0.8$, the Wald method performs best among the three method and both two transformed methods give conservative simultaneous confidence intervals . Figures 4.12 (d)–(f) indicate the very similar tendency of the interval width to those for the all pairwise comparisons at the same sample size as shown in Figure 4.11.

4.6 Summary

We compared the Wald, logit and inverse sinh approaches for constructing confidence interval of a single AUC and a difference of two AUCs. We also compared the three approaches for simultaneous confidence intervals of multiple comparisons of correlated AUCs. Confidence coverage probabilities, interval width and symmetry of tail errors were used as

criteria.

For a single AUC procedure, when the number of subjects are equal in both diseased and normal groups, the method based on inverse sinh transformation performs best amongst the three methods concerning the coverage rate and tail errors for all AUC values. Specifically, even with group size of 25, the inverse sinh transformation still gives CI having estimated coverage rate very close to the nominal level and balanced tail errors. When group sizes are unequal, inverse sinh approach performs well in most cases except when AUC is as large as 0.9 and sample sizes are small. The interval width decreases as the sample sizes or the true AUC value increase.

For a difference of two AUCs, the simulation results show that when ρ is small to moderate, the method based on inverse sinh transformation has the best performance among the three methods in terms of coverage probability and tail errors for all sample sizes combination. When ρ is as large as 0.8, the Wald method provides coverage rates that are close to the nominal level, and the two methods based on transformation are conservative particularly for small sample size study. The Wald method generates the widest interval width compared to the other two methods based on transformation, and the interval width decreases as correlation or sample sizes increase.

For multiple comparisons of correlated AUCs, the results for all pairwise comparisons and comparisons with a standard are very similar for a same parameter combination. When sample sizes are small to medium and the correlation coefficient between test groups is small to moderate, the proportion of coverage probabilities that is beyond the target range of coverage probability, (93.65%, 96.35%), is higher for the Wald method than that for the two transformed approaches. Most of the coverage probabilities of the two transformed methods fall in the target interval, and those not included in the target range are close to the range. Compared to the two transformations, the inverse sinh transformation performs better in most cases of parameter combinations in terms of coverage probability. When correlation coefficient ρ is as large as 0.8, both two transformed methods generate simultaneous confidence intervals with most of coverage probabilities above the target range, but

two methods are conservative in some cases.

For large sample sizes with small to moderate correlation among measurements, all three methods give similar coverage rates which are close to the nominal level. When correlation is as high as 0.8, both logit and inverse sinh methods provide conservative simultaneous confidence intervals and the latter method is less conservative than the former one with closer median coverage rate to the nominal level and less proportion of coverage rates above the target range.

The average width of the simultaneous confidence intervals tend to decrease as correlation increases for a given sample size. When sample sizes are small to medium, two approaches based on transformation give intervals with less variation, and the inverse sinh approach provides a slightly shorter interval width compared to logit method. When sample sizes are getting larger, the interval widths in three approaches are nearly equal.

In summary, in terms of coverage rates, tail error rates and average interval width, the performance of inverse sinh transformation method is preferable among the three methods for both confidence interval and simultaneous confidence intervals construction, especially with small to moderate correlation coefficient among test outcomes.

Table 4.1: Comparative performance of the Wald method, logit method and inverse sinh method in construction of a two-sided 95% confidence interval for the area under ROC curve based on 10,000 runs

		Wald	Logit	Inverse Sinh
(n_Y, n_X)	AUC	CV (ML, MR)% WD	CV (ML, MR)% WD	CV (ML, MR)% WD
(25, 25)	0.6	94.16 (3.88, 1.96) 0.32	96.03 (2.06, 1.91) 0.31	95.71 (2.28, 2.01) 0.30
	0.7	93.93 (4.56, 1.51) 0.29	95.93 (1.96, 2.11) 0.29	95.33 (2.36, 2.31) 0.28
	0.8	93.02 (6.04, 0.94) 0.24	96.07 (1.83, 2.10) 0.25	95.41 (2.23, 2.36) 0.24
	0.9	90.35 (9.21, 0.44) 0.16	95.36 (2.20, 2.44) 0.18	94.31 (2.97, 2.72) 0.17
(25, 50)	0.6	93.95 (3.68, 2.37) 0.27	95.23 (2.35, 2.42) 0.27	94.94 (2.54, 2.52) 0.26
	0.7	93.32 (4.76, 1.92) 0.25	95.11 (2.43, 2.46) 0.25	94.73 (2.66, 2.61) 0.24
	0.8	92.50 (6.10, 1.40) 0.21	95.14 (2.51, 2.35) 0.21	94.73 (2.75, 2.52) 0.21
	0.9	90.22 (9.12, 0.66) 0.14	94.63 (2.77, 2.60) 0.15	93.70 (3.48, 2.82) 0.15
(25, 75)	0.6	94.27 (3.48, 2.25) 0.26	95.24 (2.44, 2.32) 0.25	95.03 (2.56, 2.41) 0.25
	0.7	93.93 (4.22, 1.85) 0.24	95.05 (2.46, 2.49) 0.23	94.86 (2.58, 2.56) 0.23
	0.8	92.98 (5.75, 1.27) 0.20	95.16 (2.40, 2.44) 0.20	94.76 (2.67, 2.57) 0.19
	0.9	90.71 (8.60, 0.69) 0.13	94.42 (3.13, 2.45) 0.14	93.87 (3.54, 2.59) 0.14
(50, 50)	0.6	94.46 (3.44, 2.10) 0.22	95.20 (2.58, 2.22) 0.22	95.06 (2.63, 2.31) 0.22
	0.7	94.29 (4.02, 1.69) 0.20	95.35 (2.35, 2.30) 0.20	95.07 (2.48, 2.45) 0.20
	0.8	93.57 (5.12, 1.31) 0.17	95.28 (2.20, 2.52) 0.17	95.01 (2.39, 2.60) 0.17
	0.9	92.03 (7.12, 0.85) 0.12	94.98 (2.46, 2.56) 0.12	94.55 (2.76, 2.69) 0.12
(50, 100)	0.6	95.00 (3.01, 1.99) 0.19	95.53 (2.31, 2.16) 0.19	95.35 (2.40, 2.25) 0.19
	0.7	94.88 (3.56, 1.56) 0.18	95.49 (2.31, 2.20) 0.18	95.32 (2.43, 2.25) 0.17
	0.8	94.37 (4.42, 1.21) 0.15	95.61 (2.30, 2.09) 0.15	95.40 (2.46, 2.14) 0.15
	0.9	93.20 (6.07, 0.73) 0.10	95.46 (2.38, 2.16) 0.10	95.20 (2.55, 2.25) 0.10
(100, 100)	0.6	94.61 (3.04, 2.35) 0.16	95.11 (2.31, 2.58) 0.16	95.02 (2.37, 2.61) 0.15
	0.7	94.34 (3.47, 2.19) 0.14	95.02 (2.38, 2.60) 0.14.	94.93 (2.40, 2.67) 0.14
	0.8	93.93 (4.18, 1.89) 0.12	94.99 (2.32, 2.69) 0.12	94.89 (2.38, 2.73) 0.12
	0.9	93.19 (5.52, 1.29) 0.08	94.67 (2.39, 2.94) 0.08	94.54 (2.48, 2.98) 0.08

Note: CV means coverage probability. ML and MR denote the interval misses the true AUC from the left and the right, respectively. WD denotes average interval width.

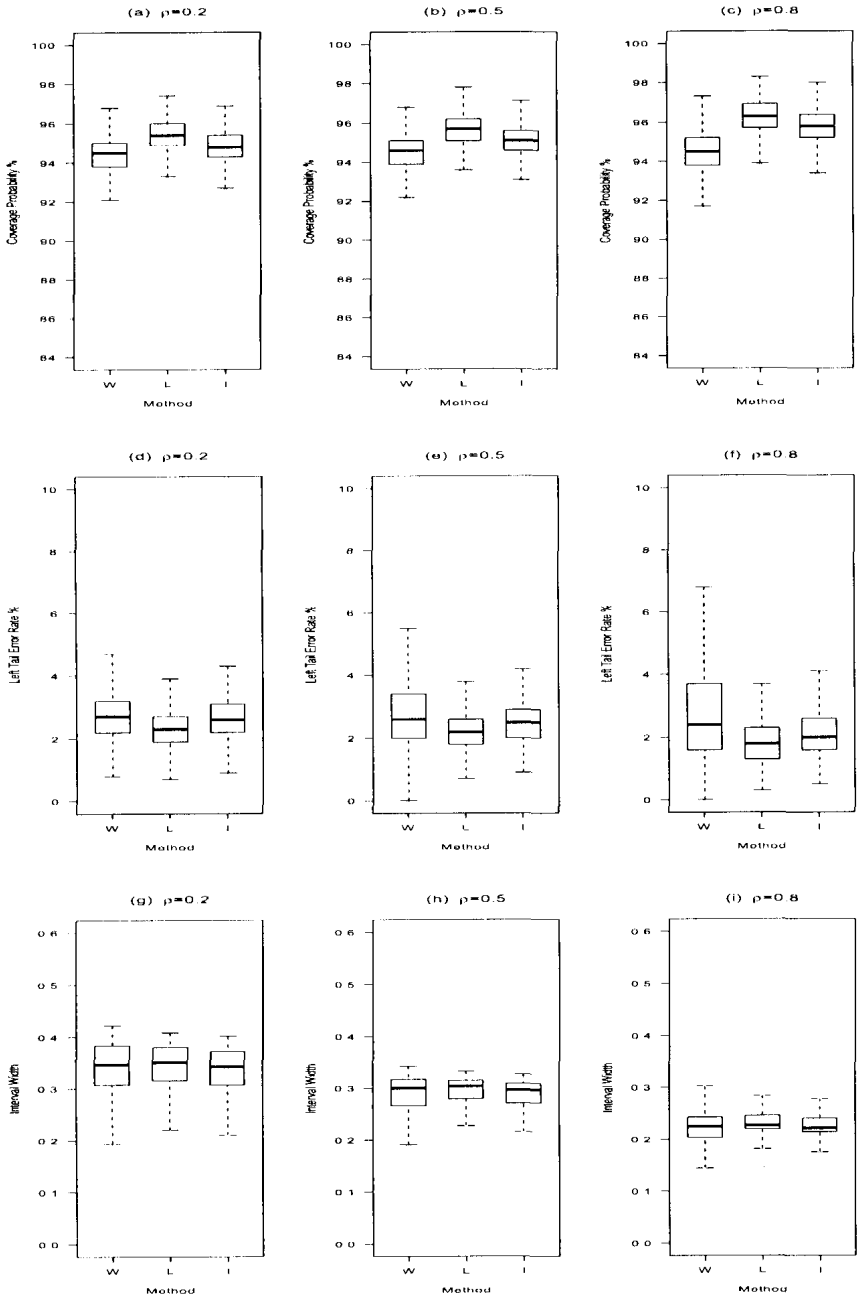

Figure 4.1: Box plots of 95% confidence interval for a difference of areas under 2 ROC curves using three methods for sample sizes $n_X = n_Y = 25$.(a)–(c) are box plots of coverage probability, (d)–(f) are box plots of left tail error rates. (g)–(i) are box plots of interval width. Each box plot was based on 1000 sets of true AUC values. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

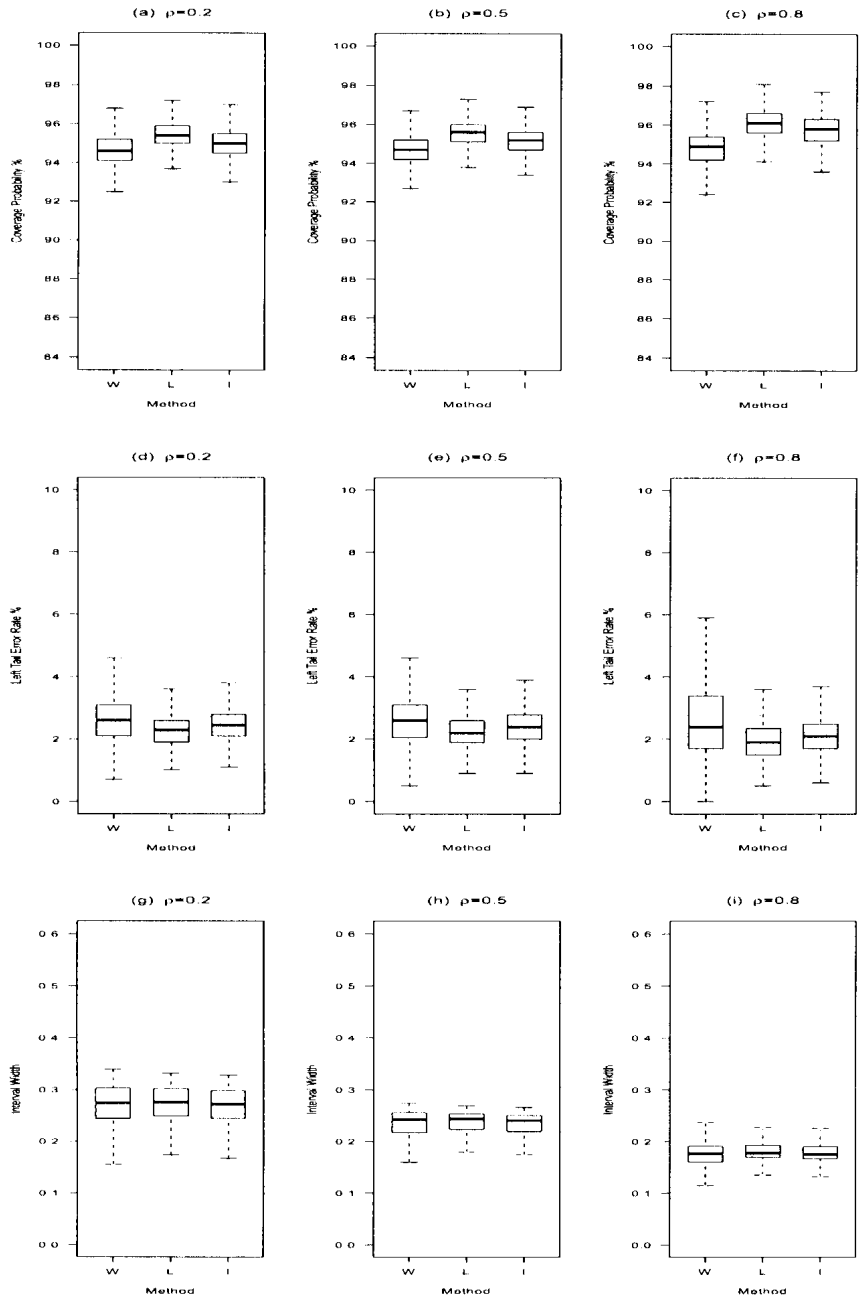

Figure 4.2: Box plots of 95% confidence interval for a difference of areas under 2 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of left tail error rates. (g)–(i) are box plots of interval width. Each box plot was based on 1000 sets of true AUC values, and sample sizes $n_X = 50, n_Y = 25$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

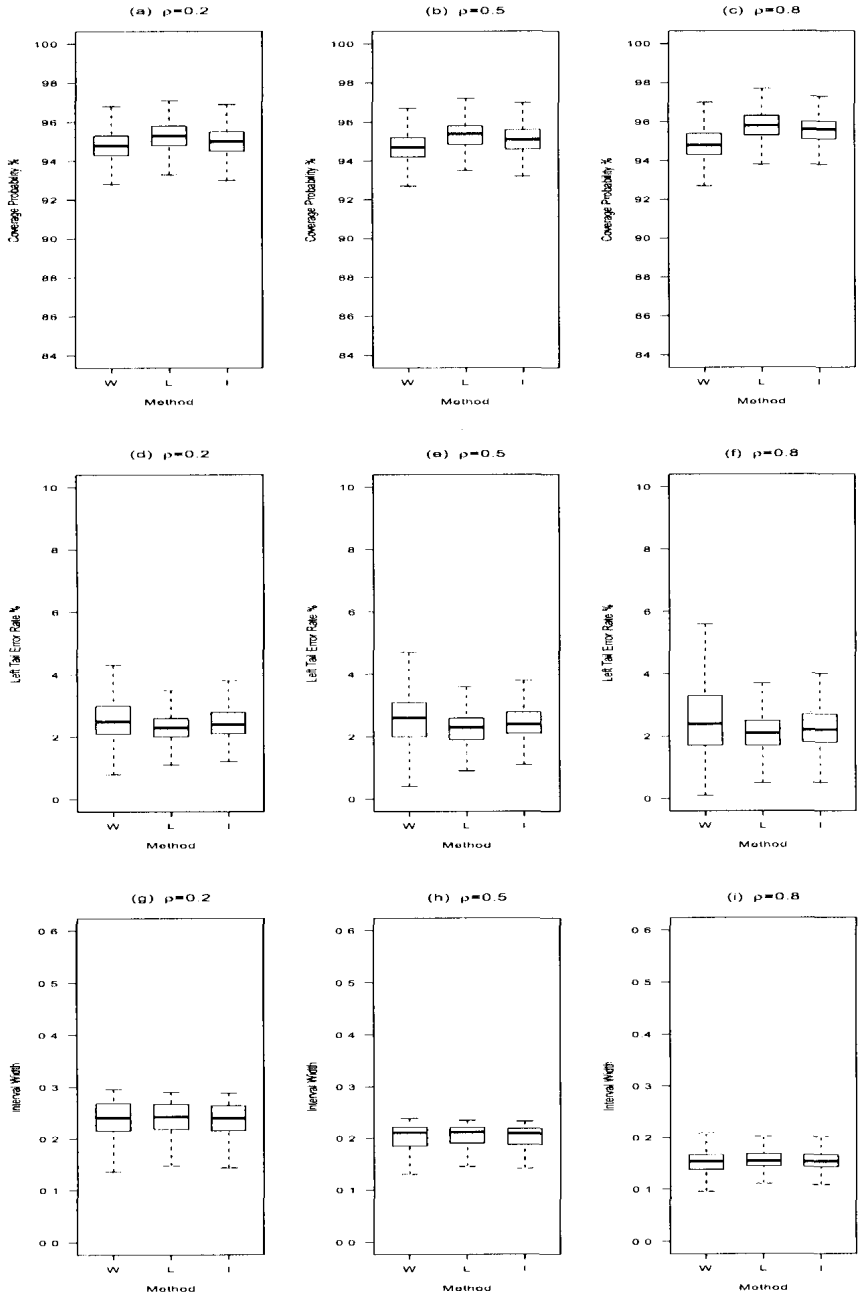

Figure 4.3: Box plots of 95% confidence interval for a difference of areas under 2 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of left tail error rates. (g)–(i) are box plots of interval width. Each box plot was based on 1000 sets of true AUC values, and sample sizes $n_X = n_Y = 50$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

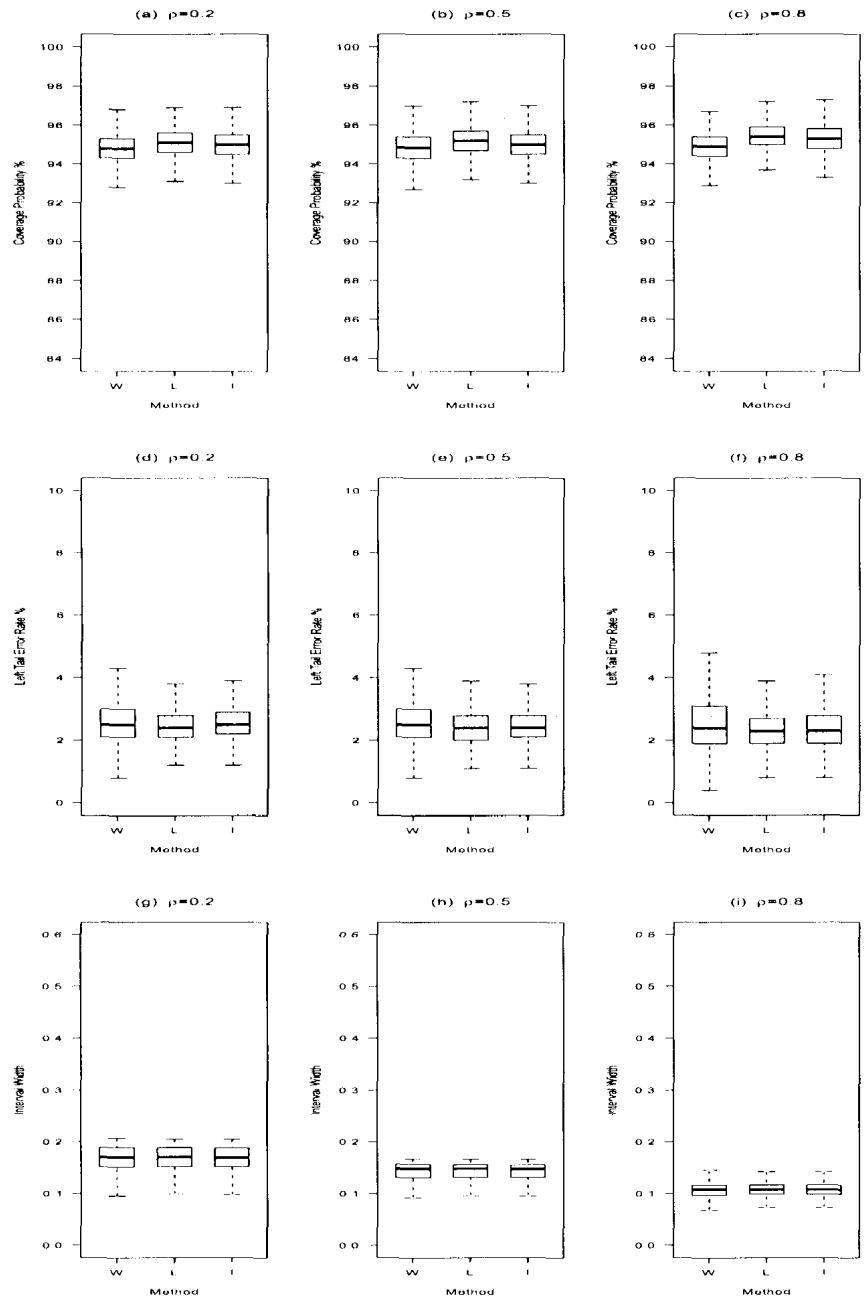

Figure 4.4: Box plots of 95% confidence interval for a difference of areas under 2 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of left tail error rates. (g)–(i) are box plots of interval width. Each box plot was based on 1000 sets of true AUC values, and sample sizes $n_X = n_Y = 100$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

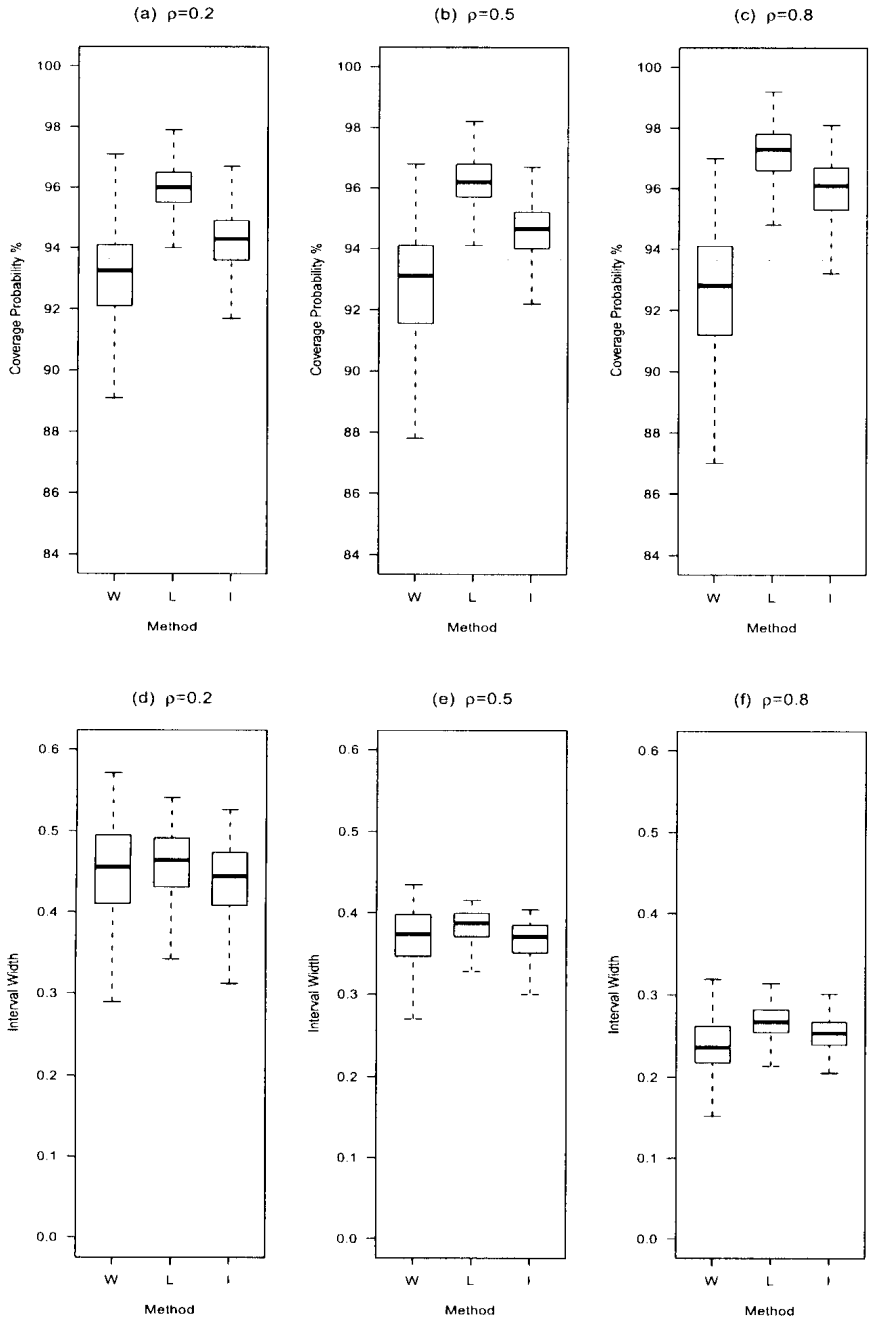

Figure 4.5: Box plots of 95% simultaneous confidence intervals for all pairwise comparisonsof areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 25$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

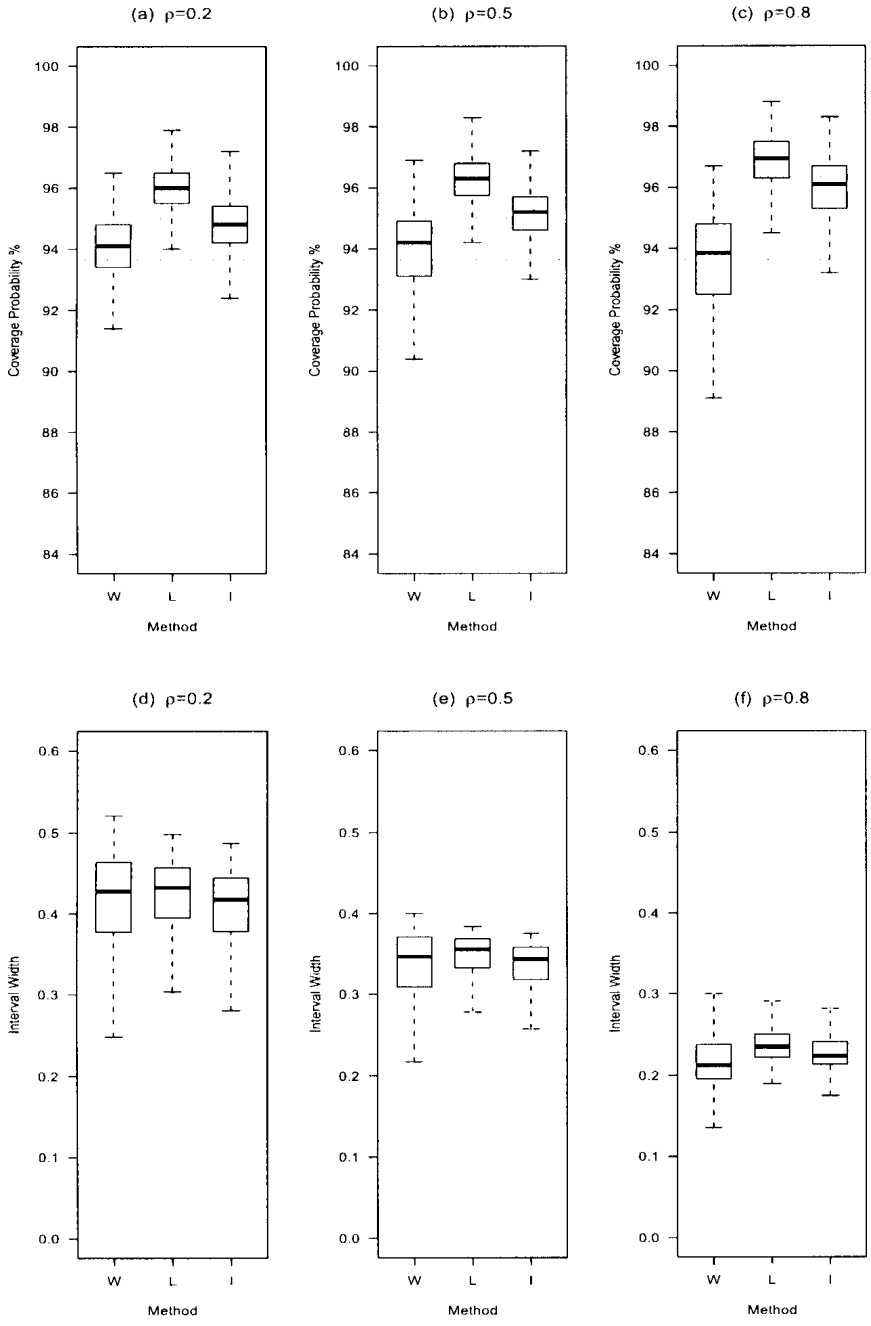

Figure 4.6: Box plots of 95% simultaneous confidence intervals for multiple comparisons with a standard of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 25$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

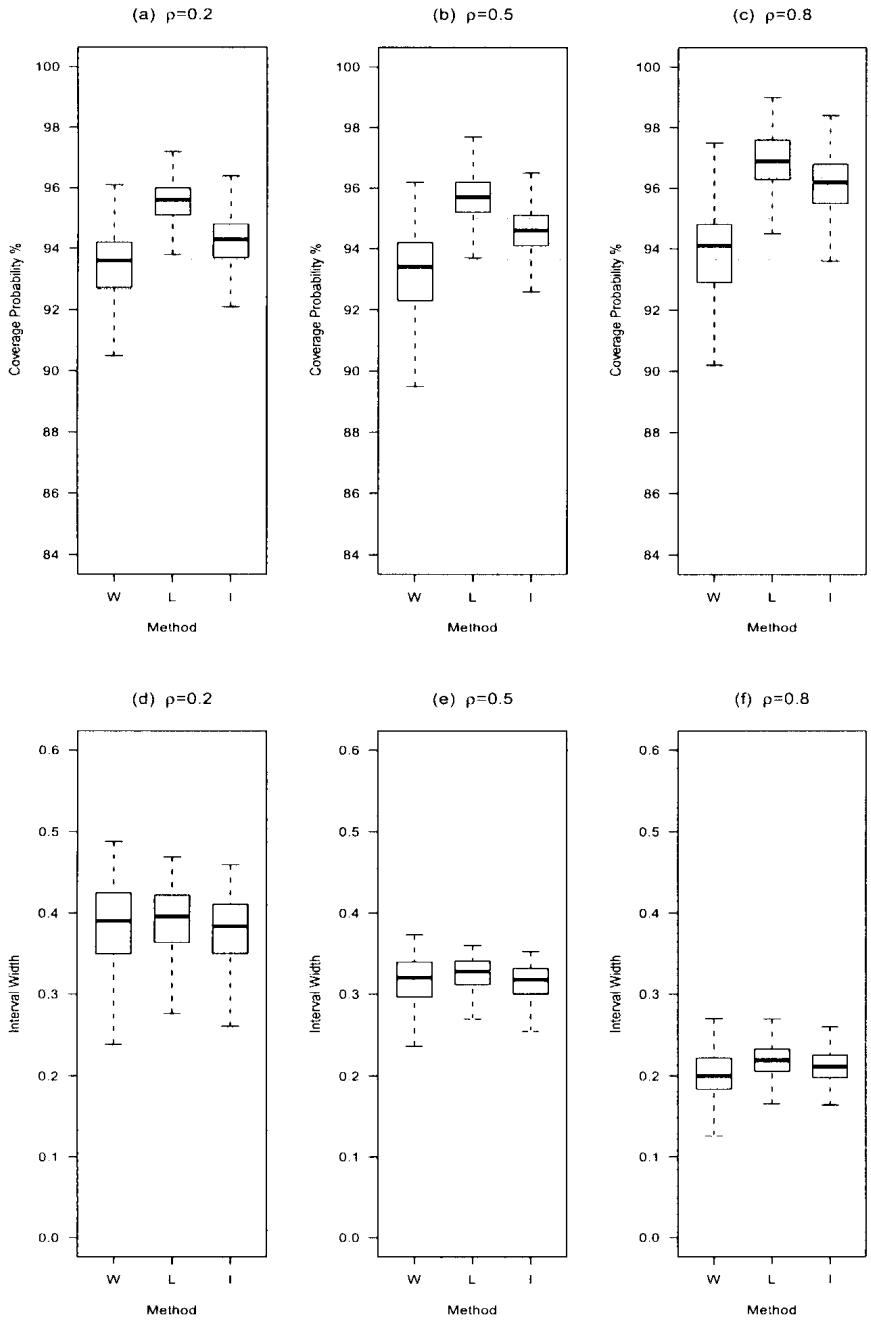

Figure 4.7: Box plots of 95% simultaneous confidence intervals for all pairwise comparisons of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = 50$, $n_Y = 25$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

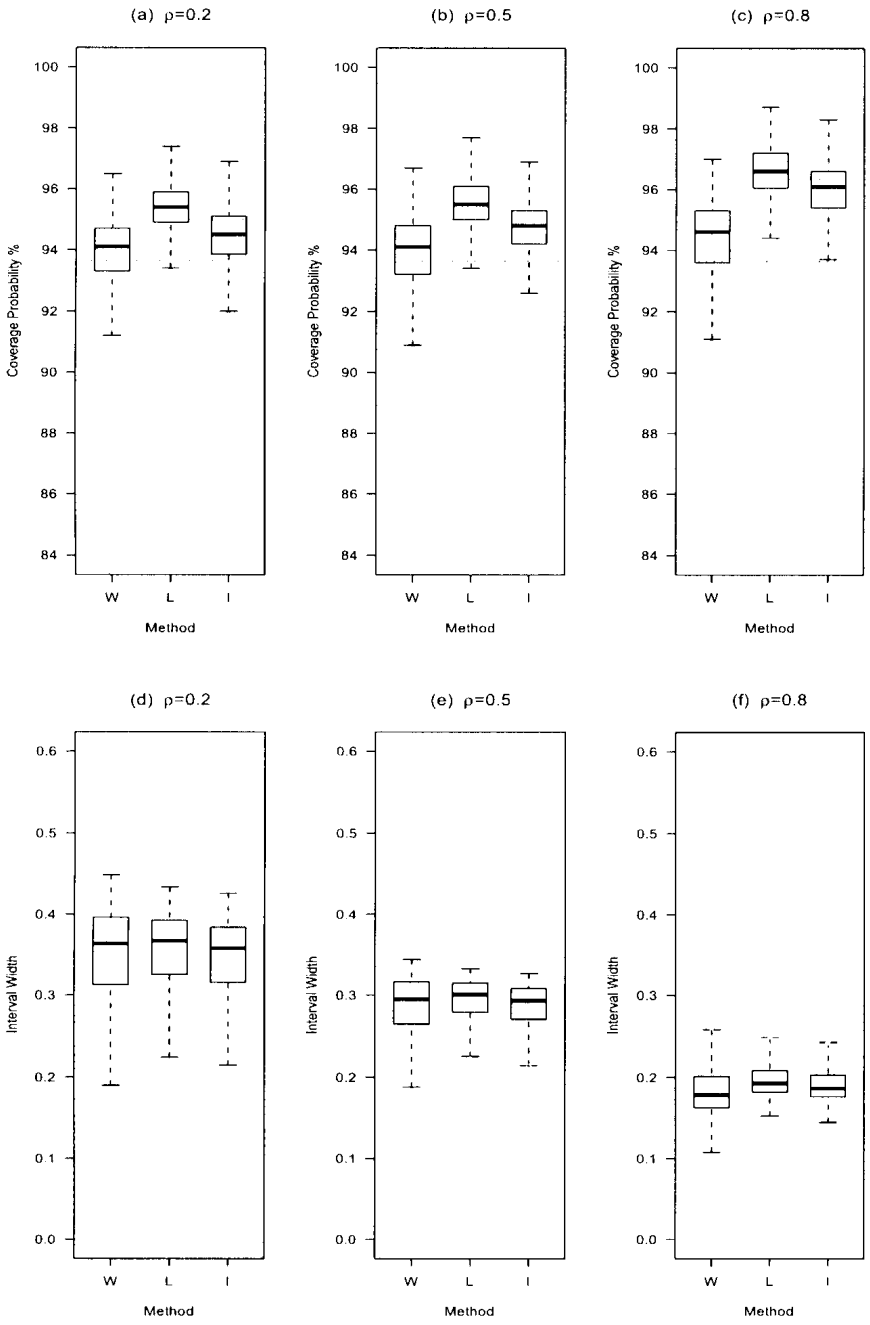

Figure 4.8: Box plots of 95% simultaneous confidence intervals for multiple comparisons with a standard of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = 50$, $n_Y = 25$. r represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

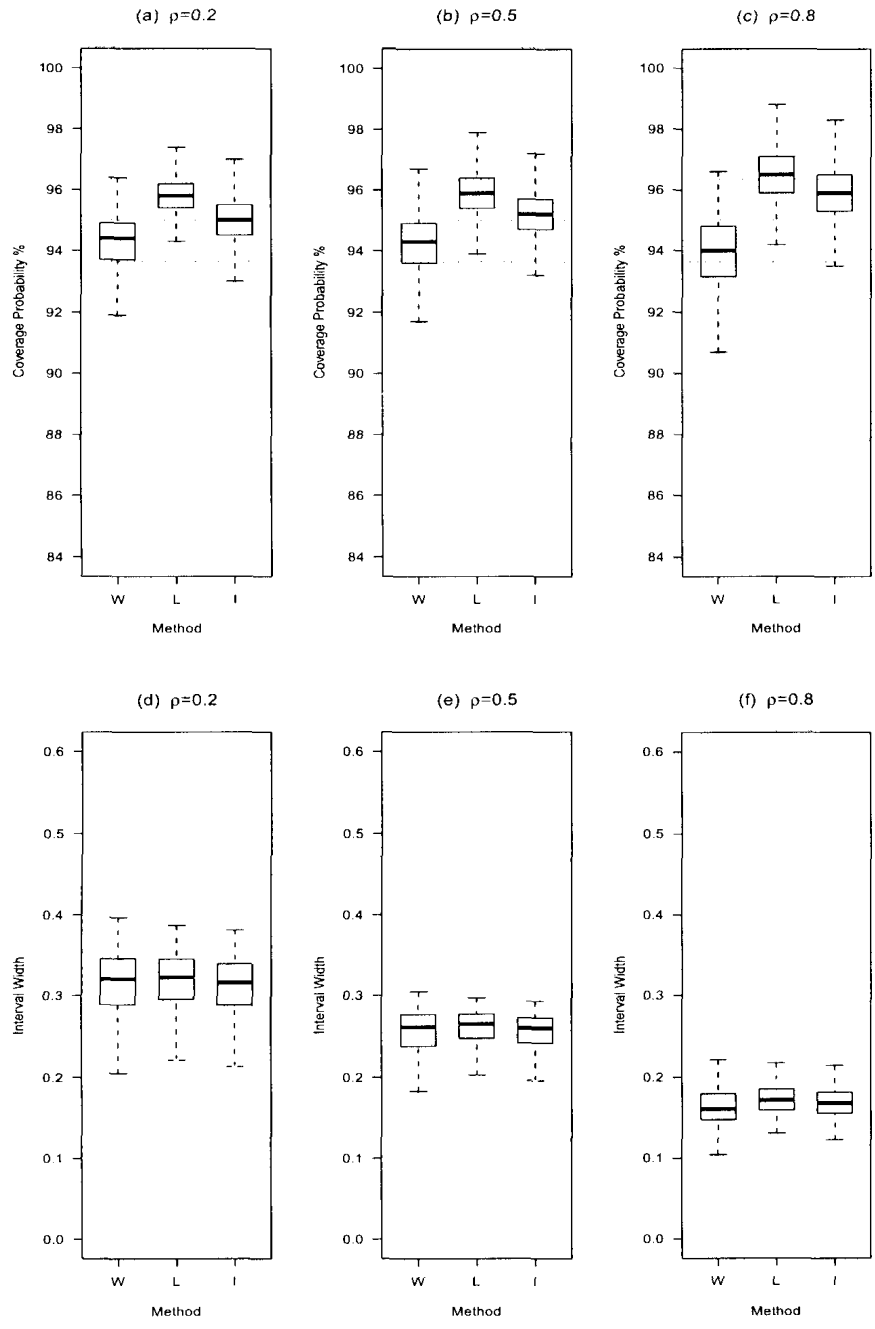

Figure 4.9: Box plots of 95% simultaneous confidence intervals for all pairwise comparisons of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 50$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

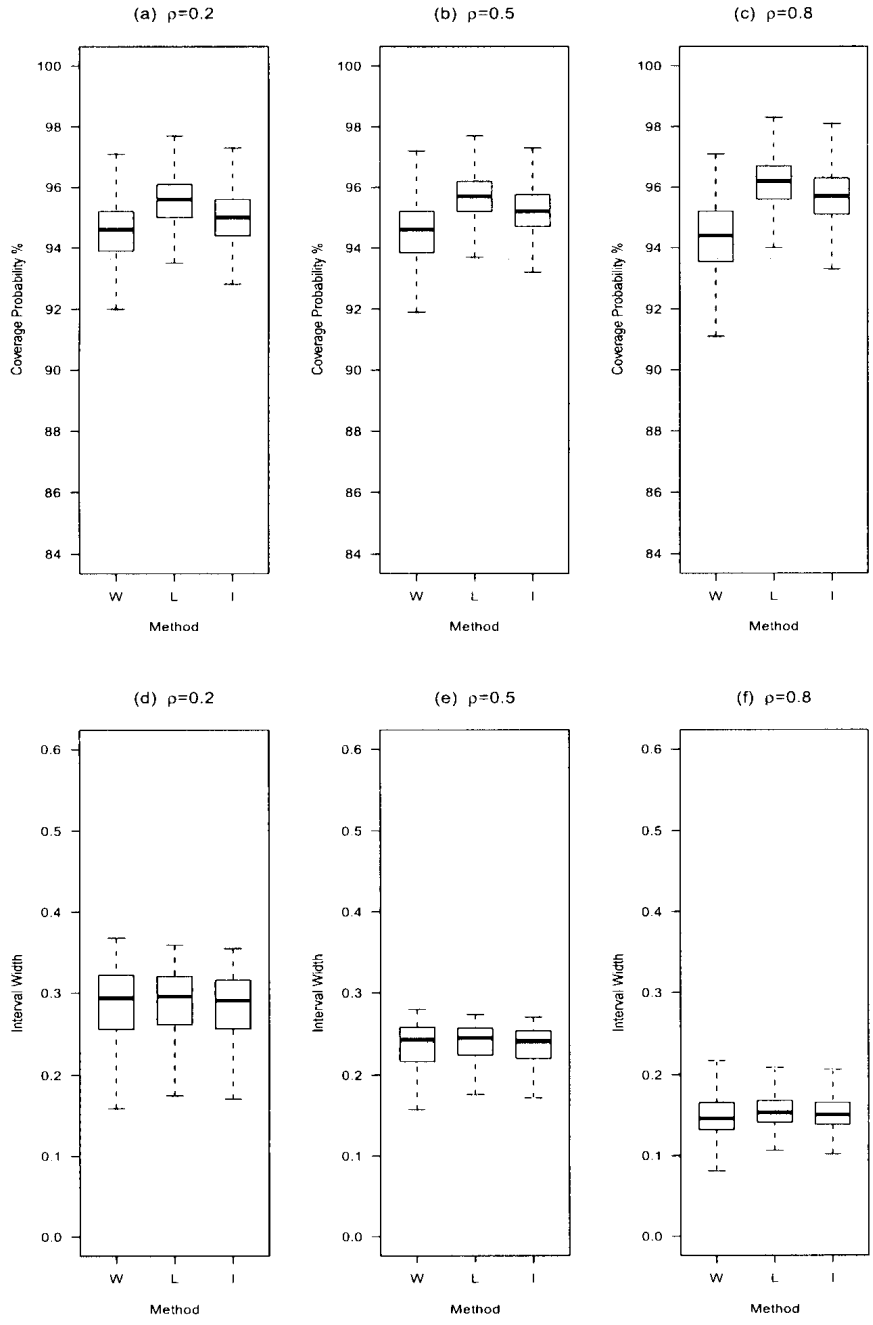

Figure 4.10: Box plots of 95% simultaneous confidence intervals for multiple comparisons with a standard of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 50$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

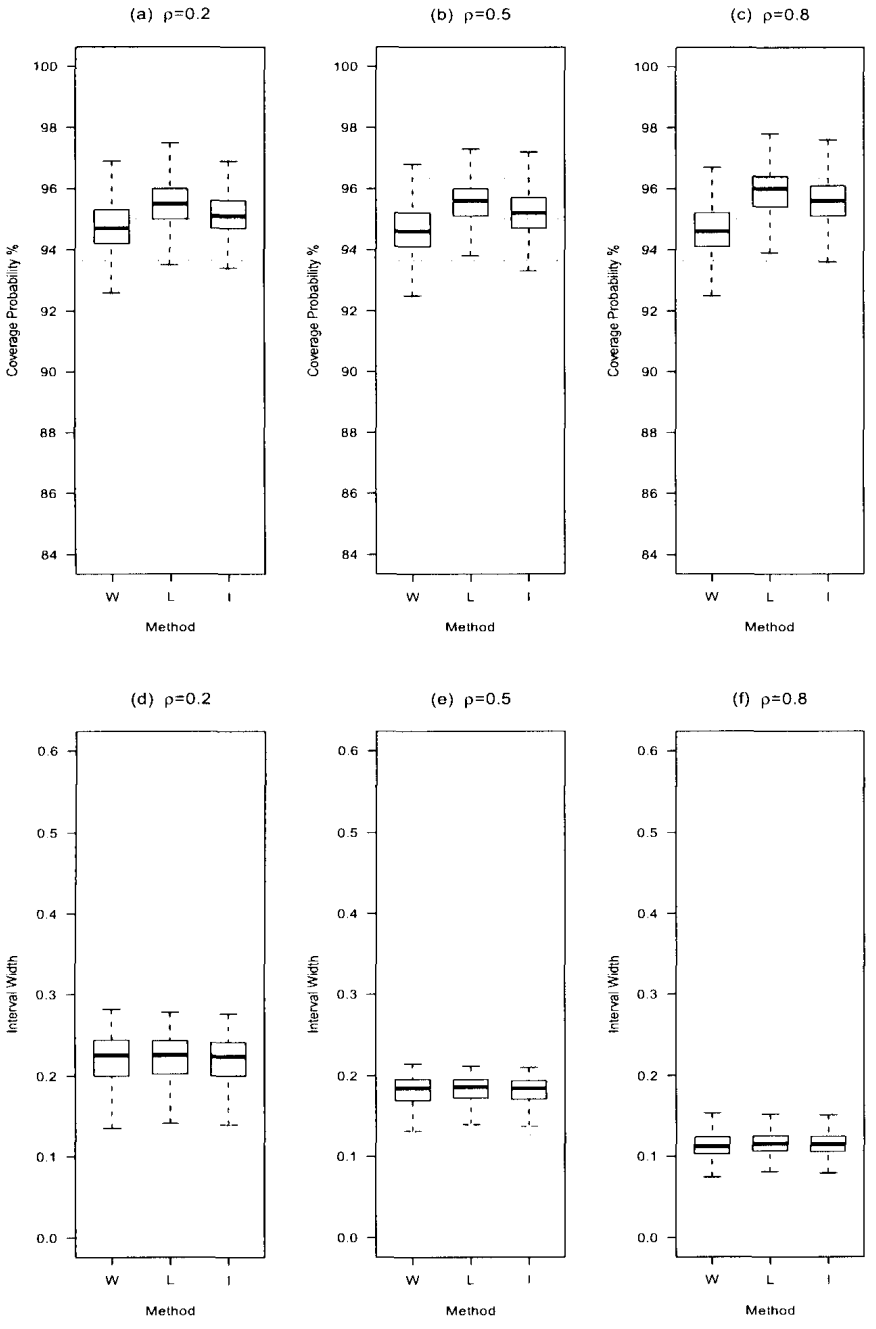

Figure 4.11: Box plots of 95% simultaneous confidence intervals for all pairwise comparisons of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 100$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

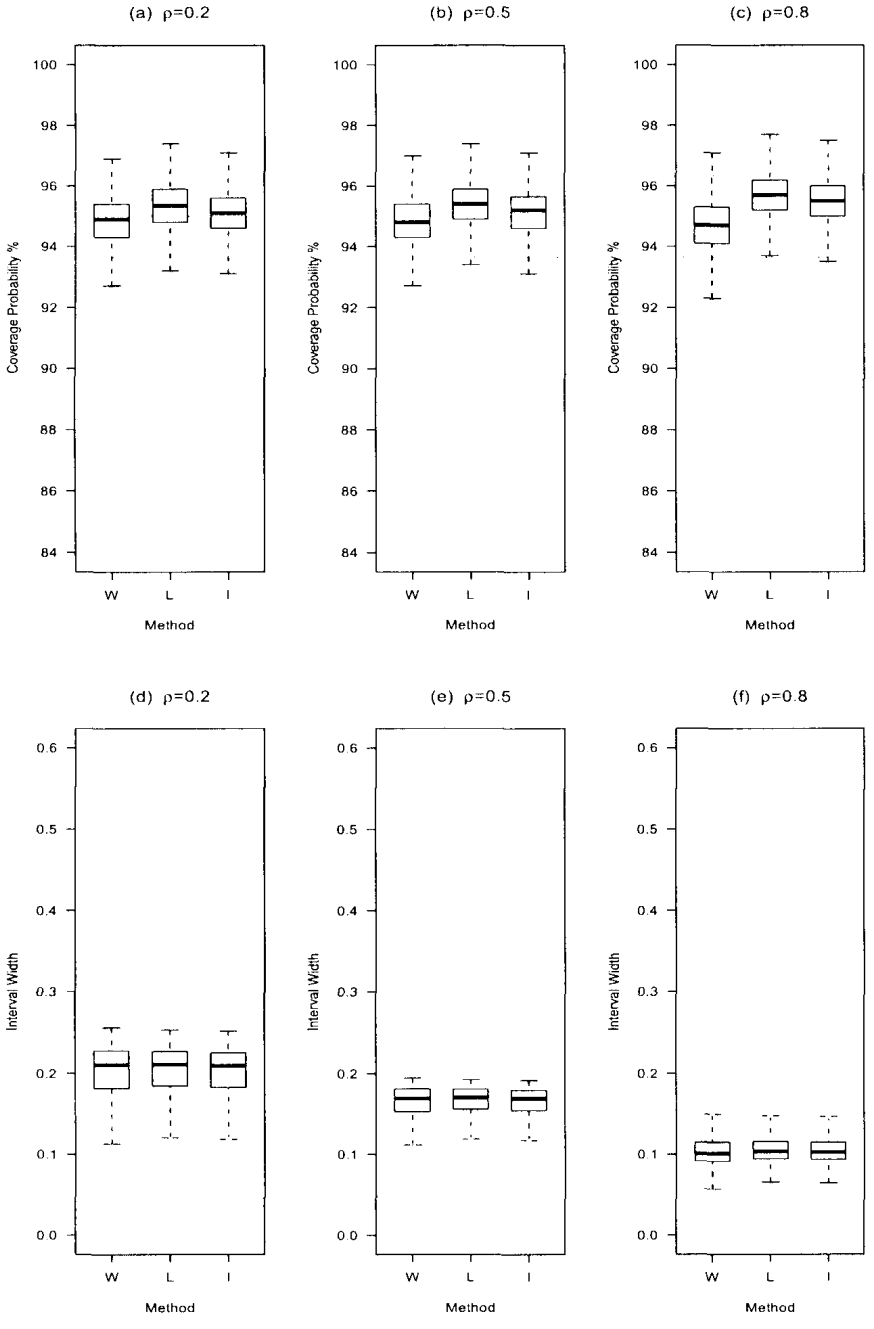

Figure 4.12: Box plots of 95% simultaneous confidence intervals for multiple comparisons with a standard of areas under 4 ROC curves using three methods. (a)–(c) are box plots of coverage probability, (d)–(f) are box plots of interval width. Each box plot was based on 1000 runs, and sample sizes $n_X = n_Y = 100$. ρ represents the correlation coefficient between different tests. Methods 'W', 'L' and 'I' represent 'Wald method', 'Logit method' and 'Inverse Sinh' method, respectively.

Chapter 5

DISCUSSION

This thesis first presented three procedures for constructing confidence interval for a single AUC, and then presented three approaches for confidence interval of a difference between two correlated AUCs, and finally developed simultaneous confidence intervals for multiple comparisons of correlated AUCs derived from the same cases. The simulation results show that for small to moderate correlation coefficients among the test measurements, the method based on the inverse sinh transformation outperforms the other two approaches for the construction of confidence interval for a single AUC in terms of coverage percentage and the balance of tail errors. The advantage of this method carried over to the case of confidence interval for a difference between two AUCs and simultaneous confidence intervals for multiple comparisons of AUCs.

The different performance of the three approaches for a single AUC can be explained by the normality assumption. Since the point estimate and its variance estimate for AUC are related, the sampling distribution for the estimate of $\widehat{\text{AUC}}$ is unlikely to be normal for small to medium sample size. As a result, the simple symmetric Wald-type CI for AUC cannot be a good approach. The method based on logit transformation improved the performance resulting in coverage rate close to the nominal level with balanced tail errors. In addition, inverse sinh transformation of AUC estimates can make the coverage rate even closer to the nominal level with reduced interval width.

For the construction of confidence interval for a difference between two AUCs and also the simultaneous confidence intervals of multiple AUCs, two key steps are needed as follows:

1. Valid confidence interval procedures for a single AUC;

2. A method that can provide confidence interval for a contrast, based on confidence limits for separate AUCs in component.

For the first step, the simulation results in the thesis have shown that the inverse sinh method can provide the most efficient CI for a single AUC among the three approaches especially when sample size for diseased and non-diseased subjects are balanced. Hence, when other conditions are unchanged, this method should also perform best for simultaneous confidence intervals constructions. Our simulation results indicate that the inverse sinh method does have best performance for simultaneous confidence intervals in most cases.

For the second step, before the introduction of MOVER, there was no simple and effective method to construct CI for a linear combination of parameters of interest based only on the confidence limits for each parameter components. (schenker, 2001; Cumming, 2009). As shown in this thesis, the MOVER only requires accurate CIs for each parameter and then we can easily derive a CI for a linear combination of parameters using closed forms. Moreover, MOVER reflects the sampling distributions of the components. In cases when sampling distributions for the parameter estimates are normal, confidence intervals by the MOVER approach are identical to those by the Wald method. The is because the variance estimates recovered using the lower limit and the upper limit are the same if the underlying sampling distribution is normal. Thus, for the simultaneous confidence intervals of multiple comparisons of AUCs based on Wald-type CI of single AUC, MOVER provides the same symmetric simultaneous confidence intervals as calculated by Wald procedure.

We considered three representative correlations $\rho = 0.2$, 0.5 and 0.8. However, in practice, a summary from a large number of studies shows that the average correlation between the observations for two diagnostic tests ranges from 0.35 to 0.59 (Rockette *et al.*, 1999). Hence, the simulation results for correlation coefficient $\rho = 0.5$ is most applicable in practice. Our results show the predominant performance for the method based on inverse sinh transformation when correlation is small to moderate. But we need to notice that, for the rare case that ρ is large, the inverse sinh method may give conservative confidence interval

estimates.

The SAS macro implementing our improved method with available simultaneous confidence intervals of multiple comparisons of AUCs is shown in the appendix.

We have only considered the case where each subject was tested only once by each instrument. Obuchowski (1997) extended the method of DeLong *et al.* (1988) for clustered data for two group comparison. Future work should evaluate the MOVER approach for multiple comparisons for clustered data arising from studies in which multiple observations are obtained by each instrument.

Bibliography

Agresti, A., and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* **54**, 280–288.

Agresti, A., and Coull, B. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery* **29**, 554–556.

Altman, D. G., and Bland, J. M. (1994). Diagnostic test 3: receiver operating characteristic plots. *British Medical Journal* **309**, 188–188.

Aoki, K., Misumi, J., Kimura, T., Zhao, W., and Xie, T. (1997). Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study. *American Journal of Epidemiology* **7**, 143–151.

Bebu, I., and Mathew, T. (2008) Comparing the means and variances of a bivariate log-normal distribution. *Statistics in Medicine* **27**, 2684–2696.

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

Chen, Y. H., and Zhou, X. H. (2006) Interval estimates for the ratio and difference of two lognormal means. *Statistics in Medicine* **25**, 4099–4113.

Cleves, M. A. (1999). Receiving operating characteristic (ROC) analysis. *Stata Technical Bulletin* **52**, 19–33.

Cleves, M. A. (2002). Comparative assessment of three common algorithms for estimating the variance of the area under the nonparmetric reciver operating characteristic curve. *The Stata Journal* **2**, 280–289.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall.

Cumming, G. (2009) Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine* **28**, 205–220.

Daly, L. E. (1998). Confidence limits made easy: interval estimation using a substitution method. *American Journal of Epidemiology* **147**, 783–790.

Deeks, J. J., and Altman, D. G. (2004). Statistics notes: diagnostic tests 4: likelihood ratios. *British Medical Journal* **329**, 168–169.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *Biometrics* **64**, 1270–1275.

Diciccio, T.J., and Efron, B. (1996). Rejoinder of 'bootstrap confidence intervals'. *Statistic Science* **11**, 223–238.

Donner, A., and Zou, G. Y. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics* **58**, 209–215.

Donner, A., and Zou, G. Y. (2010). Closed-form confidence intervals for functions of the normal mean and standard deviation. *Statistical Methods in Medical Research***3**, 320–335.

Donner, A., and Zou, G. Y. (2010). Estimating simultaneous confidence intervals for multiple contrasts of proportions by the method of variance estimates recovery. *Statistics in Biopharmaceutical Research*. (In press)

Dorfman, D. D., and Alf, E. (1969). Maximum-likelihood estmation of parameters of signal-detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology* **6**, 487–496.

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

Dorfman, D. D., Berbaum, K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., and Dagga, H. A. (1997). Proper receiver operating characteristic analysis: The bigamma model. *Academic Radiology* **4**, 138–149.

Edwards, D., and Berry, J. J. (1987). The efficiency of simulation based multiple comparisons. *Biometrics* **43**, 913928.

Efron, B., and Tibshirani, R. (1993). *An Introduction to the Boostrap.* New York : Chapman and Hall.

Erdreich, L. S., and Lee, E. T. (1981). Use of relative operating characteristic analysis in epidemiology. *American Journal of Epidemiology* **114**, 649–662.

Erkanli, A., Sung, M., Costello, E. J., and Angold, A. (2006). Bayesian semiparametric ROC analysis. *Statistics in Medicine* **25**, 3905–3928.

Goddard, M. J., and Hinberg, I. (1990). Receiver operating characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine* **9**, 325–337.

Green, D., and Swets, J. (1966). *Signal Detection Theory and Psychophysics.* New York: John Wiley and Sons, Inc.

Green, D., and Swets, J. (1988). *Signal Detection Theory and Psychophysics.* Wiley: New York, 1966. Reprinted by Peninsula Publishing: Los Altos Hills, CA.

Greenland, S. (1999). Re: "Confidence limits made easy: interval estimation using a substitution method." (Letter). *American Journal of Epidemiology* **149**, 884–884.

Greiner, M., Pfeiffer, D., and Smithc, R. D. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* **45**, 23–41.

Gribskov,M., and Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry* **20**, 25-33.

Grmec, S., and Gasparovic, V. (2001). Comparison of APACHE II, MEES and Glasgow Coma Scale in patients with nontraumatic coma for prediction of mortality. *Critical Care* **5** 19–23.

Gu, J. Z., Ghosal, S., and Roy, A. (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* **27**, 5407–5420.

Hjian-Tilaki, K. O., Hanley, J. A., Joseph, L., and Collet, J. P. (1997). Comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making* **17**, 94–102.

Hand, D. J. (2009). Evaluating diagnositic test: The area under the ROC curve and the balance of errors. *Statistics in Medicine* **29**, 1502–1510.

Hanley, J. A. (1996). The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* **15**, 1575–1585.

Hanley, J. A., and Hajian-Tilaki, J. A. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Statistics in Radiology* **4**, 49–58.

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

Hanley, J. A., and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843.

Hannig, J., Iyer, H., and Patterson, P. (2006). Fiducial generalized confidence intervals. *Journal of American Statistical Association* **101**, 254–269.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *The Annals of Statistics* **12**, 61-75.

He, Y. H., and Escobar, M. (2008). Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. *Statistics in Medicine* **27**, 5291–5308.

Holford, T. R., Walter, S. D., and Dunnett, C. W. (1989). Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *Journal of Clinical Epidemiology* **42**, 427–434.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

Hsieh, F., and Turnbull. B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* **24**, 25–40.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods.* New York : Chapman and Hall.

Hsu, J. C., Qiu, P. H., Hin, L. Y., Mutti, D.O., and Zadnik, K. (2004). Multiple comparisons with the best ROC curve. *Institute of Mathematical Statistics* **47**, 65–75.

Kaufmann, J., Werner, C., and Brunner, E. (2005). Nonparametric methods for analysing the accuracy of diagnostic tests with multiple readers. *Statistical Methods in Medical Research* **14**, 129–146.

Krebs, H. B., and Goplerud, D. R. (1983). Surgical management of bowel obstruction in advanced ovarian carcinoma. *Obstetrics and Gynecology* **61**, 327–330.

Krishnamoorthy, K., Mathew, T., and Ramachandran, G. (2006). Generalized p-values and confidence intervals: A novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene* **3**, 642–650.

Lee, W. C., and Hsiao, C. K. (1996). Alternative summary indices for the receiver operating characteristic curve. *Epidemiology* **7**, 605–611.

Li, G., Tiwari, R. C., and Wells, M. T. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. *Biometrika* **86**, 487–502.

Li, Y., Koval, J., Donner, A., and Zou, G. Y. (2010) Interval estimation for the area under the receiver operating characteristic curve when data are subject to error. *Statistics in Medicine* **29**, 2521-2531.

McClish, D. K. (1987). Comparing the areas under more than two independent ROC curves. *Medical Decision Making* **7**, 149–155.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

McNeil, B. J., and Hanley, J. A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making* **4**, 137–150.

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298.

Metz, C. E., and Kronman, H. B. (1980) Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology* **22**, 218–243.

Metz, C. E., Wang, P. L., and Kronman, H. B. (1984) 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.), 432–445. The Hague: Martinus Nijhof.

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

Miller, R. G. (1981). *Simultaneous Statistical Inference 2nd Ed.* Springer Verlag, New York.

Moher, D., Schulz, K. F., and Altman, D. G.(1998). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet* **357**, 1191–1194.

Nelson, P. R. (1989). Multiple comparisons of means using simultaneous confidence intervals. *Jounel of Quality Technology* **21**, 232–241.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**, 857–872.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **17**, 2635–2650.

Newcombe, R. G. (1999). Re: "Confidence limits made easy: interval estimation using a substitution method." (Letter). *American Journal of Epidemiology* **149**, 884-885.

Newcombe, R. G. (2001). Logit confidence interval and the inverse sinh transformation. *The American Statistician* **55**, 200–202.

Neyman, J. (1935). On the problem of confidence limits. *Annals of Mathematical Statistics* **6**, 111–116.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of Royal Society of London. Series A, Mathematical and Physical Sciences* **236**, 333–380.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* New York: Oxford Universtiy Press.

Pepe, M. S., and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**, 528-535.

Peterson, W., Birdsall, T., and Fox, W. (1954). The theory of signal detectability. *Institute of Radio Engineers Transactions* **4**, 171–212.

Piegorsch, W. W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics* **47**, 45–52.

Rao, J. N. K., and ScottSaigo, A. J. (1992). A Simple Method for the Analysis of Clustered Binary Data. *Biometrics* **48**, 577–585.

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

Saigo, H., Vert, J. P., Ueda, N., and Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics* **20**,1682-1689.

Schaarschmidt, F., Biesheuvel, E., and Hothorn, L. A. (2009). Asymptotic simultaneous confidence intervals for many-to-one comparisons of binary proportions in randomized clinical trials. *Journal of Biopharmaceutical Statistics* **19**, 292–310.

Schenker, N., and Gentleman, J. F. (2001) On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician* **55**, 182–186.

Sen, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* **10**, 1–18.

Shapiro, D. E. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research* **8** 113–134.

Spackman, K. A. (1989). *"Signal detection theory: valuable tools for evaluating inductive learning". Proceedings of the Sixth International Workshop on Machine Learning.* San Mateo, CA: Morgan Kaufmann.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Tian, L. L. (2005). Inferences on the mean of zero-inflated lognormal data: The generalized variable approach. *Statistics in Medicine* **24**, 3223–3232.

Tian, L. L., and Cappelleri, J. C. (2004). A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient:the generalized variable method. *Statistics in Medicine* **23**, 2125–2135.

Tsimikas, J. V., Bosch, R, J., Coull, B. A., Coull, B. A., and Barmi, H. E. (2002). Profile-likelihood inference for highly accurate diagnostic tests. *Biometrics* **58**, 946–956.

Venzon, D. J., and Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics* **37**, 87–94.

Vergara, I. A., Norambuena, T., Ferrada, E., Slater, A. W., and Melo, F. (2008). StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics* **9**, 265–269.

Wan, S., and Zhang, B. (2007). Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* **26**, 2565–2586.

Westfall, P., Tobias, R. D., and Rom, D. (1999). *Multiple Comparisons and Multiple Tests Using SAS.* Cary, NC: SAS Institute.

Weerahandi, S. (1991) Testing variance components in mixed models with generalized p values. *Journal of the American Statistical Association* **86**, 151–153.

Weerahandi, S. (1993) Generalized confidence intervals. *Journal of the American Statistical Association* **88**, 899–905.

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

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

Wolfe, R., and Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *Canadian Medical Association Journal* **166**, 65–66.

Zhou, X. H., and Lin, H. Z. (2008). Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests. *Statistics in Medicine* **27**, 5271-5290.

Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology* **168**, 212-224.

Zou, G. Y. (2009). Assessment of risks by predicting counterfactuals. *Statistics in Medicine* **28**, 3761-3781.

Zou, G. Y. (2010). Confidence interval estimation under inverse sampling. *Computational Statistics & Data Analysis* **54**, 55–64.

Zou, G. Y., and Donner, A. (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* **27**, 1693–1702.

Zou, G. Y., Huang, W. Y., and Zhang, X. H. (2009). A note on confidence interval estimation for a linear function of binomial proportions. *Computational Statistics & Data Analysis* **53**, 1080–1085.

Zou, G. Y., Taleban. J., and Huo, C. H. (2009). Confidence interval estimation for lognormal data with application to health economics. *Computational Statistics & Data Analysis* **53**, 3755–3764.

Zou, K. H., and Hall, W. J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* **27**, 621–631.

Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**, 2143–2156.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., and Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology* **11**, 178-189.

Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561–577.

Appendix

```
%macro roc(data=, var=, response=, contrast=, details=no,
           alpha=.05);


%let opts = %sysfunc(getoption(notes))
            _last_=%sysfunc(getoption(_last_));
options nonotes;
%let error=0;

/* Verify that DATA= option is specified */
%if &data= %then %do;
 %put ERROR: Specify DATA= containing the OUT= data sets of models
             to be compared;
 %goto exit;
%end;

/* Verify that VAR= option is specified */
%if &var= %then %do;
  %put ERROR: Specify predictor or XBETA variables in the VAR= argument;
  %goto exit;
%end;

/* Verify that RESPONSE= option is specified */
%if &response= %then %do;
    %put ERROR: Specify response variable in the RESPONSE= argument;
    %goto exit;
%end;

%let i=1;
%do %while (%scan(&data,&i) ne %str() );
  %let data&i=%scan(&data,&i);
  %let i=%eval(&i+1);
%end;
%let ndata=%eval(&i-1);

data _comp(keep = &var &response);
 %if &data=%str() or &ndata=1 %then set;
 %else merge;
  &data;
  if &response not in (0,1) then call symput('error',1);
```

```
  run;
%if &error=1 %then %do;
  %put ERROR: Response must have values 0 or 1 only.;
  %goto exit;
%end;



/* Original SAS/IML code from author follows */
proc iml;
   start mwcomp(psi,z);
    *;
    * program to compute the mann-whitney components  ;
    * z is (nn by 2);
    *  z[,1] is the column of data values;
    *  z[,2] is the column of indicator variables;
    *   z[i,2]=1 if the observation is from the x population;
    *   z[i,2]=0 if the observation is from the y population;
    *
    * psi is the returned vector of u-statistic components;

    rz  = ranktie( z[,1] );                          * average ranks;
    nx  = sum( z[,2] );                              * num. of Xs  ;
    ny  = nrow(z)-nx;                                * num of Ys   ;
    loc = loc( z[,2]=1 );                            * x indexes   ;
    psi = j(nrow(z),1,0);
    psi[loc] = (rz[loc] - ranktie(z[loc,1]))/ny;     * x components ;
    loc = loc( z[,2]=0 );                            * y indexes    ;
    psi[loc] = ( nx+ranktie(z[loc,1])-rz[loc])/nx; * y components ;
    free rz loc nx ny;                               * free space   ;
   finish;

   start mwvar(t,v,nx,ny,z);
    *;
    * compute mann-whitney statistics and variance;
    * input z, n by (k+1);
    *  z[,1:k] are the different variables;
    *  z[,k+1] are indicator values,
    *    1 if the observation is from population x and ;
    *    0 if the observation is from population y;
    * t is the k by k vector of estimated statistics;
    *  the (i,j) entry is the MannWhitney statistic for the
```

```
     *  i-th column when used with the j-th column. The only
     *  observations with nonmissing values in each column are
     *  used. The diagonal elements are, hence, based only on the
     *  single column of values.
     * v is the k by k estimated variance matrix;
     * nx is the matrix of x population counts on a pairwise basis;
     * ny is the matrix of y population counts on a pairwise basis;

     k   = ncol(z)-1;
     ind = z[,k+1];
     v   = j(k,k,0); t=v; nx=v; ny=v;

     * The following computes components after pairwise deletion of
     *  observations with missing values. If either there are no missing
     *  values or it is desired to use the components without doing
     *  pairwise deletion first, the nested do loops could be evaded.
     *;
     do i=1 to k;
       do j=1 to i;
          who = loc( (z[,i]^=.)#(z[,j]^=.) );     * nonmissing pairs;
          run mwcomp(psii,(z[,i]||ind)[who,]);    * components;
          run mwcomp(psij,(z[,j]||ind)[who,]);
          inow = ind[who,];                       * Xs and Ys;
          m = inow[+];                            * current Xs;
          n = nrow(psii)-m;                       * current Ys;
          nx[i,j] = m; ny[i,j] = n;
          mi = (psii#inow)[+] / m;                * means;
          mj = (psij#inow)[+] / m;
          t[i,j] = mi; t[j,i] = mj;
          psii = psii-mi; psij = psij-mj;         * center;
          v[i,j] = (psii#psij#inow)[+]      / (m#(m-1))
                 + (psii#psij#(1-inow))[+] / (n#(n-1));
          v[j,i] = v[i,j];
       end;
     end;
     free psii psij inow ind who;
   finish;
/* start of execution of the IML program */
   use _comp var {&var &response};
   read all into data [colname=names];
   run mwvar(t,v,nx,ny,data);                  * estimates and variances;
   vname = names[1:(ncol(names)-1)];
```

```
   manwhit = vecdiag(t);


 %if &contrast= %then %do;
  %put ROC: No contrast specified.  Pairwise contrasts of all;
  %put %str(    ) curves will be generated.;
   call symput('col',char(ncol(data)-1));
  %if &col=1 %then %str(l=1;); %else %do;
   l=(j(&col-1,1)||-i(&col-1))
    %do i=&col-2 %to 1 %by -1;
        //(j(&i,&col-&i-1,0)||j(&i,1)||-i(&i))
    %end;
   ;
   %end;
   call symput('maxrow',char(comb(max(nrow(l),2),2)));
 %end;
 %else %do;
   l = { &contrast };
   call symput('maxrow',char(nrow(l)));
 %end;

   lt=l*manwhit;
   lv=l*v*l';
   c = ginv(lv);
   chisq = lt'*c*lt;
   df = trace(c*lv);
   p = 1 - probchi( chisq, df );
/* Original SAS/IML code by author ends */


   /* Individual area stderrs and CIs */
   stderr=sqrt(vecdiag(v));
   arealcl=manwhit-probit(1-&alpha/2)*stderr;
   areaucl=manwhit+probit(1-&alpha/2)*stderr;
   areastab=putn(manwhit||stderr||arealcl||areaucl,'7.4');

   /* Pairwise difference stderrs and CIs */
   sediff=sqrt(vecdiag(lv));
   difflcl=lt-probit(1-&alpha/2)*sediff;
   diffucl=lt+probit(1-&alpha/2)*sediff;
   diffchi=(lt##2)/vecdiag(lv);
   diffp=1-probchi(diffchi,1);
```

```
  %if %upcase(%substr(&details,1,1)) ne N %then %do;
   print t [label='Pairwise Deletion Mann-Whitney Statistics'
   colname=vname  rowname=vname];
  %end;

   print areastab [label=
   "ROC Curve Areas and
   %sysevalf(100*(1-&alpha))% Confidence Intervals"
   rowname=vname colname={'ROC Area' 'Std Error'
   'Confidence' 'Limits'}];

   rname='Row1':"Row&maxrow";
%if %upcase(%substr(&details,1,1)) ne N %then %do;
   print v [label='Estimated Variance Matrix'
    colname=vname rowname=vname];
   print nx [label='X populations sample sizes'
    colname=vname rowname=vname];
   print ny [label='Y populations sample sizes'
    colname=vname rowname=vname];
   print lv [label='Variance Estimates of Contrast'
    rowname=rname colname=rname];
  %end;
   print l [label='Contrast Coefficients'
    rowname=rname colname=vname];

   fdiffchi=putn(diffchi,'9.4');
   fdiffp=putn(diffp,'pvalue.');
   diffs=putn(lt||sediff||difflcl||diffucl,'7.4');
   diffstab=diffs||fdiffchi||fdiffp;
   print diffstab [label=
 "Tests and %sysevalf(100*(1-&alpha))
   % Confidence Intervals for Contrast Rows"
   rowname=rname colname={'Estimate'
     'Std Error' 'Confidence' 'Limits'
   'Chi-square' 'Pr > ChiSq'}];

   c2=putn(chisq,'9.4');
   df2=putn(df,'3.');
   p2=putn(p,'pvalue.');
   ctest=c2||df2||p2;
   print ctest [label='Contrast Test Results'
```

```
        colname={'Chi-Square' '  DF' 'Pr > ChiSq'}];

   /* Make overall p-value available */
   %global pvalue;
   call symput('pvalue',p2);

quit;

%exit:
options &opts;
title; title2;
%mend;

 data roc;
     input alb tp totscore popind;
     totscore = 10 - totscore;
     datalines;
   3.0 5.8 10 0
   3.2 6.3  5 1
   3.9 6.8  3 1
   2.8 4.8  6 0
   3.2 5.8  3 1
   0.9 4.0  5 0
   2.5 5.7  8 0
   1.6 5.6  5 1
   3.8 5.7  5 1
   3.7 6.7  6 1
    .   .   6 1
   3.2 5.4  4 1
   3.8 6.6  6 1
   4.1 6.6  5 1
   3.6 5.7  5 1
   4.3 7.0  4 1
   3.6 6.7  4 0
   2.3 4.4  6 1
   4.2 7.6  4 0
   4.0 6.6  6 0
   3.5 5.8  6 1
   3.8 6.8  7 1
   3.0 4.7  8 0
   4.5 7.4  5 1
   3.7 7.4  5 1
```

```
   3.1 6.6  6 1
   4.1 8.2  6 1
   4.3 7.0  5 1
   4.3 6.5  4 1
   3.2 5.1  5 1
   2.6 4.7  6 1
   3.3 6.8  6 0
   1.7 4.0  7 0
    .   .   6 1
   3.7 6.1  5 1
   3.3 6.3  7 1
   4.2 7.7  6 1
   3.5 6.2  5 1
   2.9 5.7  9 0
   2.1 4.8  7 1
    .   .   8 1
   2.8 6.2  8 0
    .   .   7 1
    .   .   7 1
   4.0 7.0  7 1
   3.3 5.7  6 1
   3.7 6.9  5 1
   2.0  .   7 1
   3.6 6.6  5 1
   ;
data roc;
      input alb tp totscore popind @@;
      totscore = 10 - totscore;
      datalines;
   3.0 5.8 10 0   3.2 6.3  5 1   3.9 6.8  3 1   2.8 4.8  6 0
   3.2 5.8  3 1   0.9 4.0  5 0   2.5 5.7  8 0   1.6 5.6  5 1
   3.8 5.7  5 1   3.7 6.7  6 1   3.2 5.4  4 1   3.8 6.6  6 1
   4.1 6.6  5 1   3.6 5.7  5 1   4.3 7.0  4 1   3.6 6.7  4 0
   2.3 4.4  6 1   4.2 7.6  4 0   4.0 6.6  6 0   3.5 5.8  6 1
   3.8 6.8  7 1   3.0 4.7  8 0   4.5 7.4  5 1   3.7 7.4  5 1
   3.1 6.6  6 1   4.1 8.2  6 1   4.3 7.0  5 1   4.3 6.5  4 1
   3.2 5.1  5 1   2.6 4.7  6 1   3.3 6.8  6 0   1.7 4.0  7 0
   3.7 6.1  5 1   3.3 6.3  7 1   4.2 7.7  6 1   3.5 6.2  5 1
   2.9 5.7  9 0   2.1 4.8  7 1   2.8 6.2  8 0   4.0 7.0  7 1
   3.3 5.7  6 1   3.7 6.9  5 1   3.6 6.6  5 1
   ;
```

```
 data alb(keep=alb popind)
      ptp(keep=tp popind)
      ptots(keep=totscore popind);
    set roc;
run;

     %roc(data=alb ptp ptots,
          var=alb tp totscore,
          response=popind)
```