Published in final edited form as:

J Stat Plan Inference. 2009 ; 139(10): 3725–3732. doi:10.1016/j.jspi.2009.05.001.

Confidence interval estimation of the difference between paired AUCs based on combined biomarkers

Lili Tian^{a,*}, Albert Vexler^a, Li Yan^a, and Enrique F. Schisterman^b

^a Department of Biostatistics, University at Buffalo, 249 Farber Hall, 3435 Main St. Bldg. 26 Buffalo, NY 14214–3000, USA

^b Division of Epidemiology, Statistics and Prevention, NICHD, NIH, USA

Abstract

In many diagnostic studies, multiple diagnostic tests are performed on each subject or multiple disease markers are available. Commonly, the information should be combined to improve the diagnostic accuracy. We consider the problem of comparing the discriminatory abilities between two groups of biomarkers. Specifically, this article focuses on confidence interval estimation of the difference between paired AUCs based on optimally combined markers under the assumption of multivariate normality. Simulation studies demonstrate that the proposed generalized variable approach provides confidence intervals with satisfying coverage probabilities at finite sample sizes. The proposed method can also easily provide *P*-values for hypothesis testing. Application to analysis of a subset of data from a study on coronary heart disease illustrates the utility of the method in practice.

Keywords

Receiver operating characteristic (ROC) curve; Optimal linear combinations; Generalized pivot; Generalized test variable

1. Introduction

Receive-operating characteristic (ROC) analysis is a useful tool for evaluating the performance of diagnostic tests. The ROC curve of a biomarker is a plot of its sensitivity (true positive rate) vs. 1 minus specificity (true negative rate). The area under ROC curve (AUC) is an overall summary of diagnostic accuracy and has been widely used as a quantitative index of the performance of a biomarker. For excellent sources of information on ROC analysis, see e.g., Shapiro (1999), Zhou et al. (2002) and Pepe (2003). When multiple diagnostic tests results or multiple disease markers are available, they should be combined to improve the diagnostic accuracy. Based on multivariate normal theory, Su and Liu (1993) proposed an optimal linear combination of markers to obtain the largest AUC among all the possible linear combinations of the markers. Following the work of Su and Liu (1993), the related issues in the context of AUC have been well addressed under the assumption of multivariate normality. For example, Reiser and Faraggi (1997) provided confidence intervals for the AUC of the optimal linear combination of markers; Liu et al. (2005) considered combinations maximizing sensitivity over a range of specificity and demonstrated that the optimal combination uniformly dominate all the other linear combinations under homogeneity. There also exist some works in distribution-

free context, e.g., Pepe and Thompson (2000) discussed non-parametric approach for optimizing the AUC and generalized this approach to a few aspects; Vexler et al. (2006) proposed a kernel-smoothed approach to estimate the optimal linear combination and the corresponding separation measure.

The comparison of the overall diagnostic accuracy of two markers simultaneously measured on an individual is frequently addressed by comparing the resulting paired AUCs. For example, Obuchowski (1997) and Zhou et al. (2002) applied two one-sided tests to evaluate the two-sided equivalence of two diagnostic procedures based on the non-parametric estimates of paired AUCs developed by Delong et al. (1988); Wieand et al. (1989) proposed non-parametric and parametric tests for the same problem; Molodianovitch et al. (2006) extended Wieand et al.'s approach to non-normal data by using Box-Cox transformation; recently, Li et al. (2008) has applied generalized inference theory for confidence interval estimation of the difference between paired AUCs under the assumption of normality. For more related research in this context, see Zhang et al. (2002) and Vexler et al. (2008).

Under many circumstances, it is of interest to test if one group of biomarkers has the same discriminatory ability as another group, that is, to compare paired AUCs based on combined biomarkers. For example, in a study on coronary heart disease (CHD) Schisterman et al. (2002), evaluating the discriminatory ability of a new biomarker like TBARS (a measured of oxidative stress) and comparing it with that of several well-established biomarkers like uric acid, Body Mass Index (BMI), globulin and total protein have been shown to be clinically important for the early identification and prevention of coronary heart disease. Despite the fact that many statistical inference procedures have been developed for comparisons of diagnostic accuracy based on the paired AUCs, to our knowledge, there does not exist an inference procedure for the difference between two paired AUCs based on combined biomarkers.

Therefore, the aim of this article is to develop an approach for confidence interval estimation of the difference between paired AUCs based on optimal linear combinations of markers in the context of Su and Liu (1993). Hypothesis testing will also be addressed in the same framework. Towards this end, we will use the concepts of generalized confidence intervals and generalized P -values. The generalized P -value was introduced by Tsui and Weerahandi (1989) and the generalized confidence interval by Weerahandi (1993); see the book by Weerahandi (2003) for a detailed discussion. The concepts of generalized confidence interval and generalized P -value have been widely applied to a variety of practical settings where standard solutions do not exist and the excellent small sample properties have been demonstrated; e.g., see Weerahandi (1995), Weerahandi and Berger (1999), Krishnamoorthy and Lu (2003), Tian and Cappelleri (2004), Iyer et al. (2004), Lin et al. (2007), Li et al. (2008), Krishnamoorthy and Tian (2008), and Lin and Liao (2008). The generalized variable approaches not only can provide confidence intervals, but also can easily provide P -values for hypothesis testing. Usually the generalized approach can be carried out via a few straightforward simulation steps. Theoretically speaking, the proposed approach applies to any combinations of biomarkers. Specifically we will focus on the optimal linear combination of markers since it is of more interest than others.

This paper is organized as follows. Notations and background knowledge are introduced in Section 2. In Section 3, the generalized variable approach is proposed. In Section 4, simulation results are presented for evaluating the performance of the proposed approach on coverage probabilities. In Section 5, the proposed approach is applied to a subset of data from a study on coronary heart disease. Section 6 contains summary and discussion.

2. Preliminaries

Let X and Y denote the scores of a marker for diseased group and non-diseased groups, respectively. The area under ROC curve is equal to $P(X > Y)$ with larger AUC values indicating better diagnostic accuracy. Assume $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \xi^2)$, then

$$AUC = \Phi\left(\frac{\mu - \nu}{\sqrt{\sigma^2 + \xi^2}}\right). \quad (1)$$

Now consider the multivariate case. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y} \sim N_p(\boldsymbol{\xi}, \boldsymbol{\Xi})$ be p -dimensional vectors denoting the scores for p different markers for the diseased ($\mathbf{X}$) group and non-diseased ($\mathbf{Y}$) group, respectively. Furthermore, assuming that $\mathbf{X}$ and $\mathbf{Y}$ can be partitioned into two vectors with p_1 and p_2 ($p_1 + p_2 = p$) dimensions, i.e.

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \quad (2)$$

and

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{pmatrix}\right), \quad (3)$$

where $\mathbf{X}_1$ and $\mathbf{Y}_1$ are p_1 -dimensional vectors, as well as $\mathbf{X}_2$ and $\mathbf{Y}_2$ are p_2 -dimensional vectors. for $i = 1, 2$, consider any linear combinations $U_i = \mathbf{c}_i' \mathbf{X}_i$ and $V_i = \mathbf{c}_i' \mathbf{Y}_i$ where $\mathbf{c}_i$ is a p_i -dimensional vector as the combination coefficient, then

$$E(U_i) = \mathbf{c}_i' \mu_i \quad \text{and} \quad E(V_i) = \mathbf{c}_i' \nu_i, \quad (4)$$

$$\text{Var}(U_i) = \mathbf{c}_i' \Sigma_i \mathbf{c}_i \quad \text{and} \quad \text{Var}(V_i) = \mathbf{c}_i' \Xi_i \mathbf{c}_i. \quad (5)$$

The AUCs for these combined markers (U_1 and V_1 ; U_2 and V_2) are

$$AUC_i = \Phi\left[\frac{E(U_i) - E(V_i)}{\sqrt{\text{Var}(U_i) + \text{Var}(V_i)}}\right] = \Phi\left[\frac{\mathbf{c}_i'(\mu_i - \nu_i)}{\sqrt{\mathbf{c}_i'(\Sigma_{ii} + \Xi_{ii})\mathbf{c}_i}}\right] \quad \text{for } i=1, 2. \quad (6)$$

As shown by Su and Liu (1993), to achieve the maximum AUC_1 and AUC_2 , the optimal coefficients $\mathbf{c}_i$ are

$$\mathbf{C}_i = (\Sigma_{ii} + \Xi_{ii})^{-1}(\mu_i - \nu_i) \quad \text{for } i=1, 2 \quad (7)$$

which lead to the optimal AUC_1 based on the combined markers of first p_1 components of $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{AUC}_1 = \Phi \left[\sqrt{(\mu_1 - \nu_1)'(\Sigma_{11} + \Xi_{11})^{-1}(\mu_1 - \nu_1)} \right] \quad (8)$$

and the optimal $\mathbf{AUC}_2$ based on the combined marker of last p_2 components of $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathbf{AUC}_2 = \Phi \left[\sqrt{(\mu_2 - \nu_2)'(\Sigma_{22} + \Xi_{22})^{-1}(\mu_2 - \nu_2)} \right]. \quad (9)$$

It is of our interest to estimate confidence interval of

$$\Delta \mathbf{AUC} = \mathbf{AUC}_1 - \mathbf{AUC}_2. \quad (10)$$

3. The generalized variable approach

The goal of this paper is to provide an approach for confidence interval estimation of $\Delta \mathbf{AUC}$ based on the concepts of generalized confidence interval and generalized pivot reviewed in the Appendix. Assume that $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{n1}$ consist a sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{n2}$ consist a sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\xi})$. Denote

$$\bar{\mathbf{X}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{X}_i, \quad \mathbf{S}_x = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \quad (11)$$

and

$$\bar{\mathbf{Y}} = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbf{Y}_i, \quad \mathbf{S}_y = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})' \quad (12)$$

to be the sample mean vector and sample covariance matrix for the diseased and non-diseased groups, respectively. Furthermore, $\bar{\mathbf{X}}, \bar{\mathbf{Y}}, \mathbf{S}_x, \mathbf{S}_y$ can be partitioned as follows:

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{\mathbf{X}}_1 \\ \bar{\mathbf{X}}_2 \end{pmatrix}, \quad \mathbf{S}_x = \begin{pmatrix} \mathbf{S}_{x11} & \mathbf{S}_{x12} \\ \mathbf{S}_{x21} & \mathbf{S}_{x22} \end{pmatrix} \quad (13)$$

and

$$\bar{\mathbf{Y}} = \begin{pmatrix} \bar{\mathbf{Y}}_1 \\ \bar{\mathbf{Y}}_2 \end{pmatrix}, \quad \mathbf{S}_y = \begin{pmatrix} \mathbf{S}_{y11} & \mathbf{S}_{y12} \\ \mathbf{S}_{y21} & \mathbf{S}_{y22} \end{pmatrix}, \quad (14)$$

where $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{Y}}_1$ are p_1 -dimensional, as well as $\bar{\mathbf{X}}_2$ and $\bar{\mathbf{Y}}_2$ are p_2 -dimensional. Now define

$$\mathbf{T}_x = \frac{n_1 - 1}{n_1} \mathbf{S}_x \quad \text{and} \quad \mathbf{T}_y = \frac{n_2 - 1}{n_2} \mathbf{S}_y. \quad (15)$$

The $\mathbf{T}_x$ and $\mathbf{T}_y$ can be partitioned similarly as $\mathbf{S}_x$ and $\mathbf{S}_y$ as follows:

$$\mathbf{T}_x = \begin{pmatrix} \mathbf{T}_{x11} & \mathbf{T}_{x12} \\ \mathbf{T}_{x21} & \mathbf{T}_{x22} \end{pmatrix} \quad \text{and} \quad \mathbf{T}_y = \begin{pmatrix} \mathbf{T}_{y11} & \mathbf{T}_{y12} \\ \mathbf{T}_{y21} & \mathbf{T}_{y22} \end{pmatrix}. \quad (16)$$

Denote $\bar{\mathbf{x}}$, $\mathbf{t}_x$, $\bar{\mathbf{y}}$ and $\mathbf{t}_y$ as the observed values of $\mathbf{X}$, $\mathbf{T}_x$, $\mathbf{Y}$ and $\mathbf{T}_y$, respectively. Lin et al. (2007) developed the generalized inferences for the common mean vector of several multivariate normal population. In their paper, the generalized pivots for multivariate normal mean and the covariance matrix were also presented. Using their results, the generalized pivots for μ , v , Σ and Ξ are

$$R_\mu = \bar{\mathbf{x}} - (\mathbf{t}_x^{1/2} \mathbf{W}_x^{-1} \mathbf{t}_x^{1/2})^{1/2} \mathbf{Z}, \quad R_v = \bar{\mathbf{y}} - (\mathbf{t}_y^{1/2} \mathbf{W}_y^{-1} \mathbf{t}_y^{1/2})^{1/2} \mathbf{Z}, \quad R_\Sigma = n_1 \mathbf{t}_x^{1/2} \mathbf{W}_x^{-1} \mathbf{t}_x^{1/2}, \quad R_\Xi = n_2 \mathbf{t}_y^{1/2} \mathbf{W}_y^{-1} \mathbf{t}_y^{1/2}, \quad (17)$$

where $\mathbf{Z} \sim N_p(0, I_p)$, $\mathbf{W}_x \sim \text{Wishart}_p(n_1 - 1, I_p)$, and $\mathbf{W}_y \sim \text{Wishart}_p(n_2 - 1, I_p)$.

Partitioning R_μ , R_Σ , R_v and R_Ξ as

$$R_\mu = \begin{pmatrix} R_{\mu_1} \\ R_{\mu_2} \end{pmatrix} \quad \text{and} \quad R_\Sigma = \begin{pmatrix} R_{\Sigma_{11}} & R_{\Sigma_{12}} \\ R_{\Sigma_{21}} & R_{\Sigma_{22}} \end{pmatrix}, \quad R_v = \begin{pmatrix} R_{v_1} \\ R_{v_2} \end{pmatrix} \quad \text{and} \quad R_\Xi = \begin{pmatrix} R_{\Xi_{11}} & R_{\Xi_{12}} \\ R_{\Xi_{21}} & R_{\Xi_{22}} \end{pmatrix}, \quad (18)$$

where R_{μ_1} and R_{v_1} are p_1 -dimensional, as well as R_{μ_2} and R_{v_2} are p_2 -dimensional. The generalized pivot for the optimal coefficient $\mathbf{c}_1$ and $\mathbf{c}_2$ by (7) can be easily obtained as follows:

$$R_{\mathbf{c}_1} = (R_{\Sigma_{11}} + R_{\Xi_{11}})^{-1} (R_{\mu_{11}} - R_{v_{11}}), \quad R_{\mathbf{c}_2} = (R_{\Sigma_{22}} + R_{\Xi_{22}})^{-1} (R_{\mu_{22}} - R_{v_{22}}). \quad (19)$$

It is straightforward to check that $R_{\mathbf{c}_1}$ and $R_{\mathbf{c}_2}$ are bona fide generalized pivots for $\mathbf{c}_1$ and $\mathbf{c}_2$.

Remark 3.1. For any combinations of markers, the generalized pivot for the vector of combination coefficient can be obtained similarly; that is, by substituting μ_i , Σ_{ii} , v_i and Ξ_{ii} with R_{μ_i} , $R_{\Sigma_{ii}}$, R_{v_i} and $R_{\Xi_{ii}}$, respectively, in the vector of combination coefficient.

Denote sample means for U_i and V_i as $\bar{U}_i$ and $\bar{V}_i$ for $i = 1, 2$. Note that

$$\frac{(\bar{U}_i - \bar{V}_i) - (E(U_i) - E(V_i))}{\sqrt{\text{Var}(U_i)/n_1 + \text{Var}(V_i)/n_2}} \sim N(0, 1), \quad (20)$$

then the generalized pivot for $E(U_i) - E(V_i)$ is

$$R_{E(U_i) - E(V_i)} = R_{\mathbf{c}_i}, \quad \left(\bar{\mathbf{x}}_i - \bar{\mathbf{y}}_i \right) - \sqrt{\frac{R_{\text{Var}(U_i)}}{n_1} + \frac{R_{\text{Var}(V_i)}}{n_2}} \mathbf{Z}, \quad (21)$$

where $R_{\text{Var}(U_i)}$ and $R_{\text{Var}(V_i)}$ are the generalized pivots for $\text{Var}(U_i)$ and $\text{Var}(V_i)$ defined below, $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{y}}_i$ are the observed sample mean vectors for $\mathbf{X}_i$ and $\mathbf{Y}_i$, $\mathbf{s}_{xii}$ and $\mathbf{s}_{yii}$ are the observed sample covariance matrix for $\mathbf{S}_{Xii}$ and $\mathbf{S}_{Yii}$, and $\mathbf{Z} \sim N(0, 1)$.

Denote the sample variances for U_i and V_i as SSU_i and SSV_i . It is easy to see

$$SSU_i = \mathbf{c}_i' S x_{ii} \mathbf{c}_i, \quad SSV_i = \mathbf{c}_i' S y_{ii} \mathbf{c}_i. \quad (22)$$

Note the facts that

$$\frac{(n_1 - 1) SSU_i}{Var(U_i)} = \chi_{n_1-1}^2, \quad \frac{(n_2 - 1) SSV_i}{Var(V_i)} = \chi_{n_2-1}^2, \quad (23)$$

where $Var(U_i)$ and $Var(V_i)$ are defined as (4). Therefore, the generalized pivots for $Var(U_i)$ and $Var(V_i)$ are

$$R_{Var(U_i)} = \frac{R_{\mathbf{c}_i}' (n_1 - 1) s_{xii} R_{\mathbf{c}_i}}{\frac{(n_1 - 1) \mathbf{c}_i' S x_{ii} \mathbf{c}_i}{Var(U_i)}}, \quad R_{Var(V_i)} = \frac{R_{\mathbf{c}_i}' (n_2 - 1) s_{yii} R_{\mathbf{c}_i}}{\frac{(n_2 - 1) \mathbf{c}_i' S y_{ii} \mathbf{c}_i}{Var(V_i)}}. \quad (24)$$

It is easy to verify that $R_{Var(U_i)}$ and $R_{Var(V_i)}$ are bona fide generalized pivots.

Furthermore, the generalized pivots for AUC_i are

$$R_{AUC_i} = \Phi \left(\frac{R_{E(U_i) - E(V_i)}}{\sqrt{R_{Var(U_i)} + R_{Var(V_i)}}} \right) \quad \text{for } i = 1, 2. \quad (25)$$

Eventually, the generalized pivot for $\Delta AUC = AUC_1 - AUC_2$ is

$$R_{\Delta AUC} = R_{AUC_1} - R_{AUC_2}. \quad (26)$$

Note that for given $\mathbf{x}$, $\mathbf{t}_x$, $\mathbf{y}$ and $\mathbf{t}_y$, the following holds: (1) the distribution of $R_{\Delta AUC}$ is independent of any unknown parameters, and (2) the value of $R_{\Delta AUC}$ is ΔAUC as $\mathbf{X} = \mathbf{x}$, $\mathbf{T}_X = \mathbf{t}_x$, $\mathbf{Y} = \mathbf{y}$ and $\mathbf{T}_Y = \mathbf{t}_y$. Therefore, $R_{\Delta AUC}$ is a bona fide generalized pivotal quantities for constructing confidence interval of ΔAUC . For testing the hypothesis $H_0 : \Delta AUC \geq \Delta AUC_0$ vs. $H_1 : \Delta AUC < \Delta AUC_0$, the generalized variable is defined as

$$T_{\Delta AUC} = R_{\Delta AUC} - \Delta AUC. \quad (27)$$

It is easy to see that $T_{\Delta AUC}$ is a bona fide generalized test variable.

Remark 3.2. There exist many bona fide generalized pivots for ΔAUC . However, not all of them can provide reasonable confidence interval estimates. For example, a bona fide generalized pivot for ΔAUC can be directly obtained by first substituting μ_i , Σ_{ii} , Ξ_i and Ξ_{ii} with R_{μ_i} , $R_{\Sigma_{ii}}$, R_{Ξ_i} and $R_{\Xi_{ii}}$, respectively, in (8) and (9) to obtain R_{AUC_1} and R_{AUC_2} , and then substituting AUC_1 and AUC_2 with R_{AUC_1} and R_{AUC_2} in (10). However, simulation study demonstrated that such obtained generalized pivot generate very liberal confidence intervals. Therefore, this approach is not pursued further.

3.1. Simulation steps

For a given data set $\mathbf{x}=(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_1})'$ from diseased group and $\mathbf{y}_2=(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{n_2})'$ from non-diseased group, the generalized confidence intervals for ΔAUC based on the optimal combinations can be computed via the following steps:

1. Compute $\bar{\mathbf{x}}, \bar{\mathbf{t}}_{\mathbf{x}}, \bar{\mathbf{y}}$ and $\bar{\mathbf{t}}_{\mathbf{y}}$.
2. Generate $\mathbf{Z}_x \sim N_p(0, I_p)$, $\mathbf{W}_x \sim W_p(n_1 - 1, I_p)$, $\mathbf{Z}_y \sim N_p(0, I_p)$ and $\mathbf{W}_y \sim W_p(n_1 - 1, I_p)$. Calculate R_μ, R_Σ, R_V and R_Ξ following (17).
3. Calculate R_{c_1} and R_{c_2} following (19).
4. Generate $Z \sim N(0, 1)$. Calculate $R_{E(U_i)-E(V_i)}$ following (21) for $i = 1, 2$.
5. Generate $V_1\tilde{\chi}_{n_1-1}^2$ and $V_2\tilde{\chi}_{n_2-1}^2$. Calculate $R_{var(U_i)}$ and $R_{var(V_i)}$ as (24) for $i = 1, 2$.
6. Compute R_{AUC_i} following (25). Compute $R_{\Delta\text{AUC}}$ following (26.)
7. Repeat Steps 2–6 a total m times and obtain an array of $R_{\Delta\text{AUC}}$'s.
8. Rank this array of $R_{\Delta\text{AUC}}$'s and from small to large.

The 100 α th percentile of $R_{\Delta\text{AUC}}$'s, namely $R_{\Delta\text{AUC}}(\alpha)$, is an estimate of the lower bound of the one-sided 100(1 - α)% confidence interval and ($R_{\Delta\text{AUC}}(\alpha/2)$, $R_{\Delta\text{AUC}}(1 - \alpha/2)$) is an estimate of the two-sided 100(1 - α)% confidence interval of ΔAUC . The percentage that $R_{\Delta\text{AUC}}$'s are larger than or equal to ΔAUC_0 is a Monte Carlo estimate of the generalized P -value for testing $\Delta\text{AUC} = \Delta\text{AUC}_0$ vs. $\Delta\text{AUC} < \Delta\text{AUC}_0$. Similarly, the generalized P -values for testing $\Delta\text{AUC} = \Delta\text{AUC}_0$ vs. $\Delta\text{AUC} \neq \Delta\text{AUC}_0$ can be obtained.

4. Simulation studies

Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y} \sim N_p(\mathbf{v}, \boldsymbol{\Xi})$ be p -dimensional vectors denoting the scores for p different markers for the diseased ($\mathbf{X}$) group and non-diseased ($\mathbf{Y}$) group, respectively. Reiser and Faraggi (1997) argued that for any $\mathbf{X}$ and $\mathbf{Y}$, there always exist a non-singular transformation matrix D which makes the covariance matrix A for $D\mathbf{X}$ to be diagonal and that equals $I - A$ for $D\mathbf{Y}$. Hence, inference on ΔAUC based on $\mathbf{X}$ and $\mathbf{Y}$ is equivalent to that based on $D\mathbf{X}$ and $D\mathbf{Y}$. Therefore, in this simulation study, a variety of scenarios for diagonal covariance matrix A were set up for $\mathbf{X}$ and $I - A$ for $\mathbf{Y}$. The setting for the covariance matrix is listed in Tables 1 and 2. We consider following scenarios: $p = 3$ with $p_1 = 2$ and $p_2 = 1$; as well as $p = 5$ with $p_1 = 3$ and $p_2 = 2$. The mean vector $\mathbf{m}$ for $\mathbf{Y}$ is set as $\mathbf{0}$. For the structure of the mean vector $\boldsymbol{\mu}$ for $\mathbf{X}$, two scenarios are considered: (a) all the components of $\boldsymbol{\mu}$ are equal; (b) as $p = 3$, $\mu_1 = 2\mu_2 = 2\mu_3$; as $p = 5$, $\mu_1 = \mu_2 = 2\mu_3 = 2\mu_4 = 2\mu_5$. For $p = 3$, $n_1 = n_2 = 30, 50, 100$ and $n_1 = n_2 = 50, 100, 200$ for $p = 5$.

For each parameter setting, 2000 random samples were generated. To estimate the confidence intervals using the proposed approach, within each of the 2000 random samples, 2000 $R_{\Delta\text{AUC}}$'s were obtained via the algorithms presented in Section 3.1. According to large sample theory, a coverage probability between 88.7% and 96.3% was considered as satisfying for nominal level 90% confidence interval and a coverage probability between 94% and 96% was considered as satisfying for nominal level 95% confidence interval.

Tables 1 and 2 display the coverage probabilities of the estimated coverage probabilities by the proposed confidence intervals when $p = 3$, $p_1 = 2$, $p_2 = 1$ and $p = 5$, $p_1 = 3$, $p_2 = 2$, respectively. The results presented in Tables 1 and 2 are all for the scenario (a) in terms of the structure of the mean vector $\boldsymbol{\mu}$ defined above. Overall speaking, the proposed approach provides

confidence intervals with satisfying coverage probabilities for ΔAUC except that it tends to be slightly liberal for a few scenarios.

In addition to the results presented in Tables 1 and 2, there are many more scenarios considered in this simulation study. The simulation results are omitted here, but they are available upon request. The following Remarks 4.1–4.3 provide a summary.

Remark 4.1. Simulation study was also carried out for the scenario (b) for the mean vector μ defined above. The conclusions about the performance of the proposed approach for confidence interval estimations are similar to those for the scenario (a).

Remark 4.2. Simulation study was also carried out for unbalanced designs, that is, as $n_1 \neq n_2$. The conclusions about the performance of the proposed approach for confidence interval estimation are similar as that for balanced designs.

Remark 4.3. As stated in Section 3, the proposed approach can easily provide P -values for hypothesis testing. Simulation study shows that the proposed test has satisfying type-I error control.

5. A coronary heart disease example

We illustrate the proposed approach using data from a study on atherosclerotic process of coronary heart disease. A population-based sample of randomly selected residents of Erie and Niagara counties of New York state, USA, 35–79 years of age, was the focus of this study by Schisterman et al. (2002). Free radicals have been implicated in CHD. Well-developed laboratory methods may make available a large number of biomarkers of individual oxidative stress and antioxidant status. Such markers are able to quantify different phases of the oxidative stress and antioxidant status of an individual. However, limited knowledge is available on how to combine these biomarkers to best discriminate between individuals with and without CHD. Finding such a combination can be cost effective and clinically important for the early identification and prevention of coronary heart disease. It was previously found the combination of all these multiple markers of oxidative stress does not greatly improve ability to differentiate between individuals with and without CHD compared to the use of TBARS alone (Schisterman et al., 2002). However, the comparison of the discriminatory ability of TBARS with that of combinations of standard biomarkers like uric acid, Body Mass Index, globulin and total protein, as indicators of CHD, has never been addressed. To illustrate the proposed approach, we obtain a subset of this data set with 50 cases and 50 controls. We are interested in comparing the discriminatory abilities between TBARS and the combination of uric acid, BMI, globulin and total protein (comparison #1). Furthermore, for illustrative purposes, we consider the comparison of the combination of TBARS and uric acid vs. combination of BMI, globulin and total protein (comparison #2). The data were first log-transformed to tested for multivariate normality using method proposed by Mudholkar et al. (1992). The sample means and variances $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, $\mathbf{s}_x$, $\mathbf{s}_y$ of the variables (TBARS, uric acid, BMI, globulin, total protein) for CHD group and control group, respectively, are as follows:

$$\bar{\mathbf{x}} = \begin{pmatrix} 0.3982 \\ 1.8248 \\ 3.3934 \\ 0.9878 \\ 1.9679 \end{pmatrix} \quad \text{and} \quad \mathbf{s}_x = \begin{pmatrix} 0.0878 & -0.0018 & 0.0088 & 0.0158 & 0.0047 \\ -0.0018 & 0.0539 & 0.0135 & 0.0071 & 0.0019 \\ 0.0088 & 0.0135 & 0.0214 & 0.0069 & 0.0013 \\ 0.0158 & 0.0071 & 0.0069 & 0.0292 & 0.0100 \\ 0.0047 & 0.0019 & 0.0013 & 0.0100 & 0.0045 \end{pmatrix},$$

$$\bar{\mathbf{y}} = \begin{pmatrix} 0.2655 \\ 1.7286 \\ 3.2904 \\ 0.9280 \\ 1.9406 \end{pmatrix} \quad \text{and} \quad \mathbf{s}_y = \begin{pmatrix} 0.0500 & -0.0012 & 0.0073 & 0.0092 & 0.0042 \\ -0.0012 & 0.0615 & 0.0090 & -0.0062 & -0.0012 \\ 0.0073 & 0.0090 & 0.0257 & 0.0008 & -0.0005 \\ 0.0092 & -0.0062 & 0.0008 & 0.0347 & 0.0112 \\ 0.0042 & -0.0012 & -0.0005 & 0.0112 & 0.0053 \end{pmatrix}.$$

The estimated overall AUC using all of these five markers equals 0.7291. For comparison #1, the estimated AUC based only on TBARS is 0.6389, and 0.7135 for the optimal combination of the other four markers. Using the proposed approach, the 95% confidence interval for the difference between these two paired AUCs is $(-0.2053, 0.1111)$. The P -value equals to 0.4910 for two-sided hypothesis testing $H_0 : \Delta\text{AUC} = 0$ vs. $H_1 : \Delta\text{AUC} \neq 0$. Therefore, one can conclude there is no significant difference between the discriminatory abilities of TBARS and the combination of uric acid, BMI, globulin and total protein. For comparison #2, the estimated AUCs based on the optimal combination of TBARS and uric acid is 0.6777, and 0.7077 for the optimal combination of the other three markers. The 95% confidence interval by the proposed method for the difference between the paired AUCs as $(-0.1730, 0.1344)$. The P -value equals to 0.7400 for two-sided hypothesis testing $H_0 : \Delta\text{AUC} = 0$ vs. $H_1 : \Delta\text{AUC} \neq 0$. Therefore, one can conclude there is no significant difference between the discriminatory abilities of combination of TBARS and uric acid and the combination of BMI, globulin and total protein.

6. Summary and discussion

This paper focuses on the confidence interval estimation of ΔAUC , i.e. the difference between paired AUCs based on optimally combined markers. We present an approach based on generalized inference theory. The simulation study demonstrated that the proposed approach can be considered as a good candidate for confidence interval estimation of ΔAUC except that it tends to be slightly liberal as sample sizes are very small. The proposed generalized approach is a numeric approach which involves generating multivariate normal data and can be performed using a few straightforward simulation steps. To run the example on a PC with processor 2.00 GHz and 1.00 GB of RAM, it takes less than a second. Therefore, the computational speed is not an issue with standard equipped computers.

In addition to confidence interval estimation, the proposed approach can also easily provide P -values for hypothesis testing. Simulation studies demonstrated that the proposed approach generally has satisfactory type-I error control. It is also worth noting that the proposed approach can be easily adapted to any types of combinations of biomarkers or to any contrasts of paired AUCs based on combined biomarkers.

This aim of this paper is to provide a procedure for confidence interval estimation for ΔAUC in the framework of multivariate normality. All the procedures are developed under this specific parametric assumption. While the investigation of the robustness of procedures might be of interest, it does not fit in the framework of this paper. Furthermore, it is of interest to extend the proposed approach to the comparison of discriminatory ability of two groups of biomarkers with overlapping components. We plan to tackle these issues in future research.

Appendix A. Generalized confidence interval and generalized P-values

In the following, the basic concepts for generalized inference developed by Tsui and Weerahandi (1989) and Weerahandi (1993) are described.

Suppose that $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ form a random sample from a distribution which depends on the parameters $\theta(\psi, \mathbf{v})$ where ψ is the parameter of interest and $\mathbf{v}'$ is a vector of nuisance parameters. A generalized pivot $R(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$, where $\mathbf{y}$ is an observed value of $\mathbf{Y}$, for interval estimation defined in Weerahandi (1993), has the following two properties:

1. $R(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$ has a distribution free of unknown parameters.
2. The value of $R(\mathbf{y}; \mathbf{y}, \psi, \mathbf{v})$ is ψ .

Let that R_α be the 100 α th percentile of R . Then R_α becomes the 100(1 - α)% lower bound for ψ and $(R_{\alpha/2}, R_{1-\alpha/2})$ becomes a 100(1 - α)% two-sided generalized confidence interval for ψ .

Now consider testing $H_0: \psi = \psi_0$ vs. $H_1: \psi < \psi_0$ where ψ_0 is a specified quantity. A generalized test variable of the form $T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$, where $\mathbf{y}$ is an observed value of $\mathbf{Y}$, is chosen to satisfy the following three conditions (Tsui and Weerahandi, 1989):

1. For fixed $\mathbf{y}$, the distribution of $T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$ is free of the vector of nuisance parameters $\mathbf{v}$.
2. The value of $T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$ at $\mathbf{Y} = \mathbf{y}$ is free of any unknown parameters.
3. For fixed $\mathbf{y}$ and $\mathbf{v}$, and for all t , $\Pr[T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v}) > t]$ is either an increasing or a decreasing function of ψ .

A generalized extreme region is defined as $C = [\mathbf{Y} : T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v}) \leq T(\mathbf{y}; \mathbf{y}, \psi, \mathbf{v})]$ if $T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$ is stochastically increasing in ψ . If $T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v})$ is stochastically decreasing in ψ , a generalized extreme region is defined as $C = [\mathbf{Y} : T(\mathbf{Y}; \mathbf{y}, \psi, \mathbf{v}) \geq T(\mathbf{y}; \mathbf{y}, \psi, \mathbf{v})]$. The generalized P -value is defined as $P(C|\psi_0)$.

References

- Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrika* 1988;44:837–845.
- Iyer HK, Wang CM, Mathew T. Models and confidence intervals for true values in interlaboratory trials. *J. Acoust. Soc. Amer* 2004;99:1060–1071.
- Krishnamoorthy K, Lu Y. Inference on the common means of several normal populations based on the generalized variable method. *Biometrics* 2003;59:237–247. [PubMed: 12926708]
- Krishnamoorthy K, Tian L. Inferences on the difference and ratio of the means of two inverse Gaussian distributions. *J. Statist. Plann. Inference* 2008;138:2082–2089.
- Li CR, Liao CT, Liu J. On the exact interval estimation for the difference in paired areas under the ROC curves. *Statist. Med* 2008;27:224–242.
- Lin SH, Lee JC, Wang RS. Generalized inferences on the common mean vector of several multivariate normal populations. *J. Statist. Plann. Inference* 2007;137:2240–2249.
- Lin SH, Liao CT. Prediction intervals for general balanced linear random models. *J. Statist. Plann. Inference* 2008;138:3164–3175.
- Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Statist. Med* 2005;24:37–47.
- Molodianovitch K, Faraggi D, Reiser B. Comparing the areas under two correlated ROC curves: parametric and non-parametric approaches. *Biometrical J* 2006;48:745–757.
- Mudholkar GS, McDermott M, Srivastava DK. A test of p-variate normality. *Biometrika* 1992;79:850–854.
- Obuchowski N. Testing for equivalence of diagnostic tests. *Amer. J. Radiol* 1997;168:13–17.

- Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy Margaret. *Biostatistics* 2000;1–2:123–140.
- Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series 2003:28.
- Reiser B, Faraggi D. Confidence intervals for the generalized ROC criterion. *Biometrics* 1997;53:644–652. [PubMed: 9235120]
- Schisterman EF, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Reiser B, Trevisan M. Minimal and best linear combination of oxidative stress and antioxidant biomarkers to discriminate cardiovascular disease. *Nutr. Metab. Cardiovasc. Dis* 2002;12:259–266. [PubMed: 12616805]
- Shapiro D. The interpretation of diagnostic tests. *Statist. Methods Med. Res* 1999;8:113–134.
- Su J, Liu JS. Linear combinations of multiple diagnostic markers. *J. Amer. Statist. Assoc* 1993;88:1350–1355.
- Tsui KW, Weerahandi S. Generalized P-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Statist. Assoc* 1989;84:602–607.
- Tian L, Cappelleri JC. A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method. *Statist. Med* 2004;23:2125–2135.
- Vexler A, Liu A, Schisterman EF, Wu C. Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *J. Nonparametric Statist* 2006;18:145–158.
- Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* 2008;64:895–903. [PubMed: 18047527]
- Weerahandi S. Generalized confidence intervals. *J. Amer. Statist. Assoc* 1993;88:899–905.
- Weerahandi S. ANOVA under unequal error variances. *Biometrics* 1995;51:589–599.
- Weerahandi S, Berger VW. Exact inference for growth curves with intraclass correlation structure. *Biometrics* 1999;55:921–924. [PubMed: 11315029]
- Weerahandi, S. *Exact Statistical Methods for Data Analysis*. Springer; New York: 2003.
- Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989;76:585–592.
- Zhang DD, Zhou XH, Freeman DH, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statist. Med* 2002;21:701–715.
- Zhou, X-H.; Obuchowski, N.; McClish, D. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.

Table 1

Empirical coverage probabilities ($\times 1000$) of approximate 90% (CV^1) and 95% (CV^2) two-sided confidence bounds for $\Delta AUC = AUC_1 - AUC_2$ when $p_1 = 2$ and $p_2 = 1$ (based on 2000 simulations).

$(\lambda_1, \lambda_2, \lambda_2)$	AUC	Sample sizes					
		(30,30)		(50,50)		(100,100)	
		CV^1	CV^2	CV^1	CV^2	CV^1	CV^2
(0.1, 0.1, 0.1)	0.9	906	950	911	950	894	951
	0.8	903	951	912	949	893	948
	0.7	915	960	897	950	895	947
(0.5, 0.5, 0.5)	0.9	909	958	912	956	902	948
	0.8	908	954	910	956	897	947
	0.7	920	965	905	951	896	947
(0.9, 0.9, 0.9)	0.9	900	948	910	957	906	954
	0.8	901	947	906	949	906	949
	0.7	909	956	905	955	905	949
(0.1, 0.1, 0.9)	0.9	903	949	902	952	900	949
	0.8	903	947	908	955	900	945
	0.7	912	958	902	952	899	947
(0.9, 0.9, 0.1)	0.9	903	955	902	952	897	958
	0.8	905	951	904	951	905	953
	0.7	914	961	899	957	900	948
(0.9, 0.1, 0.1)	0.9	904	949	908	953	909	949
	0.8	903	952	902	949	901	948
	0.7	915	960	902	948	900	948
(0.9, 0.1, 0.9)	0.9	903	950	908	955	905	949
	0.8	898	951	909	955	903	954
	0.7	910	961	909	954	897	946

Table 2

Empirical coverage probabilities ($\times 1000$) of approximate 90% (CV^1) and 95% (CV^2) two-sided confidence bounds for $\Delta AUC = AUC_1 - AUC_2$ when $p_1 = 3$ and $p_2 = 2$ (based on 2000 simulations).

(λ_1-s)	AUC	Sample sizes			CV^1	CV^2	(100,100)			CV^1	CV^2	(200,200)		
		(50,50)	CV^1	CV^2			CV^1	CV^2	CV^1	CV^2	CV^1	CV^2	CV^1	CV^2
(0.1, 0.1, 0.1, 0.1, 0.9)	0.9	902		958		912		960		899		945		
	0.8	908		956		901		947		900		945		
	0.7	917		962		902		948		902		945		
(0.5, 0.5, 0.5, 0.5, 0.9)	0.9	904		957		907		957		896		956		
	0.8	908		954		905		956		899		947		
	0.7	920		964		893		943		900		951		
(0.9, 0.9, 0.9, 0.9, 0.9)	0.9	903		945		899		951		901		948		
	0.8	899		951		893		947		900		946		
	0.7	913		957		892		950		899		945		
(0.1, 0.1, 0.1, 0.9, 0.9)	0.9	906		952		989		947		893		947		
	0.8	902		953		900		946		894		947		
	0.7	911		959		906		951		894		946		
(0.9, 0.9, 0.9, 0.1, 0.1)	0.9	902		952		902		951		897		943		
	0.8	911		957		894		949		897		948		
	0.7	918		964		900		951		895		948		
(0.9, 0.1, 0.5, 0.9, 0.9)	0.9	906		954		906		950		902		954		
	0.8	906		953		905		952		901		953		
	0.7	919		963		901		954		905		953		
(0.9, 0.1, 0.5, 0.1, 0.9)	0.9	899		950		906		955		898		948		
	0.8	900		953		892		943		895		947		
	0.7	912		959		892		943		895		947		
(0.1, 0.1, 0.9, 0.1, 0.1)	0.9	903		959		902		951		901		949		
	0.8	906		958		897		953		900		947		
	0.7	918		964		910		951		899		948		
(0.1, 0.1, 0.9, 0.9, 0.9)	0.9	909		956		900		946		896		950		
	0.8	911		959		899		948		898		951		
	0.7	916		966		902		950		900		953		
(0.1, 0.9, 0.9, 0.1, 0.1)	0.9	910		960		895		953		906		950		
	0.8	909		962		900		948		900		947		
	0.7	918		967		910		955		900		949		
(0.1, 0.9, 0.9, 0.9, 0.9)	0.9	903		955		894		951		899		952		
	0.8	904		956		892		947		901		952		
	0.7	912		959		904		957		900		953		