Published in final edited form as:

Biom J. 2010 June ; 52(3): 417–435. doi:10.1002/bimj.200900228.

Evaluating the Improvement in Diagnostic Utility from Adding New Predictors

Caixia Li^{1,2} and Ying Lu^{*,1,3,4,5}

¹Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, CA 94143-0946, USA

²Department of Medical Statistics and Epidemiology, Sun Yat-Sen University, Guangzhou, Guangdong, 510275, P.R.China

³Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94143, USA

⁴Veterans Affairs Cooperative Studies Program Palo Alto Coordinating Center, VA Palo Alto Health Care System, Mountain View, CA 94043, USA

⁵Department of Health Research and Policy, Stanford University, CA 94305-5405, USA

Summary

Multiple diagnostic tests and risk factors are commonly available for many diseases. This information can be either redundant or complimentary. Combinations of these tests and risk factors may improve the diagnostic/predictive accuracy but also may unnecessarily increase complexity, risks, and/or costs. The improved accuracy gained by including additional variables can be evaluated by the increment of the area under (AUC) the receiver operating characteristic (ROC) curves with and without the new variable(s). In this paper, we derive a new test statistic to accurately and efficiently determine the statistical significance of this incremental AUC under a multivariate normality assumption. Our test links the difference in AUC to a quadratic form of a standardized mean shift in a unit of the inverse covariance matrix through a properly linear transformation of all diagnostic variables. The distribution of the estimator of the quadratic form is related to the multivariate Behrens-Fisher problem. We provide explicit mathematical solutions of the estimator and its approximate noncentral F-distribution, type I error rate, and sample size formula. We use simulation studies to prove that our new test maintains prespecified type I error rates as well as reasonable statistical power under practical sample sizes. We use data from the Study of Osteoporotic Fractures (SOF) as an application example to illustrate our method.

Keywords

Area under ROC curves; Behrens-Fisher problem; Noncentral F distribution; Receiver operating characteristic curve

Conflict of Interests Statement

The authors have declared no conflict of interest.

1 Introduction

The utility of a diagnostic test is measured in sensitivity and specificity. For ordinal diagnostic variables, a receiver operating characteristic (ROC) curve plots the false positive rate (FPR, or 1-specificity) and the true positive rate (TPR, or sensitivity) over the diagnostic decision range. The area under the ROC curve (AUC), ranging from 0.5 to 1, is a summary measure of overall diagnostic utility (Pepe et al., 2004; Wang et al., 2006; Ware, 2006). For many diseases, more than one diagnostic test is available, and combining them may improve the diagnostic accuracy. While the optimal method to summarize multiple variables is the likelihood ratio score (McIntosh and Pepe, 2002; Neyman and Pearson, 1933), the most common medical applications use simple linear combinations. Jin and Lu proposed a bootstrap-based non-parametric test (Jin and Lu, 2008) and a permutation test (Jin and Lu, 2009a) to confirm the non-inferiority of linear combinations to the optimal likelihood ratio score. In the remaining paper, we will focus only on linear combinations of diagnostic variables. The ROC curve corresponding to linear combinations has been called the generalized ROC (GROC) curve by Reiser and Faraggi (Reiser and Faraggi, 1997).

Under an assumption of multivariate normally distributed diagnostic variables for both diseased and non-diseased populations, Su and Liu (Su and Liu, 1993) proved that the Fisher's discriminant function is the optimal linear combination (OLC) that maximizes the AUC of the ROC curves. Jin and Lu (Jin and Lu, 2009b) extend Su and Liu to the generalized linear models. Pepe and colleagues (Pepe, Cai and Longton, 2006; Pepe and Thompson, 2000) relaxed the multivariate normality assumption and found linear forms which maximize the non-parametric estimate of the AUC. Schisterman, Schisterman, Faraggi and Reiser (Schisterman, Faraggi and Reiser, 2004) discussed linear combinations of multivariate normally distributed diagnostic variables with an adjustment for covariates. Liu, Schisterman and Zhu (Liu, Schisterman and Zhu, 2005) proposed the OLC of multivariate normal variables for a given range of specificity. Cai and Cheng (Cai and Cheng, 2008) proposed robust combination of multiple diagnostic tests for censored data. While these papers constructed the OLCs, they did not evaluate the incremental AUC by adding new variables to a panel of possibly prior multiple diagnostic variables.

The OLCs with and without new variables are different. Adding new variables to a panel of multiple diagnostic variables always increases the AUC of the ROC curve. However, the increment in the apparent AUC may be too small to justify the increase in cost, risk, or logistic complications. This paper addresses statistical estimation and inference for the incremental AUC. We focus our comparisons on linear combinations of multivariate normal diagnostic variables based on linear discriminant function. Our test compares two AUCs from two combinations based on an incremental standardized quadratic form. Explicit analytic estimator and test statistics are derived. For continuous diagnostic variables that fail the normal distributional assumption, a Box-Cox type power transformation can be used prior to analysis.

Our paper is organized as follows. In Section 2, we describe the ROC model for linear combinations and define the index of incremental utility. In Section 3, we discuss the estimator and its approximate distribution. In Section 4, we propose a hypothesis testing procedure and sample size formula. Section 5 applies the proposed methods to the Study of Osteoporotic Fractures (SOF). Section 6 presents simulation experiments, and Section 7 offers discussion and conclusions.

2 Index of Diagnostic Utility Improvement

We use **bold** letters for vectors and matrices. Let $\mathbf{X}^{(k)} = (X_1^{(k)}, X_2^{(k)}, \dots, X_p^{(k)})'$ be p -dimensional normally distributed diagnostic variables. A superscript indicates observations from diseased ($k=1$) or non-diseased ($k=0$) populations; thus $\mathbf{X}^{(0)}$ is independent of $\mathbf{X}^{(1)}$ (i.e.,

$\mathbf{X}^{(0)} \perp \mathbf{X}^{(1)}$). Let $\mathbf{X}^{(k)} \sim N(\mu_x^{(k)}, \Sigma_{xx}^{(k)})$. For a p -dimensional vector $\mathbf{a}$, the AUC of the ROC curve constructed by the linear combination $\mathbf{a}'\mathbf{X}$ is

$$\text{AUC}(\mathbf{a}'\mathbf{X}) = \Phi\left(\frac{\mathbf{a}'\mu_x}{\sqrt{\mathbf{a}'\Sigma_{xx}\mathbf{a}}}\right),$$

where $\mu_x = \mu_x^{(1)} - \mu_x^{(0)}$, $\Sigma_{xx} = \Sigma_{xx}^{(1)} + \Sigma_{xx}^{(0)}$, and Φ is the cumulative distribution function (CDF) of the standard normal distribution. As Su and Liu (Su and Liu, 1993) pointed out, the linear discriminant function

$$L(\mathbf{X}) = (\mu_x)' \Sigma_{xx}^{-1} \mathbf{X} \quad (1)$$

(or $\mathbf{a} = \Sigma_{xx}^{-1} \mu_x$) maximizes the AUC as

$$\theta_{L(\mathbf{X})} = \max_{\mathbf{a}} \text{AUC}(\mathbf{a}'\mathbf{X}) = \text{AUC}\left(\mu_x' \Sigma_{xx}^{-1} \mathbf{X}\right) = \Phi\left(\sqrt{\mu_x' \Sigma_{xx}^{-1} \mu_x}\right) \quad (2)$$

Thus the AUC of the OLC is a monotonic function of the quadratic form of a standardized mean shift in a unit of the inverse covariance matrix, $\mu_x' \Sigma_{xx}^{-1} \mu_x$. The ROC curve of the OLC is

$$\begin{aligned} \text{ROC}(t) = & \Phi\left(\frac{\mu_x' \Sigma_{xx}^{-1} \mu_x}{\sqrt{\mu_x' \Sigma_{xx}^{-1} \Sigma_{xx}^{(1)} \Sigma_{xx}^{-1} \mu_x}}\right) \\ & + \frac{\sqrt{\mu_x' \Sigma_{xx}^{-1} \Sigma_{xx}^{(0)} \Sigma_{xx}^{-1} \mu_x}}{\sqrt{\mu_x' \Sigma_{xx}^{-1} \Sigma_{xx}^{(1)} \Sigma_{xx}^{-1} \mu_x}} \Phi^{-1}(t) \quad t \in [0, 1] \end{aligned} \quad (3)$$

Adding a combination of q new variables $\mathbf{Y} = (Y_1, Y_2, \dots, Y_q)'$ to $\mathbf{X}$ will increase the AUC of the ROC curve. To determine the significance of the difference in the AUCs for the OLC of $\mathbf{X}$ versus OLC of $(\mathbf{X}', \mathbf{Y}')'$, we first give the following theorem that characterizes the amount of AUC increment by decomposing the quadratic form

$\mu_{(\mathbf{X}', \mathbf{Y}')'}' \Sigma_{(\mathbf{X}', \mathbf{Y}')'}^{-1} \mu_{(\mathbf{X}', \mathbf{Y}')'}$ of the standardized mean shift of $L(\mathbf{X}, \mathbf{Y})$ in (2) in comparison to that of $L(\mathbf{X})$.

Theorem 2.1

Let $\theta_{L(X)}$ and $\theta_{L(X,Y)}$ be the maximum AUC of the OLCs of X and (X', Y') , denoted as $L(X)$ and $L(X, Y)$, respectively. Suppose $(X^{(k)}, Y^{(k)})' \sim N(\mu^{(k)}, \Sigma^{(k)})$ and the OLC residuals of Y on X , $Z^{(k)} = Y^{(k)} - \sum_{YX} \sum_{XX}^{-1} X^{(k)}$. Then

$$\theta_{L(X,Y)} = \theta_{L(X,Z)} = \Phi \left[\left(\mu_X' \sum_{XX}^{-1} \mu_X + \mu_Z' \sum_{ZZ}^{-1} \mu_Z \right)^{1/2} \right] \quad (4)$$

where $\mu_Z = \mu_Z^{(1)} - \mu_Z^{(0)} = \mu_Y - \sum_{YX} \sum_{XX}^{-1} \mu_X$ and $\sum_{ZZ} = \sum_{ZZ}^{(1)} + \sum_{ZZ}^{(0)} = \sum_{YY} - \sum_{YX} \sum_{XX}^{-1} \sum_{XY}$.

Therefore the incremental utility of adding Y to X in the AUC can be determined by

$$\lambda(Y|X) = \left[\Phi^{-1}(\theta_{L(X,Y)}) \right]^2 - \left[\Phi^{-1}(\theta_{L(X)}) \right]^2 = \mu_{(X',Y')}' \sum_{(X',Y')'(X',Y')}^{-1} \mu_{(X',Y')}' - \mu_X' \sum_{XX}^{-1} \mu_X = \mu_Z' \sum_{ZZ}^{-1} \mu_Z \quad (5)$$

in a quadratic form $\mu_Z' \sum_{ZZ}^{-1} \mu_Z$ of the residuals $Z = Y - \sum_{YX} \sum_{XX}^{-1} X$, which is a monotone function of difference between squared roots of AUCs and represents the additionally standardized mean shift for OLC due to Y after removing all contributions of X .

Note that the regression coefficient $\beta = \sum_{XX}^{-1} \sum_{XY}$ is different from the ordinal least squared estimator of the regression coefficient of Y on X .

From (5), we have the following theorem.

Theorem 2.2

Suppose $X^{(k)}$, $Y_1^{(k)}$, and $Y_2^{(k)}$ be the p , q , and r -dimensional normally distributed diagnostic variables ($k=0,1$ for non-diseased and diseased subjects, respectively), then the incremental utility of (X, Y_1, Y_2) over X is additive, i.e.,

$$\lambda((Y_1, Y_2)|X) = \lambda(Y_1|X) + \lambda(Y_2|(X, Y_1)) = \lambda(Y_2|X) + \lambda(Y_1|(X, Y_2)) \quad (6)$$

Proof of Theorems is in Appendix A.

3 Estimation

Under the multivariate normality assumption, the incremental AUC of (X, Y) over X alone is uniquely determined by (5). Suppose that we have independent and identically-distributed

(i.i.d) random samples $(X_i^{(k)}, Y_i^{(k)})' = (X_{1i}^{(k)}, X_{2i}^{(k)}, \dots, X_{pi}^{(k)}, Y_{1i}^{(k)}, Y_{2i}^{(k)}, \dots, Y_{qi}^{(k)})'$, $i = 1, 2, \dots, n_k$ from $N(\mu^{(k)}, \Sigma^{(k)})$, where the subscripts indicate vector components from the i th subject in the k th group. Denoting sample means and covariance matrices for each group k as, $\bar{X}^{(k)}$, $\bar{Y}^{(k)}$, $S_{XX}^{(k)}$, $S_{XY}^{(k)}$, and $S_{YY}^{(k)}$, $\hat{\beta} = S_{XX}^{-1} S_{XY}$ is the consistent estimator of regression coefficient β , with $S_{XX} = S_{XX}^{(0)} + S_{XX}^{(1)}$, $S_{XY} = S_{XY}^{(0)} + S_{XY}^{(1)}$, and

$$\hat{Z}_i^{(k)} = Y_i^{(k)} - \hat{\beta}' X_i^{(k)}, \quad i=1, 2, \dots, n_k \quad (7)$$

be the residuals. We define the sample mean $\bar{\mathbf{Z}}$ and covariance $\mathbf{S}_{zz}^{(k)}$ of $\{\widehat{\mathbf{Z}}_i^{(k)}\}$ as

$$\bar{\mathbf{Z}}^{(k)} = \bar{\mathbf{Y}}^{(k)} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \bar{\mathbf{X}}^{(k)}, \quad k=0, 1, \quad (8)$$

$$\mathbf{S}_{zz} = \mathbf{S}_{zz}^{(0)} + \mathbf{S}_{zz}^{(1)} = \mathbf{S}_{yy} - \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \quad (9)$$

We estimate λ (5) by

$$\widehat{\lambda} = \left(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \right)' \mathbf{S}_{zz}^{-1} \left(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \right). \quad (10)$$

Conditioning on $\widehat{\boldsymbol{\beta}}, \{\widehat{\mathbf{Z}}_i^{(k)}, i=1, 2, \dots, n_k\}$ can be regarded as i.i.d samples. Therefore the distribution of statistic $\widehat{\lambda}$ in (10) is related to a noncentral F -distribution, according to Reiser and Faraggi (Reiser and Faraggi, 1997). However, the common estimator $\widehat{\boldsymbol{\beta}}$

introduces additional correlations among $\{\widehat{\mathbf{Z}}_i^{(k)}, i=1, 2, \dots, n_k\}$, and thus $\text{var}(\bar{\mathbf{Z}}^{(k)}) \neq \sum_{zz}^{(k)}/n_k$. In the Appendix B, we prove

$$\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \sim N(\mu_z, \sum_{zz}), \quad (11)$$

where $\sum_{zz} = \frac{\sum_{zz}^{(0)}}{n_0} + \frac{\sum_{zz}^{(1)}}{n_1} + \mathbf{V}_1 + \mathbf{V}_2$, with

$$\mathbf{V}_1 = \sum_{k=0}^1 \frac{1}{n_k - 1} \left[\mu_x' \sum_{xx}^{-1} \sum_{xx}^{(k)} \sum_{xx}^{-1} \mu_x \right] \otimes \text{Var}(\mathbf{e}^{(k)}), \quad (12)$$

$$\mathbf{V}_2 = \left[\mu_x' \otimes (\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)})' \right] \text{Var} [\text{vec} \{ \mathbf{S}_{xx}^{(0)} \mathbf{S}_{xx}^{-1} \}] \left[\mu_x \otimes (\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}) \right] \quad (13)$$

and

$$\begin{aligned} & \text{Var} [\text{vec} (\mathbf{S}_{xx}^{(0)} \mathbf{S}_{xx}^{-1})] \\ & \approx \sum_{k=0}^1 \frac{n_k}{(n_k-1)^2} \left[\sum_{xx}^{-1} \otimes (\sum_{xx}^{(1-k)} \sum_{xx}^{-1}) \right] \left[(\mathbf{I}_{p^2} + \mathbf{K}_{pp}) (\sum_{xx}^{(k)} \otimes \sum_{xx}^{(k)}) \right] \left[\sum_{xx}^{-1} \otimes (\sum_{xx}^{-1} \sum_{xx}^{(1-k)}) \right], \quad (14) \end{aligned}$$

where vec means vector operator, $\otimes$ is the Kronecker product, $\mathbf{I}_{p^2}$ is a $p^2 \times p^2$ identity matrix, $\mathbf{K}_{pp}$ is a $p^2 \times p^2$ commutation matrix. We can estimate $\mathbf{V}_1, \mathbf{V}_2$ by substituting sample means, sample variances, mean squared residual, and $\widehat{\boldsymbol{\beta}}^{(k)}$ for μ 's, $\mathbf{S}$'s, $\text{var}(\mathbf{e})$ and $\boldsymbol{\beta}^{(k)}$ in (13) and (14), and denote them by $\widehat{\mathbf{V}}_1, \widehat{\mathbf{V}}_2$. Let

$$\widehat{\sum}_{zz} = \frac{\widehat{\sum}_{zz}^{(0)}}{n_0} + \frac{\widehat{\sum}_{zz}^{(1)}}{n_1} + \widehat{\mathbf{V}}_1 + \widehat{\mathbf{V}}_2, \quad (15)$$

where $\widehat{\sum}_{zz}^{(k)} = MSE^{(k)} + (\widehat{\beta}^{(k)} - \widehat{\beta})' S_{xx}^{(k)} (\widehat{\beta}^{(k)} - \widehat{\beta})$, and $MSE^{(k)}$ is the mean squared error of observed residuals. In addition, from the Appendix C, we can derive the distribution of $\widehat{\lambda}$ as

$$\frac{\widehat{f} - p - q + 1}{(\widehat{f} - p)\widehat{v}} \widehat{c}\widehat{\lambda} \sim F\left(\widehat{v}, \widehat{f} - p - q + 1, \frac{\widehat{f}}{\widehat{f} - p} \widehat{c}\widehat{\lambda}\right), \quad (16)$$

where

$$\widehat{f} = \frac{\left[\text{tr}\left\{ (S^{(0)} + S^{(1)})^2 \right\} + \left\{ \text{tr}(S^{(0)} + S^{(1)}) \right\}^2 \right]}{\left[\text{tr}\left\{ (S^{(0)})^2 \right\} + \left\{ \text{tr}(S^{(0)}) \right\}^2 \right] / (n_0 - 1) + \left[\text{tr}\left\{ (S^{(1)})^2 \right\} + \left\{ \text{tr}(S^{(1)}) \right\}^2 \right] / (n_1 - 1)},$$

$$\widehat{c} = \frac{\text{tr}(S_{zz}^{-1} \widehat{\sum}_{zz}) + 2 \left(\overline{\overline{Z}}^{(1)} \overline{\overline{Z}}^{(0)} \right)' S_{zz}^1 \left(\overline{\overline{Z}}^{(1)} \overline{\overline{Z}}^{(0)} \right)}{\text{tr}\left[(S_{zz}^{-1} \widehat{\sum}_{zz})^2 \right] + 2 \left(\overline{\overline{Z}}^{(1)} \overline{\overline{Z}}^{(0)} \right)' S_{zz}^1 \widehat{\sum}_{zz} S_{zz}^{-1} \left(\overline{\overline{Z}}^{(1)} \overline{\overline{Z}}^{(0)} \right)}.$$

$$\widehat{v} = \widehat{c} \cdot \text{tr}(S_{zz}^{-1} \widehat{\sum}_{zz})$$

It's easy to get $\widehat{v} = 1$ when only adding one variable Y is added over X .

In the special case when $\lambda=0$, i.e., there is no improvement from Y beyond X , $(\widehat{f} - p - q + 1) / [(\widehat{f} - p) \widehat{v}] \widehat{c}\widehat{\lambda}$ follows central F distribution with degrees of freedom $\widehat{v}$ and $\widehat{f} - p - q + 1$.

4 Hypothesis Test

Statistical inference of whether adding new variables (diagnostic test 1) is superior to existing variables (test 2) is based on testing a one-sided hypothesis about the difference in their AUCs, i.e., $H_0: \theta_1 - \theta_2 \leq \Delta$ versus $H_1: \theta_1 - \theta_2 > \Delta$, where $\Delta \geq 0$ is a prespecified superiority margin and θ_i 's are the corresponding AUCs. A conventional test statistic is

$$st = \frac{\widehat{\theta}_1 - \widehat{\theta}_2 - \Delta}{\sqrt{\text{var}(\widehat{\theta}_1 - \widehat{\theta}_2)}}, \quad (17)$$

which asymptotically follows a standard normal distribution. The key of this conventional approach is to derive the unbiased estimates of AUCs $\widehat{\theta}_1$ and $\widehat{\theta}_2$ respectively for the best combinations of vectors X and $(X', Y')'$. A naïve approach to use test statistic (17) is to use the nonparametric formula by DeLong, DeLong and Clarke-Pearson (DeLong, DeLong and Clarke-Pearson, 1988) to estimate difference in AUCs and its variance based on the two linear discriminant function scores for combinations of markers. When the diagnostic procedures follow bivariate normal distributions, there is also a parametric approach based on the parameters from normal distributions (Metz, 1986). Both approaches ignore uncertainties introduced in estimating linear discriminant functions, and the overly optimistic estimates of AUCs because of the same data being used for fitting and evaluating models. Although Janes, Longton and Pepe (Janes, Longton and Pepe, 2009) proposed use

of cross-validation (CV) to correct for the over optimism, they used naïve estimates from Stata in their examples to compare the ROC curves for the two linear prediction models for all subjects that still utilized same data for fitting and testing.

Alternative to direct comparisons of AUCs, Liu et al (Liu, et al, 2006) proposed inverse normal transformation of AUC and evaluate difference $\Phi^{-1}(\hat{\theta}_1) - \Phi^{-1}(\hat{\theta}_2)$ to assess the equivalence or non-inferiority of two diagnostic variables. The difference reflects the standardized mean shift between two ROC models. Liu and colleagues demonstrated improvement in statistical consistency and efficiency by using the standardized difference in comparison to AUCs. Our λ in (5) is a natural generalization from their standardized difference to multivariate linear combinations of $L(X)$ and $L(X, Y)$. The hypotheses to be tested are

$$H_0: \lambda \leq \delta \text{ versus } H_1: \lambda > \delta \quad (18)$$

where the margin δ can be determined from AUC difference Δ by the relationship of

$$\delta = [\Phi^{-1}(\theta_{L(X)} + \Delta)]^2 - [\Phi^{-1}(\theta_{L(X)})]^2.$$

Using $\hat{\lambda}$ defined in (10) and distribution property of (16), we reject the null hypothesis for a too large $(\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda}$. That is to reject null $H_0: \lambda \leq \delta$ at significance level α if

$$\frac{(\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda}}{(\hat{v}, \hat{f} - p - q + 1, \hat{f}/(\hat{f} - p)\hat{c}\delta)} > F_{1-\alpha}^{(-1)} \quad (19)$$

where $F_{\alpha}^{(-1)}(df_1, df_2, ncp)$ denotes the 100α quintile of the F distribution F ; df_1, df_2, ncp and the p-value is $1 - F((\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda}; \hat{v}, \hat{f} - p - q + 1, \hat{f}/(\hat{f} - p)\hat{c}\delta)$. The statistical power is

$$P\left\{\frac{(\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda}}{(\hat{v}, \hat{f} - p - q + 1, \hat{f}/(\hat{f} - p)\hat{c}\delta)} > F_{1-\alpha}^{(-1)} \mid \lambda > \delta\right\} \\ = 1 - F\left\{F_{1-\alpha}^{(-1)}\left(\hat{v}, \hat{f} - p - q + 1, \frac{\hat{f}}{\hat{f} - p}\hat{c}\delta\right); \hat{v}, \hat{f} - p - q + 1, \frac{\hat{f}}{\hat{f} - p}\hat{c}\lambda \mid \lambda > \delta\right\} \quad (20)$$

It is easy to see that the power depends on the degrees of freedom or sample size, superiority margin δ , and the true amount of improvement λ . Thus, for a given type I error rate α , expected power $1 - \beta$, λ under the alternative hypothesis δ , and covariance matrices, we can get the degrees of freedom $\hat{f}$ and corresponding required sample size.

When superiority margin $\delta \rightarrow 0$, the superiority test is simplified as the conventional hypothesis test for $H_0: \lambda = 0$ versus $H_1: \lambda > 0$. Test statistics $(\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda}$ is central F ; $\hat{v}, \hat{f} - p - q + 1, 0$ distributed under $H_0: \lambda = 0$. Therefore combining Y with X presents an improvement beyond X alone when

$$(\hat{f} - p - q + 1)/[(\hat{f} - p)\hat{v}]\hat{c}\hat{\lambda} > F_{1-\alpha}^{(-1)}(\hat{v}, \hat{f} - p - q + 1, 0).$$

5 An Example of Application

Between September 1986 and October 1988, 9,704 ambulatory white women aged 65 years or older without bilateral hip replacements were recruited from population-based listings in four United States cities in the Study of Osteoporotic Fractures (SOF, <http://sof.ucsf.edu/>)

Interface/). At the second visit (1988-89), 7,784 women (82% of the survivors at that time) had bone mineral density (BMD) measurements of the posterior-anterior (PA) spine (SBMD) and proximal femoral neck (NBMD), trochanteric (TRBMD), intertrochanteric (INTRBMD), and total hip (TBMD) using dual x-ray absorptiometry (DXA) scanners, and at the calcaneus, distal radius (DBMD), and proximal radius (PBMD) using single photon absorptiometry (SPA) scanners. Besides BMD, additional information, including age, weight, height, and body mass index (BMI), was also collected at all visits. The women were followed for hip fracture by letter or telephone every 4 months for an average of 4.1 years with 99 percent completion. Reported hip fractures were confirmed by a radiologist from preoperative radiographs. Details of the study design and summary statistics of SOF have been published previously (Cummings et al., 1993; Cummings et al., 1995) and available online.

To apply our method to the SOF data, we examined whether TBMD measured by a hip DXA scan or PBMD measured by a SXA forearm scan or their combination can provide incremental diagnostic information for 10-year hip fracture risk beyond BMI and height. The case group consists of women who developed hip fracture during 10 years follow-up. The control group consists of women who had been followed for more than 10 years without hip fracture. To illustrate our method, we select a random subset of 1275 women with a 1:2 age-matched cases and controls. Table 1 summarizes our data samples.

To improve the normality fit, we made a natural log transformation of BMI ($\ln\text{BMI}$). Quantile-quantile plots for TBMD, PBMD, $\ln\text{BMI}$, and height fit normal distributions well. Based on (Su and Liu, 1993), the OLC is the linear discriminant functions (LDF). For $X = (\ln\text{BMI}, \text{Height})$, LDF is $-2.139 \times \ln\text{BMI} - 0.0008 \times \text{height}$, with $\text{AUC}=0.560$. For $(X, Y) = (\ln\text{BMI}, \text{height}, \text{TBMD})$, LDF is $1.580 \times \ln\text{BMI} + 0.016 \times \text{height} - 3.600 \times \text{TBMD}$, with $\text{AUC}=0.706$. For $(X, Y) = (\ln\text{BMI}, \text{height}, \text{PBMD})$ LDF is $-1.519 \times \ln\text{BMI} - 0.004 \times \text{height} - 1.105 \times \text{PBMD}$, with $\text{AUC}=0.584$. The ROC curves for the OLCs are shown in Figure 1.

Suppose the superiority margin $\Delta=0.05$ for AUC difference, the corresponding margin δ for λ is

$$\delta = [\Phi^{-1}(0.560 + \Delta)]^2 - [\Phi^{-1}(0.560)]^2 = 0.055.$$

Table 2 gives the estimates $\hat{\lambda}$, $\hat{c}$, $\hat{f}$ for adding TBMD or PBMD. The degree of freedom from numerator may be different from the number of added diagnostic variables, although it is always 1 from adding only 1 variable. For the given margin $\delta=0.055$, the one-sided p-values are <0.0001 for $\hat{\lambda}(\text{TBMD}|\ln\text{BMI}, \text{height})$, 0.978 for $\hat{\lambda}(\text{PBMD}|\ln\text{BMI}, \text{height})$, and <0.0001 for $\hat{\lambda}(\text{TBMD}, \text{PBMD}|\ln\text{BMI}, \text{height})$ based on equation (16). So we reject $H_0: \lambda \leq \delta$ for adding TBMD but accept H_0 for PBMD. Therefore, we conclude that at the 0.05 significance level (1) the addition of TBMD by hip DXA scan significantly increase diagnostic accuracy from using only $\ln\text{BMI}$ and height; (2) the addition of PBMD by SXA scan doesn't significantly improve diagnostic accuracy from using only $\ln\text{BMI}$ and height. Furthermore, by testing $\hat{\lambda}(\text{TBMD}|\text{PBMD}, \ln\text{BMI}, \text{height})$ directly, we compared $\hat{\lambda}(\text{TBMD}, \text{PBMD}|\ln\text{BMI}, \text{height}) = 0.2882$ versus $\hat{\lambda}(\text{PBMD}|\ln\text{BMI}, \text{height})$ according to Theorem 2.2 and showed significant benefit of including TBMD on top of $\ln\text{BMI}$, height, and PBMD.

Using formula (20), we can calculate sample size needed in planning a study. For example, to use a superiority margin $\delta=0.147$, $\alpha=0.05$, and $\lambda=0.271$, we will need 248 cases and 496 controls (in 1:2 ratio) or 347 cases and 347 controls (in 1:1 ratio) to assure 80% statistical power of the study.

6 Simulation Studies

In simulation studies under Gaussian assumption, we specify mean vectors and covariance matrices based on the values in the SOF sample. Let $\mathbf{X} = (\ln\text{BMI}, \text{height})'$; $k = 1$ for hip fracture within 10 years and $k = 0$ otherwise; and $(Y_1, Y_2)' = (\text{TBMD}, \text{PBMD})'$.

Accordingly, the means and covariance matrices for $(\ln\text{BMI}, \text{height}, \text{TBMD}, \text{PBMD})'$ are $\boldsymbol{\mu}^{(0)} = (1.179, 158.612, 0.755, 0.629)'$, $\boldsymbol{\mu}^{(1)} = (1.169, 158.604, 0.668, 0.602)'$, and

$$\mathbf{C}^{(0)} = \begin{pmatrix} 0.002 & -0.029 & 0.003 & 0.001 \\ -0.029 & 33.236 & 0.157 & 0.141 \\ 0.003 & 0.157 & 0.016 & 0.008 \\ 0.001 & 0.141 & 0.008 & 0.011 \end{pmatrix}, \quad \mathbf{C}^{(1)} = \begin{pmatrix} 0.003 & 0.003 & 0.003 & 0.002 \\ 0.003 & 42.135 & 0.160 & 0.164 \\ 0.003 & 0.160 & 0.011 & 0.006 \\ 0.002 & 0.164 & 0.006 & 0.011 \end{pmatrix}$$

for controls and cases, respectively. In addition, we add a $Y_3 = \ln\text{BMI} + \varepsilon$, where ε is Gaussian noise $N(0, 4 \text{ Var}(\ln\text{BMI}))$. So Y_3 is conditional independent with the fracture outcome for given $\mathbf{X}$. Therefore there is no benefit adding Y_3 over $\mathbf{X}$. Let $\mathbf{Y} = (Y_1, Y_2, Y_3)'$. Because the equality of population covariance matrices affects the distribution of the estimator, we simulate two scenarios:

Experiment 6.1: $\boldsymbol{\Sigma}^{(0)} = \mathbf{C}^{(0)}$, $\boldsymbol{\Sigma}^{(1)} = \mathbf{C}^{(1)}$ and $\sum_{zz}^{(0)} \neq \sum_{zz}^{(1)}$ and

Experiment 6.2: $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}^{(1)} = (\mathbf{C}^{(0)} + \mathbf{C}^{(1)})/2$ and $\sum_{zz}^{(0)} = \sum_{zz}^{(1)}$.

Simulations are run by generating observations from a multivariate normal distribution. For each scenario, we reduplicated 10,000 times. Each replication generates $n = n_0 + n_1$ observations of cases (n_1) and controls (n_0). The proportion of replications rejecting H_0 is used to estimate the probability of type I error or power when $\alpha = 0.05$ and superiority margin δ are specified. When we specify $\delta = 0.1 \lambda(Y_1 | X_1, X_2)$ and $n_0 : n_1 = 1$ in Experiment 6.1, to achieve 80% power, $n = 104$ subjects are needed to test the additive utility Y_1 over X_1, X_2 , for our proposed method. Thus we choose $n = 120$ as the largest sample size in simulation studies.

Firstly, we evaluate possible estimation bias of effect size for our quadratic index and estimation of AUC difference using non-parametric AUC estimators (DeLong et al., 1988). One of non-parametric estimations is derived from naïve procedure where we fit OLCs and then assess the non-parametric AUCs from the predicted values for all subjects in the same dataset. Another non-parametric estimation is derived via cross-validation (CV) technique (Janes, Longton and Pepe, 2009), where different datasets are used in OLCs constructing and OLCs predicting. In a K -fold cross-validation, the original subset is partitioned into K subsets with almost equal sample size. Leave out one subset by turn, we construct OLCs from the remaining $K - 1$ subsets and applied the fitted linear functions OLCs to the leave-out subset and then get the predicted values for the subjects in the leave-out subset. Repeat the process K times, we'll get the predicted values for all subjects, and then derive the corresponding non-parametric AUC estimations. Tables 3 reports the bias of the estimations of the three approaches for utility of (Y_1, Y_2) over (X_1, X_2) when $(n_0, n_1) = (60, 60)$ under Experiment 6.1. Here, for CV procedure, we applied 60-fold CV to the dataset with one case and one control in each subset. The bias is defined as the difference between sample mean estimation versus the theoretical mean value in 10,000 runs. According to F distribution, the bias for $\hat{\lambda}$ is calculated by

$$\bar{\lambda} - E(\hat{\lambda}) = \bar{\lambda} - \frac{(\hat{f} - p)\hat{v} + \hat{f}\hat{c}\lambda}{\hat{c}(\hat{f} - p - q - 1)}.$$

The bias in AUC difference was asymptotically normal and estimated by $\bar{\hat{\theta}} - \theta$. The relative bias is defined as bias relative to the corresponding true value.

We find the non-parametric estimations of the AUCs from both naïve procedure and CV procedure to be biased. The naïve procedure over-estimates, whereas 60-fold CV procedure under-estimates the AUCs of the OLCs, i.e. 60-fold CV procedure over corrects the overoptimism from naïve procedure. Both under-estimates the additive performance of (Y_1, Y_2) over (X_1, X_2) . For example, when $n_0 = n_1 = 60$ in Experiment 6.1, the mean AUC difference by naïve estimation from 10,000 runs is 0.1441, whereas the true difference is 0.1516. The agreement in λ is substantially better. Figure 2 shows histograms for the λ -based and naïve non-parametric θ -based test statistics for direct comparison of AUC difference (17) (DeLong et al., 1988) from 10,000 runs to test the additive utility of (Y_1, Y_2) over (X_1, X_2) when $n_0 = n_1 = 60$ in Experiment 6.1. Even after bias correction, the empirical type I error rate is still about half of 0.05. Our simulations show the naïve non-parametric statistic is positive skewed from a standard normal distribution when sample size is not big enough. Because of the bias or departure from null distribution, the AUC-based method from naïve procedure or CV procedure will not maintain proper the I error rate, and will not get correct power either. So next, we only present empirical type I error rate and power for the newly proposed method.

Tables 4 and 5 report the empirical type I error rate and power to test the incremental utilities of adding $Y_1, Y_2, Y_3, (Y_1, Y_2)', (Y_1, Y_2, Y_3)', (Y_1, Y_2, Y_3)'$ and $(Y_1, Y_2, Y_3)'$ over $X = (X_1, X_2)'$ under Gaussian assumption with experimental conditions 6.1 and 6.2 with sample sizes $(n_0, n_1) = (30, 30), (40, 20), (60, 60)$, and $(80, 40)$. They demonstrate accurate type I error rates as well as practically acceptable statistical power of the newly proposed method for assessing the added diagnostic utility. The simulations show when the variable Y_3 with no benefit over X is added to Y_1 or Y_2 , the power decreased as expected.

7 Discussion

The AUC under an ROC curve measures diagnostic accuracy for discriminating diseased from nondiseased subjects. In a binormal ROC model, the difference in AUC can be more accurately and efficiently compared via the mean shift relative to the variance (Liu et al., 2006). The incremental diagnostic accuracy by adding new variables has to be evaluated based on the increase in the AUC of ROC curves (Pepe et al., 2004; Wang et al., 2006; Ware, 2006). Parametric approaches used the binormal distribution models (Metz, 1986) or non-parametric estimation by DeLong et al (DeLong et al., 1988) can evaluate differences in AUCs from randomized clinical trials if one group diagnosed with X only and another group uses the combination of X and Y . However, in a paired design commonly used in diagnostic trials, we need to use the data first to fit OLCs based on the linear discriminant functions of X and the combination of X and Y . Theorem 2.1 proves that the gain in AUCs is a function of regression residuals of Y on X , and the regression coefficient estimator $\hat{\beta}$ introduces non-linear correlated in AUC differences based on OLCs. If we use the predicted OLC values directly in evaluation of AUC difference, we will result in biased estimation of AUC gain and variance in practical sample sizes. The resulting statistical tests will be biased. Fitting and evaluating models on the same data is known to produce overly optimistic estimates of model performance. CV procedure will correct the overoptimism but also have a risk of underestimation the true difference under practical sample sizes. Our simulations show 60-

fold CV procedure over corrects the overoptimism for the specific simulation data. Our proposed index is a function of transformed AUC differences and has desired statistical consistency and efficiency demonstrated in our simulation studies. Pencina et al. (Pencina et al., 2008) introduced two alternative measures to assess the clinical benefits of adding new diagnostic variables, one based on integrated sensitivity and specificity, and the other on event-specific reclassification tables. Further research is needed to evaluate these metrics.

We formulated the incremental diagnostic utility of adding new variables $\mathbf{Y} = (Y_1, Y_2, \dots, Y_q)'$ to the existing diagnostic variables $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ under a multivariate normal distributional assumption by quadratic form $\lambda = (\mu_z^{(1)} - \mu_z^{(0)})' (\sum_{zz}^{(0)} + \sum_{zz}^{(1)})^{-1} (\mu_z^{(1)} - \mu_z^{(0)})'$ with $\mathbf{Z} = \mathbf{Y} - \sum_{yx} \sum_{xx}^{-1} \mathbf{X}$, where λ reflects a quadratic form of mean shift relative to the variance from $\mathbf{Z}$. We account the uncertainties introduced in estimating linear discriminant functions into the distribution of $\mathbf{Z}$. The inference based on $\hat{\lambda}$ estimates fewer parameters than non-parametric approaches to compare AUCs of $L(\mathbf{X})$, $L(\mathbf{X}, \mathbf{Y})$, which require estimates of variance and covariance of the AUCs. In addition, it has no ceiling (1) nor floor (0.5) effect in this index. As with the one-dimensional case (Liu et al., 2006), the λ -based method maintains the correct type I error rate and is more efficient than AUC-based methods.

The distribution estimator $\hat{\lambda}$ is related to the multivariate Behrens–Fisher problem. In deriving the distribution, we did not limit ourselves to special conditions, such as $\sum_{zz}^{(0)} = \sum_{zz}^{(1)}$. For special conditions of homogenous covariance matrixes ($\sum_{zz}^{(0)} = \sum_{zz}^{(1)}$), there are more efficient estimators, such as the conventional Hotelling T^2 test. Our simulations cover both conditions of with and without homogeneous covariance, and the results show that the test statistics (19) maintain the proper type I error rate under both conditions. It can also be used to evaluate sample size needed for such studies.

One major limitation of current approach is the assumption of multivariate normal distributions of diagnostic variables. Because the most important derivations relate to the residual $\mathbf{Z}$, it should be possible to relax the multivariate normal assumptions by requiring multivariate normal distribution assumption only to the residual $\mathbf{Z}$. Further studies are necessary to confirm this conjecture. If we fail to make proper variable transformation to fit a multivariate normal distribution, we can always use the bootstrap method to derive the empirical p-value. The bootstrap method, however, will not help the sample size calculation when we plan a study.

In summary, we derived a new test statistic to efficiently determine the statistical significance of an incremental AUC under a multivariate normal assumption. Our test links the difference in AUC to a quadratic form of standardized mean shift relative to the covariance after a properly linear transformation of all diagnostic variables. The distribution of the estimator of the quadratic form is related to the multivariate Behrens–Fisher problem. We provide an explicit mathematical solution of the estimator and its approximate noncentral F-distribution, type I error rate, and sample size formula. Our simulation studies demonstrate that our new test maintains the prespecified type I error rates as well as satisfactory statistical power under practical sample sizes. In addition, our simulation study suggests that the proposed estimator and test statistics maintain the correct type I. We use an example of predicting osteoporotic hip fracture to illustrate our method.

Acknowledgments

This work is supported by National Institutes of Health R01EB 004079 to the University of California, San Francisco (P.I. Lu). The authors thank Mr. Phil Chu and Dr. John Kornak for their helpful comments. They also

thank for the associate editor and anonymous reviewer for their constructive suggestions that significantly improved the paper.

Appendix A

Proof of Theorem 2.1

We give two lemmas first.

Lemma 1

If $\tilde{\mathbf{X}}$ is a nonsingular linear transformation of $\mathbf{X}$, the maximum AUC for the linear combination of $\tilde{\mathbf{X}}$ is the same as that of $\mathbf{X}$, i.e. $\theta_{L(\tilde{\mathbf{X}})} = \theta_{L(\mathbf{X})}$.

The lemma is obvious because $\tilde{\mathbf{X}}$ and $\mathbf{X}$ are linear combinations of each other. Let

$$\mu^{(k)} = \begin{pmatrix} \mu_X^{(k)} \\ \mu_Y^{(k)} \end{pmatrix}, \quad \Sigma^{(k)} = \begin{pmatrix} \Sigma_{XX}^{(k)} & \Sigma_{XY}^{(k)} \\ \Sigma_{YX}^{(k)} & \Sigma_{YY}^{(k)} \end{pmatrix}, \text{ and } \mu = \mu^{(1)} - \mu^{(0)}, \Sigma = \Sigma^{(1)} + \Sigma^{(0)}.$$

Lemma 2

If $(\mathbf{X}'^{(k)}, \mathbf{Y}'^{(k)})' : N(\mu^{(k)}, \Sigma^{(k)})$, $\mathbf{Z}^{(k)} = \mathbf{Y}^{(k)} - \Sigma_{YX}^{-1} \Sigma_{XX} \mathbf{X}^{(k)}$ with $\Sigma_{YX} = \Sigma_{YX}^{(1)} + \Sigma_{YX}^{(0)}$ and $\Sigma_{XX} = \Sigma_{XX}^{(1)} + \Sigma_{XX}^{(0)}$, then

$$\begin{pmatrix} \mathbf{X}^{(k)} \\ \mathbf{Z}^{(k)} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_X^{(k)} \\ \mu_Z^{(k)} \end{pmatrix}, \begin{pmatrix} \Sigma_{XX}^{(k)} & \Sigma_{XZ}^{(k)} \\ \Sigma_{ZX}^{(k)} & \Sigma_{ZZ}^{(k)} \end{pmatrix} \right)$$

with $\Sigma_{XZ} = \Sigma_{XZ}^{(0)} + \Sigma_{XZ}^{(1)} = 0$, and $\Sigma_{ZZ} = \Sigma_{ZZ}^{(0)} + \Sigma_{ZZ}^{(1)} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$.
Hence

$$\begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{XX}^{-1} & 0 \\ 0 & \Sigma_{ZZ}^{-1} \end{pmatrix}.$$

Because $(\mathbf{X}', \mathbf{Z}')'$ is a linear transformation of $(\mathbf{X}', \mathbf{Y}')'$, according to the above Lemmas 1 and 2, we proved Theorem 2.1.

Appendix B

Moments of $\bar{\bar{\mathbf{Z}}}^{(1)} - \bar{\bar{\mathbf{Z}}}^{(0)}$

Because

$$\bar{\bar{\mathbf{Z}}}^{(1)} - \bar{\bar{\mathbf{Z}}}^{(0)} = \bar{\mathbf{Y}}^{(1)} - \bar{\mathbf{Y}}^{(0)} - \hat{\beta} (\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)}),$$

with

$$\hat{\beta} = S_{XX}^{-1} S_{XY} = S_{XX}^{-1} [S_{XX}^{(0)} \hat{\beta}^{(0)} + S_{XX}^{(1)} \hat{\beta}^{(1)}],$$

$$E(\widehat{\beta}|\mathbf{X}) = \mathbf{S}_{XX}^{-1} [\mathbf{S}_{XX}^{(0)} \beta^{(0)} + \mathbf{S}_{XX}^{(1)} \beta^{(1)}].$$

$\widehat{\beta}$ is an asymptotic unbiased estimator of β since $\mathbf{S}$ is an unbiased estimator of Σ , and $\mathbf{S}^{-1}$ is an asymptotic unbiased estimator of Σ^{-1} . So $\widehat{\Sigma}^{(1)} - \widehat{\Sigma}^{(0)}$ is a consistent estimator of $\mu_Z^{(1)} - \mu_Z^{(0)}$.

Note that $\beta^{(k)} = \left(\sum_{XX}^{(k)} \right)^{-1} \sum_{XY}^{(k)}$ is the regression coefficient for $Y^{(k)} = \beta_0^{(k)} + (\beta^{(k)})' X^{(k)} + e^{(k)}$, where $\beta_0^{(k)} = \mu_Y^{(0)} - (\beta^{(k)})' \mu_X^{(0)}$, and the error term $e^{(k)}$ follows a q -dimensional normal distribution with mean vector 0 and $e^{(k)} \perp X^{(k)}$.

$$\text{Var}[\text{vec}(\widehat{\beta})] = E[\text{Var}(\text{vec}(\widehat{\beta})|\mathbf{X})] + \text{Var}[E(\text{vec}(\widehat{\beta})|\mathbf{X})], \quad (\text{A.1})$$

where, vec means vector operator, and

$$\begin{aligned} & \text{Var}[\text{vec}(\widehat{\beta})|\mathbf{X}] \\ &= \sum_{k=0}^1 \left(\mathbf{S}_{XX}^{-1} \mathbf{S}_{YX}^{(k)} \otimes \mathbf{I}_q \right) \left\{ \text{Var}[\text{vec}(\widehat{\beta}^{(k)})|\mathbf{X}] \right\} \left(\mathbf{S}_{YX}^{(k)} \mathbf{S}_{XX}^{-1} \otimes \mathbf{I}_q \right) \\ &= \sum_{k=0}^1 \frac{1}{n_k-1} \left(\mathbf{S}_{XX}^{-1} \mathbf{S}_{YX}^{(k)} \otimes \mathbf{I}_q \right) \left\{ [\mathbf{S}_{XX}^{(k)}]^{-1} \otimes \text{Var}(e^{(k)}) \right\} \left(\mathbf{S}_{YX}^{(k)} \mathbf{S}_{XX}^{-1} \otimes \mathbf{I}_q \right) \quad (\text{A.2}) \\ &= \sum_{k=0}^1 \frac{1}{n_k-1} \left(\mathbf{S}_{XX}^{-1} \mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right) \otimes \text{Var}(e^{(k)}), \end{aligned}$$

$$\begin{aligned} & \text{Var}[E\{\text{vec}(\widehat{\beta})|\mathbf{X}\}] \\ &= \text{Var}[E\{\text{vec}(\beta^{(k)} - \widehat{\beta}|\mathbf{X})\}] \\ &= \text{Var}\left[\text{vec}\left\{\beta^{(k)} - \left(\beta^{(0)} \mathbf{S}_{XX}^{(0)} + \beta^{(1)} \mathbf{S}_{XX}^{(1)}\right) \mathbf{S}_{XX}^{-1}\right\}\right] \quad (\text{A.3}) \\ &= \text{Var}\left[\text{vec}\left\{\left(\beta^{(k)} - \beta^{(1-k)}\right)' \mathbf{S}_{XX}^{(1-k)} \mathbf{S}_{XX}^{-1}\right\}\right] \\ &= \left[\mathbf{I}_p \otimes (\beta^{(1)} - \beta^{(0)})' \right] \text{Var}\left[\text{vec}\left(\mathbf{S}_{XX}^{(1-k)} \mathbf{S}_{XX}^{-1}\right)\right] \left[\mathbf{I}_p \otimes (\beta^{(1)} - \beta^{(0)}) \right]. \end{aligned}$$

$\text{Var}[\text{vec}(\mathbf{S}_{XX}^{(1-k)} \mathbf{S}_{XX}^{-1})]$ can be derived from the delta method. Note that

$$\begin{aligned} \frac{\partial[\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1}]}{\partial \mathbf{S}_{XX}^{(k)}} &= \mathbf{S}_{XX}^{-1} \otimes \mathbf{I}_p - \mathbf{S}_{XX}^{-1} \otimes (\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1}) = \mathbf{S}_{XX}^{-1} \otimes (\mathbf{S}_{XX}^{(1-k)} \mathbf{S}_{XX}^{-1}), \\ \frac{\partial[\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1}]}{\partial \mathbf{S}_{XX}^{(1-k)}} &= -\mathbf{S}_{XX}^{-1} \otimes (\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1}). \end{aligned}$$

Adopting the first-order Taylor approximation,

$$\begin{aligned}
\text{Var} \left[\text{vec} \left(\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right) \right] &= \text{Var} \left[\text{vec} \left\{ \mathbf{S}_{XX}^{(k)} \left(\mathbf{S}_{XX}^{(0)} + \mathbf{S}_{XX}^{(1)} \right)^{-1} \right\} \right] \\
&\approx \frac{\partial \left[\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right]}{\partial \mathbf{S}_{XX}^{(k)}} \Big|_{S=\Sigma} \text{Var} \left[\text{vec}(\mathbf{S}_{XX}^{(k)}) \right] \left[\frac{\partial \left(\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right)}{\partial \mathbf{S}_{XX}^{(k)}} \Big|_{S=\Sigma} \right]' + \\
&\quad \frac{\partial \left[\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right]}{\partial \mathbf{S}_{XX}^{(1-k)}} \Big|_{S=\Sigma} \text{Var} \left[\text{vec}(\mathbf{S}_{XX}^{(1-k)}) \right] \left[\frac{\partial \left(\mathbf{S}_{XX}^{(k)} \mathbf{S}_{XX}^{-1} \right)}{\partial \mathbf{S}_{XX}^{(1-k)}} \Big|_{S=\Sigma} \right]' \\
&= \sum_{l=0}^1 \frac{n_l}{(n_l-1)^2} \left[\Sigma_{XX}^{-1} \otimes \left(\Sigma_{XX}^{(1-l)} \Sigma_{XX}^{-1} \right) \right] \left[\left(\mathbf{I}_{p^2} + \mathbf{K}_{pp} \right) \left(\Sigma_{XX}^{(l)} \otimes \Sigma_{XX}^{(l)} \right) \right] \left[\Sigma_{XX}^{-1} \otimes \left(\Sigma_{XX}^{-1} \Sigma_{XX}^{(1-l)} \right) \right],
\end{aligned}$$

where $\mathbf{K}_{pp}$ is denoted the $p^2 \times p^2$ commutation matrix, defined as

$$\mathbf{K}_{pp} = \sum_{i,j=1}^p (\mathbf{H}_{ij} \otimes \mathbf{H}_{ij}'),$$

where $\mathbf{H}_{ij}$ denote the $p \times p$ matrix with the element $h_{ij}=1$ and all other elements equal to 0. Note that $\hat{\beta} \perp (\bar{\mathbf{Y}}^{(1)}, \bar{\mathbf{Y}}^{(0)}, \bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(0)})$, hence

$$\begin{aligned}
&\text{Var} \left(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \right) \\
&= \text{Var} \left[\text{E} \left(\bar{\mathbf{Y}}^{(1)} - \bar{\mathbf{Y}}^{(0)} - \hat{\beta} \left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)} \right) \right) \hat{\beta} \right] + \text{E} \left[\text{Var} \left(\bar{\mathbf{Y}}^{(1)} - \bar{\mathbf{Y}}^{(0)} - \hat{\beta} \left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)} \right) \right) \hat{\beta}_c \right] \\
&= \text{Var} \left(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \right) + \text{E} \left[\left\{ \left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)} \right) \otimes \mathbf{I}_q \right\} \text{Var} \left\{ \text{vec}(\hat{\beta}) \right\} \left(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(0)} \right) \otimes \mathbf{I}_q \right] \\
&\approx \frac{\Sigma_{ZZ}^{(0)}}{n_0} + \frac{\Sigma_{ZZ}^{(1)}}{n_1} + \left\{ \left(\mu_X^{(1)} - \mu_X^{(0)} \right)' \otimes \mathbf{I}_q \right\} \text{Var} \left[\text{vec}(\hat{\beta}) \right] \left\{ \left(\mu_X^{(1)} - \mu_X^{(0)} \right) \otimes \mathbf{I}_q \right\}.
\end{aligned}$$

Here, we only omit a term $O(\min(n_0, n_1)^{-2})$ in the above approximation. Plugging in $\text{Var}[\text{vec}(\hat{\beta})]$ from (A.1)-(A.3), we get

$$\text{Var} \left(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)} \right) = \frac{\Sigma_{ZZ}^{(0)}}{n_0} + \frac{\Sigma_{ZZ}^{(1)}}{n_1} + V_1 + V_2, \quad (\text{A.4})$$

where

$$V_1 = \sum_{k=0}^1 \frac{1}{n_k - 1} \left[\mu_X' \Sigma_{XX}^{-1} \sum_{XX}^{(k)} \sum_{XX}^{-1} \mu_X \right] \otimes \text{Var}(e^{(k)}),$$

$$V_2 = \left[\mu_X' \otimes (\beta^{(1)} - \beta^{(0)})' \right] \text{Var} \left[\text{vec} \left\{ \mathbf{S}_{XX}^{(0)} \mathbf{S}_{XX}^{-1} \right\} \right] \left[\mu_X \otimes (\beta^{(1)} - \beta^{(0)}) \right].$$

Appendix C

Distribution of Estimator $\hat{\lambda}$

Since

$$S_{ZZ} = S_{ZZ}^{(0)} + S_{ZZ}^{(1)} = S_{YY} - S_{YX} S_{XX}^{-1} S_{XY}.$$

Let

$$S = \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix} = S^{(0)} + S^{(1)}, \quad \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix} = \Sigma^{(0)} + \Sigma^{(1)},$$

where

$$S^{(k)} = \begin{pmatrix} S_{XX}^{(k)} & S_{XY}^{(k)} \\ S_{YX}^{(k)} & S_{YY}^{(k)} \end{pmatrix}, \quad \Sigma^{(k)} = \begin{pmatrix} \Sigma_{XX}^{(k)} & \Sigma_{XY}^{(k)} \\ \Sigma_{YX}^{(k)} & \Sigma_{YY}^{(k)} \end{pmatrix} \quad k=0, 1,$$

The distribution of S is related to the multivariate Behrens-Fisher problem.

$$S = S^{(0)} + S^{(1)} \sim \text{Wishart} \left(n_0 - 1, \frac{\Sigma^{(0)}}{n_0 - 1} \right) + \text{Wishart} \left(n_1 - 1, \frac{\Sigma^{(1)}}{n_1 - 1} \right),$$

Following (Nel and Merwe, 1986), approximately,

$$S \sim \text{Wishart} \left(f, \frac{\Sigma}{f} \right),$$

where

$$f = \frac{\text{tr} \{ (\Sigma)^2 \} + \{ \text{tr} (\Sigma) \}^2}{\left[\text{tr} \{ (\Sigma^{(0)})^2 \} + \{ \text{tr} (\Sigma^{(0)}) \}^2 \right] / (n_0 - 1) + \left[\text{tr} \{ (\Sigma^{(1)})^2 \} + \{ \text{tr} (\Sigma^{(1)}) \}^2 \right] / (n_1 - 1)},$$

where “tr” means trace. According to the theorem from Muirhead p. 93-97 (Muirhead, 1982),

$$S_{ZZ} = S_{YY} - S_{YX} S_{XX}^{-1} S_{XY} \sim \text{Wishart} \left(f - p, \frac{\Sigma_{ZZ}}{f} \right).$$

So $E(S_{ZZ}) = (f - p) \Sigma_{ZZ} / f$. The variance of $S_{ZZ}^{(k)}$ tends to 0 as $n^{(k)}$ approaches infinity. So S_{ZZ} converges in the probability to the value of $(f - p) \Sigma_{ZZ} / f$. On the basis of the convergence theorem, the asymptotic distribution of $\hat{\lambda}$ is the same as the quadratic statistics $f / (f - p) \Lambda_1$ for large sample size, where

$$\Lambda_1 = \left(\bar{\bar{Z}}^{(1)} - \bar{\bar{Z}}^{(0)} \right)' \Sigma_{ZZ}^{-1} \left(\bar{\bar{Z}}^{(1)} - \bar{\bar{Z}}^{(0)} \right).$$

Many exact and approximate methods are available to calculate the distribution of quadratic forms in normal variables, including Patnaik's two-moment central χ^2 -approximation (Box, 1954; Patnaik, 1949), Pearson's three-moment central χ^2 approximation etc. (Imhof, 1961; Pearson, 1959). All these estimations were derived either based on an assumption of equal means for $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(0)}$ (thus zero differences) or simply approximate noncentral χ^2 by a central χ^2 approximation. They are inappropriate to our problem because we want to test the quadratic statistics above a positive non-zero constant. Instead, we use a two-moment noncentral χ^2 approximation to calculate the distribution of Λ_1 . We approximate

$$c^* \Lambda_1 \approx \chi^2(v, c^* \lambda)$$

with degrees of freedom v and a noncentral parameter $c^* \lambda$, where c^* , v are derived from the two-moment equations

$$c^* \left[\text{tr}(\Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ}) + \mu_Z' \Sigma_{ZZ}^{-1} \mu \right] = \left(v + c^* \mu_Z' \Sigma_{ZZ}^{-1} \mu \right)$$

and

$$2\text{tr}[(\Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ})^2] + 4\mu_Z' \Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ} \Sigma_{ZZ}^{-1} \mu = \frac{1}{(c^*)^2} (2v + 4c^* \mu_Z' \Sigma_{ZZ}^{-1} \mu),$$

where $\tilde{\Sigma}_{ZZ} = \text{Var}(\bar{\mathbf{Z}}^{(1)} - \bar{\mathbf{Z}}^{(0)}) = \frac{\Sigma_{ZZ}^{(0)}}{n_0} + \frac{\Sigma_{ZZ}^{(1)}}{n_1} + \mathbf{V}_1 + \mathbf{V}_2$, Therefore we get

$$c^* = \frac{\text{tr}(\Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ}) + 2\mu_Z' \Sigma_{ZZ}^{-1} \mu}{\text{tr}[(\Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ})^2] + 2\mu_Z' \Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ} \Sigma_{ZZ}^{-1} \mu},$$

and

$$v = c^* \cdot \text{tr}(\Sigma_{ZZ}^{-1} \tilde{\Sigma}_{ZZ})$$

In the special case of $\Sigma_{ZZ} = I\Sigma_{ZZ}$ for a positive constant I , which is always true for one-dimensional $q = 1$, $c^* = I$, $v = q$, $I\Lambda_1 = \chi^2(q, I\lambda)$.

So, when n_0, n_1 are large enough, $f/(f-p)\Lambda_1$ can directly estimate λ with approximate distribution

$$c^* \Lambda_1 \dot{\sim} \chi^2(v, c^* \lambda).$$

For small sample size, however, we cannot simply replace Σ_{ZZ} by $f(f-p)S_{ZZ}$. Let

$$\Lambda_2 = \frac{\left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)' \Sigma_{ZZ}^{-1} \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)}{\left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)' \mathbf{S}_{ZZ}^{-1} \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)}.$$

Then $\hat{\lambda} = \Lambda_1/\Lambda_2$ is a Hotelling- T^2 type statistic. Although $\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)}$ and $\mathbf{S}_{ZZ}$ may correlate, the correlation between is small as $O(\min(n_0, n_1)^{-1})$ and we can ignore it. According to the theorem from (Muirhead, 1982), $f\Lambda_2$ follows a central χ^2 -distribution with degrees of freedom $f - p - q + 1$. Therefore

$$\frac{f - p - q + 1}{fv} c^* \hat{\lambda} = \frac{c^* \Lambda_1 / v}{f\Lambda_2 / (f - p - q + 1)} \sim F(v, f - p - q + 1, c^* \lambda), \quad (\text{A.5})$$

with degrees of freedom v , $f - p - q + 1$ and noncentrality parameter $c^* \lambda$.

We can estimate f , c^* , and v by substituting $\mathbf{S}^{(k)}$, $f(f - p)\mathbf{S}_{ZZ}$, and $\widehat{\sum_{ZZ}}$ for $\mathbf{S}^{(k)}$, $\mathbf{S}_{ZZ}$ and Σ_{ZZ} ,

$$\widehat{f} = \frac{\left[\text{tr} \left\{ (\mathbf{S}^{(0)} + \mathbf{S}^{(1)})^2 \right\} + \left\{ \text{tr} (\mathbf{S}^{(0)} + \mathbf{S}^{(1)}) \right\}^2 \right]}{\left[\text{tr} \left\{ (\mathbf{S}^{(0)})^2 \right\} + \left\{ \text{tr} (\mathbf{S}^{(0)}) \right\}^2 \right] / (n_0 - 1) + \left[\text{tr} \left\{ (\mathbf{S}^{(1)})^2 \right\} + \left\{ \text{tr} (\mathbf{S}^{(1)}) \right\}^2 \right] / (n_1 - 1)}$$

$$\widehat{c^*} = \frac{\widehat{f}}{\widehat{f} - p} \widehat{c},$$

$$\widehat{v} = \widehat{c} \cdot \text{tr}(\mathbf{S}_{ZZ}^{-1} \widehat{\sum_{ZZ}}),$$

with

$$\widehat{c} = \frac{\text{tr}(\mathbf{S}_{ZZ}^{-1} \widehat{\sum_{ZZ}}) + 2 \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)' \mathbf{S}_{ZZ}^{-1} \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)}{\text{tr}[(\mathbf{S}_{ZZ}^{-1} \widehat{\sum_{ZZ}})^2] + 2 \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)' \mathbf{S}_{ZZ}^{-1} \widehat{\sum_{ZZ}} \mathbf{S}_{ZZ}^{-1} \left(\overline{\mathbf{Z}}^{(1)} - \overline{\mathbf{Z}}^{(0)} \right)}.$$

Then (A.5) can be written as

$$\frac{\widehat{f} - p - q + 1}{(\widehat{f} - p)\widehat{v}} \widehat{c} \hat{\lambda} \sim F \left(\widehat{v}, \widehat{f} - p - q + 1, \frac{\widehat{f}}{\widehat{f} - p} \widehat{c} \lambda \right). \quad (\text{A.6})$$

References

- Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*. 1954; 25:290–302.
- Cai T, Cheng S. Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics*. 2008; 9:216–233. [PubMed: 18056687]
- Cummings SR, Black DM, Nevitt MC, et al. Bone density at various sites for prediction of hip fractures. The Study of Osteoporotic Fractures Research Group. *Lancet*. 1993; 341:72–75. [PubMed: 8093403]
- Cummings SR, Nevitt MC, Browner WS, et al. Risk factors for hip fracture in white women. Study of Osteoporotic Fractures Research Group. *N Engl J Med*. 1995; 332:767–773. [PubMed: 7862179]
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
- Imhof JP. Computing the distribution of quadratic forms in normal variables. *Biometrika*. 1961; 48:419–426.
- Janes H, Longton GM, Pepe MS. Accommodating Covariates in ROC Analysis. *The Stata Journal*. 2009; 9:17–39. [PubMed: 20046933]
- Jin H, Lu Y. A procedure for determining whether a simple combination of diagnostic tests may be noninferior to the theoretical optimum combination. *Med Decis Making*. 2008; 28:909–916. [PubMed: 18556633]
- Jin H, Lu Y. Permutation test for non-inferiority of the linear to the optimal combination of multiple tests. *Statistics & Probability Letters*. 2009a; 79:664–669. [PubMed: 20161260]
- Jin H, Lu Y. The optimal linear combination of multiple predictors under the generalized linear models. *Statistics and Probability Letters*. 2009b; 79:2321–2327. [PubMed: 20160887]
- Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Stat Med*. 2005; 24:37–47. [PubMed: 15515132]
- Liu JP, Ma MC, Wu CY, Tai JY. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Stat Med*. 2006; 25:1219–1238. [PubMed: 16158400]
- McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics*. 2002; 58:657–664. [PubMed: 12230001]
- Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology*. 1986; 21:720–733. [PubMed: 3095258]
- Muirhead, RJ. *Aspects of Multivariate Statistical Theory*. New York: John Wiley & Sons, Inc; 1982. Aspects of Multivariate Statistical Theory; p. 93-97.
- Nel DG, Merwe Vd. A solution to the multivariate Behrens-Fisher problem. *Communications in Statistics-- Theory Methods*. 1986; 15:3719–3735.
- Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A*. 1933; 24:289–337.
- Patnaik PB. The non-central χ^2 - and F-distributions and their applications. *Biometrika*. 1949; 36:202–232. [PubMed: 18132092]
- Pearson ES. Note on an approximation to the distribution of non-central χ^2 . *Biometrika*. 1959; 46:364.
- Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27:157–172. discussion 207-112. [PubMed: 17569110]
- Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*. 2006; 62:221–229. [PubMed: 16542249]
- Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004; 159:882–890. [PubMed: 15105181]

- Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics*. 2000; 1:123–140. [PubMed: 12933515]
- Reiser B, Faraggi D. Confidence intervals for the generalized ROC criterion. *Biometrics*. 1997; 53:644–652. [PubMed: 9235120]
- Schisterman EF, Faraggi D, Reiser B. Adjusting the generalized ROC curve for covariates. *Stat Med*. 2004; 23:3319–3331. [PubMed: 15490426]
- Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical association*. 1993; 88:1350–1355.
- Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006; 355:2631–2639. [PubMed: 17182988]
- Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006; 355:2615–2617. [PubMed: 17182986]

Fig. 1.
ROC curves for OLCs for (1) lnBMI, height (dotted line); (2) lnBMI, height, TBMD (solid line); (3) lnBMI, height, PBMD (dashed line).

Fig. 2.

Histograms of λ -based test statistic and naïve non-parametric θ -based test statistic from 10,000 runs to test additive utility of (Y_1, Y_2) over (X_1, X_2) when $n_0=n_1=60$ in Experiment 6.1.

(1) Histograms of quadratic-form-based test statistic.

Noncentral distribution under null hypothesis (solid curve); 95% quantile of the null distribution(dashed line).

(2) Histograms of classic non-parametric AUC-based test statistic.

Standard normal distribution under null hypothesis (solid curve); 5% and 95% quantiles of standard normal distribution (dashed line); Bias corrected normal distribution under null hypothesis (dotted curve); 5% and 95% quantiles of bias corrected normal distribution (dotted line).*

Table 1Summary statistics of BMI, height, TBMD and PBMD (Mean $\pm$ SD)

	Controls (n ₀ =850)	Fractured Subjects (n ₁ =425)
BMI	26.269 $\pm$ 4.232	25.431 $\pm$ 4.345
Height	158.612 $\pm$ 5.765	158.604 $\pm$ 6.491
TBMD	0.755 $\pm$ 0.127	0.668 $\pm$ 0.106
PBMD	0.629 $\pm$ 0.103	0.602 $\pm$ 0.103

Table 2

The estimations $\hat{\lambda}$, $\hat{c}$, $\hat{f}$, and p values for the SOF Example for $X=(\ln\text{BMI,height})$

	TBMD X	PBMD X	(TBMD,PBMD) X	TBMD (PBMD, X)
$\hat{\lambda}$	0.2710	0.0224	0.2882	0.2658
$\hat{c}$	597.110	560.732	594.803	578.422
$\hat{v}$	1	1	2.076	1
$\hat{f}$	1035.129	1035.096	1035.144	1035.144
$P(\Delta=0.05)^a$	0.0000	0.9780	0.0000	0.0000
$P(\Delta=0.10)$	0.0007	1.0000	0.0002	0.0012

^aSuperiority margin $\Delta=0.05$ ($\delta=0.055$) or $\Delta=0.10$ ($\delta=0.147$) for superiority test

Table 3

Bias and standard error (SE) of estimations for utility of (Y_1, Y_2) over (X_1, X_2) in Experiment 6.1 when $(n_0, n_1) = (60, 60)$ in 10,000 Runs

	Naive AUC Bias $\pm$ SE (relative bias)	CV-based AUC Bias $\pm$ SE (relative bias)	Quadratic-form-based λ Bias $\pm$ SE (relative bias)
(X_1, X_2) ($\theta=0.5599$)	0.0255 $\pm$ 0.000427 (4.55%) [*]	-0.0312 $\pm$ 0.000512 (-5.58%)	
(X_1, X_2, Y_1, Y_2) ($\theta=0.7114$)	0.0180 $\pm$ 0.000442 (2.53%)	-0.0646 $\pm$ 0.000563 (-9.09%)	
(Y_1, Y_2) over (X_1, X_2) ($\Delta\theta=0.1516$, $\lambda=0.2882$)	-0.0075 $\pm$ 0.000515 (-4.93%)	-0.0334 $\pm$ 0.000564 (-22.05%)	0.0007 $\pm$ 0.001292 (0.25%)

^{*} Relative bias is bias divided by the true value expressed in percentage.

Table 4
Empirical type I error rates and power for Experiment 6.1 under Gaussian distribution (10,000 Runs)

Margin δ	$Y_3 X\lambda=0$	$Y_1 X\lambda=0.271$	$(Y_1, Y_3) X\lambda=0.271$	$Y_2 X\lambda=0.022$	$(Y_2, Y_3) X\lambda=0.022$	$(Y_1, Y_2) X\lambda=0.288$	$(Y_1, Y_2, Y_3) X\lambda=0.288$
$(n_0, n_1)=(30, 30)$							
$\delta=\lambda$	0.0448	0.0436	0.0460	0.0430	0.0453	0.0437	0.0458
$\delta=0.5\lambda$	---	0.1769	0.1726	0.0683	0.0621	0.1753	0.1684
$\delta=0.1\lambda$	---	0.5638	0.5153	0.1023	0.0845	0.5222	0.4873
$(n_0, n_1)=(40, 20)$							
$\delta=\lambda$	0.0481	0.0430	0.0433	0.0439	0.0445	0.0443	0.0415
$\delta=0.5\lambda$	---	0.1734	0.1538	0.0662	0.0595	0.1647	0.1495
$\delta=0.1\lambda$	---	0.5362	0.4579	0.0958	0.0769	0.4817	0.4245
$(n_0, n_1)=(60, 60)$							
$\delta=\lambda$	0.0473	0.0493	0.0472	0.0483	0.0492	0.0493	0.0475
$\delta=0.5\lambda$	---	0.2897	0.2846	0.0918	0.0853	0.2966	0.2882
$\delta=0.1\lambda$	---	0.8446	0.8145	0.1695	0.1296	0.8307	0.8069
$(n_0, n_1)=(80, 40)$							
$\delta=\lambda$	0.0497	0.0442	0.0472	0.0479	0.0510	0.0458	0.0460
$\delta=0.5\lambda$	---	0.2765	0.2701	0.0836	0.0834	0.2751	0.2710
$\delta=0.1\lambda$	---	0.8182	0.7798	0.1533	0.1241	0.8034	0.7665

\$watermark-text

\$watermark-text

\$watermark-text

Table 5
Empirical type I error rates and power for Experiment 6.2 under Gaussian distribution (10,000 Runs)

Margin δ	$Y_3 X \lambda=0$	$Y_1 X \lambda=0.271$	$(Y_1, Y_3) X \lambda=0.271$	$Y_2 X \lambda=0.022$	$(Y_2, Y_3) X \lambda=0.022$	$(Y_1, Y_2) X \lambda=0.288$	$(Y_1, Y_2, Y_3) X \lambda=0.288$
$(n_0, n_1)=(30, 30)$							
$\delta=\lambda$	0.0461	0.0460	0.0456	0.0445	0.0463	0.0457	0.0447
$\delta=0.5\lambda$	---	0.1817	0.1719	0.0701	0.0625	0.1763	0.1694
$\delta=0.1\lambda$	---	0.5732	0.5044	0.1091	0.0826	0.5291	0.4780
$(n_0, n_1)=(40, 20)$							
$\delta=\lambda$	0.0444	0.0485	0.0434	0.0470	0.0431	0.0461	0.0456
$\delta=0.5\lambda$	---	0.1704	0.1584	0.0702	0.0615	0.1656	0.1535
$\delta=0.1\lambda$	---	0.5110	0.4591	0.0999	0.0777	0.4709	0.4277
$(n_0, n_1)=(60, 60)$							
$\delta=\lambda$	0.0474	0.0479	0.0425	0.0511	0.0469	0.0466	0.0445
$\delta=0.5\lambda$	---	0.2893	0.2772	0.0935	0.0806	0.2966	0.2780
$\delta=0.1\lambda$	---	0.8367	0.8113	0.1746	0.1312	0.8299	0.8100
$(n_0, n_1)=(80, 40)$							
$\delta=\lambda$	0.0497	0.0444	0.0493	0.0488	0.0488	0.0454	0.0489
$\delta=0.5\lambda$	---	0.2643	0.2642	0.0856	0.0779	0.2626	0.2650
$\delta=0.1\lambda$	---	0.7997	0.7555	0.1561	0.1267	0.7845	0.7459