Evaluations and comparisons of treatment effects based on best combinations of biomarkers with applications to biomedical studies

Albert Vexler[*], Xiwei Chen and Jihnhee Yu

Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA

*avexler@buffalo.edu

ABSTRACT

Many clinical and biomedical studies evaluate treatment effects based on multiple biomarkers that commonly consist of pre- and post-treatment measurements. Some biomarkers can show significant positive treatment effects while other biomarkers can reflect no effects or even negative effects of the treatments, giving rise to a necessity to develop methodologies that may correctly and efficiently evaluate the treatment effects based on multiple biomarkers as a whole. In the setting of pre- and post-treatment measurements of multiple biomarkers, we propose to apply a receiver operating characteristic (ROC) curve methodology based on the best combination of biomarkers maximizing the AUC-type criterion among all possible linear combinations. In the particular case with independent pre- and post-treatment measurements, we show that the proposed method represents the well-known Su and Liu's (1993) result. Further, proceeding from derived best combinations of biomarkers' measurements, we propose an efficient technique via likelihood ratio tests to compare treatment effects. We show an extensive Monte Carlo study that confirms the superiority of the proposed test in comparison of treatment effects based on multiple biomarkers in a paired data setting. For practical applications, the proposed method is illustrated with a randomized trial of chlorhexidine gluconate on oral bacterial pathogens in mechanically ventilated patients as well as a treatment study for children with ADHD and severe mood dysregulation (SMD).

Keywords: Area under curve (AUC); Best linear combination; Likelihood ratio test; Paired data; Receiver operating characteristic (ROC); Treatment effect.

1. Introduction

Biomarkers have been important tools in disease diagnosis, drug development and research. In the area of drug development, biomarkers' measurements can be applied to reflect drug effects, and thus are often used to compare different treatment groups. Biomarkers can show treatment effects in different magnitudes or even different directions, necessitating methodologies to examine the treatment effects based on multiple biomarkers jointly. Many studies compare treatment effects based on multiple biomarkers' measurements of independent case group and control group. This paper targets to propose methodologies that can correctly and efficiently evaluate the treatment effects based on pre- and post-treatment measurements of multiple biomarkers as a whole, and to further develop an efficient statistical testing methodology to compare independent treatment groups with paired data. One of the motivating examples in this paper is as follows. The chlorhexidine gluconate on oral bacterial pathogens study was conducted on patients admitted to the 18-bed trauma intensive care unit (TICU) of the Erie County Medical Center (ECMC) where patients were mechanically ventilated. These patients were of particular interest since they have a high risk of ventilator-associated pneumonia. While it is true that these patients are ill and thus may be more susceptible to infection, they also have the greatest need for prevention of infection. A randomized, double-blind, and placebo-controlled clinical trial tested oral topical 0.12% chlorhexidine gluconate (treatment group) and placebo with vehicle alone (control group), applied twice a day by staff nurses. Quantitation of colonization of the oral cavity by respiratory pathogens on left teeth and right teeth was measured. The aim of the study was to determine the best regimen of oral hygiene in the TICU to reduce oral colonization by potential respiratory bacterial pathogens (PRPs). In this paper, we propose to combine the oral plaque quantification on left teeth and right teeth to maximize an AUC-type quantity based on pre- and post-treatment observations in the evaluation of the treatment effect on oral bacterial pathogens in mechanically ventilated patients.

For biomarkers whose values are measured on a continuous scale, its diagnostic performance in identifying diseased subjects is commonly assessed via receiver operating characteristic (ROC) curves, e.g., Pepe (2006) and Vexler (2008). Suppose values of a biomarker from the diseased population (X) and the healthy population (Y) are independent and identically distributed samples from two different distributions with cumulative distribution functions $F(\cdot)$ and $G(\cdot)$, respectively. A ROC curve plots sensitivity (true positive rate, $1 - F(t)$) versus one minus specificity (true negative rate, $1 - G(t)$) for various values of the threshold t. The mathematical formula is $ROC(t) = 1 - F(G^{-1}(t))$, where $t \in [0, 1]$. The area under the ROC curve (AUC) is a common index of the diagnostic performance of a biomarker. Bamber (1975) noted that the area under this curve is equal to $Pr(X > Y)$.

Some recent biostatistical literature (e.g., Tian, 2008; Tian et al., 2012; Hauck et al., 2000) proposes to consider the quantity $Pr(X > Y)$ in the context of a generalized treatment effect, when X and Y denote continuous outcome variables for treatment arm and control arm, respectively. Hauck *et al.* (200) introduced the use of $Pr(X > Y)$ in clinical trials as a statistical measurement of describing treatment effects, namely, the generalized treatment effect, and derived a method for confidence interval estimation of $Pr(X > Y)$ with normally distributed outcomes. Tian (2008) compared large sample approach, a generalized variable approach and a bootstrap approach for confidences interval estimation of generalized treatment effects in linear models. Tian *et al.* (2012) proposed to utilize the generalized variable method for testing equality of generalized treatment effects.

The standard ROC methodology as well as generalized treatment effects mentioned above is commonly considered with respect to case-control studies. In the case of independent populations, e.g., cases and controls, various approaches have been proposed to evaluate and compare the performance of bivariate and/or multivariate biomarkers. McClish (1987) and DeLong *et al.* (1988) proposed comparisons of diagnostic biomarkers based on the difference of areas under ROC curves. Wieand *et al.* (1989) proposed statistics for comparisons of ROC curves based on a weighted average of sensitivities. Considering the combination of multiple biomarkers as a single composite score, Pepe and Thompson (2000), as well as Vexler *et al.* (2006) have considered empirical solutions to the optimal linear combinations of biomarkers in the context of nonparametric maximizations of corresponding AUCs.

Su and Liu (1993) derived the optimal linear combinations yielding the largest area under the ROC curves if the values of the biomarkers in the diseased (case) and the non-diseased (control) population both follow multivariate normal distributions. We will extend to consider the generalized treatment effect of optimally combined biomarkers in a more general situation with paired data (X and Y are correlated). In this paper, we consider the best linear combination of pre- and post-treatment measurements of biomarkers in the sense that the AUC-type measures of treatment effects of this combination is maximized among all possible linear combinations. In a particular case, when pre- and post-treatment biomarkers' measurements are independent, the proposed method corresponds to the well-addressed result of Su and Liu's (1993).

Additionally in this paper, to compare effects of treatments between two independent groups based on pre- and post-treatment measurements of groups of biomarkers, we propose a test statistic using the concept of the efficient maximum likelihood ratio methodology, which carries out group comparisons of AUC-type measures of the optimal linear combination of biomarkers.

Primarily, the proposed approach is applied to a randomized trial of chlorhexidine gluconate on oral bacterial pathogens in mechanically ventilated patients. Also, we demonstrate an excellent applicability of the proposed method to any relevant multiple outcomes beyond biomarker studies via a treatment study for children with ADHD and severe mood dysregulation (SMD). ADHD is the most commonly diagnosed behavioral disorder of childhood. Most children with ADHD also have at least one other developmental or behavioral problem. They may also have a psychiatric problem, such as depression or bipolar disorder. Severe mood dysregulation is a syndrome defined to capture the symptomatology of children whose diagnostic status with respect to bipolar disorder is uncertain, that is, those who have severe, nonepisodic irritability and the hyperarousal symptoms characteristic of mania but who lack the well-demarcated periods of elevated or irritable mood characteristic of bipolar disorder. For each child enrolled in the study, Depression Rating Scale (CDRS) scores and Young Mania Rating Scale (YMRS) scores were taken at the baseline and the endpoint. The objective of the study was to compare total treatment effects based on pre- and post-treatment measurements of CDRS-R and YMRS between the experimental group-based therapy program and the community psychosocial treatment (i.e., control). For more related research in this context, see Vexler *et al.* (2012). In this paper, we propose to combine the measured values maximizing an AUC-type quantity based on pre- and post-treatment observations in evaluation of treatment effects in the study for children with ADHD and SMD.

This paper is organized as follows. In Section 2, we define the AUC-type measure and the estimation of the best linear combination of biomarkers. The maximum likelihood ratio test is proposed in Section 2 as well. Section 3 shows an extensive Monte Carlo study for the proposed methods. Section 4 illustrates applications to a randomized trial of chlorhexidine gluconate on oral bacterial pathogens in mechanically ventilated patients as well as a treatment study for children with ADHD and severe mood dysregulation (SMD). In Section 5, we conclude the article with remarks.

2. Methods

When distributions of two independent populations, say, case and control, are compared based on measurements of multiple biomarkers, it is desirable to combine the measurements of different biomarkers (e.g., Su and Liu, 1993), since markers usually represent different aspects of diseases. Using combined scores of biomarkers can increase the diagnostic accuracy of the set of medical tests. Commonly, biomarkers' values are proposed to be combined with respect to the maximization of AUCs (e.g., Vexler *et al.*, 2006; Liu *et al.*, 2011). In this paper, we derive best linear combinations of pre- and post-treatment measurements of biomarkers. The likelihood ratio test is used to compare two treatment groups (e.g., case and control) based on the AUC-type criterion computed with respect to the best linear combinations of biomarkers' values.

2.1 Best linear combination

Without loss of generality and with respect to our practical examples, suppose two biomarkers involved in a study to analyze treatment effects. Let X_{1i} and X_{2i} be the pre- and post-treatment measurements of one biomarker, respectively, with respect to the i-th $(i = 1, \dots, n)$ patient for a certain treatment. Let Y_{1i} and Y_{2i} be the pre- and post-treatment measurements of another biomarker, respectively, with respect to the i-th $(i = 1, \dots, n)$ patient. Assume that $(X_1, X_2, Y_1, Y_2)^T$ (here T stands for the transpose operation) follows a multivariate normal distribution with the mean vector $\boldsymbol{\mu} = (\mu_{X_1}, \mu_{X_2}, \mu_{Y_1}, \mu_{Y_2})^T$ and the covariance matrix $\Sigma = (\sigma_{hl}), 1 \leq h \leq 4, 1 \leq l \leq 4$. To represent a simple measure of treatment effects, we are interested in reducing dimensionality by constructing an effective linear combination of biomarkers with values X s and Y s. This implies that we derive certain optimal linear coefficients (λ_1, λ_2) so that for groups of markers' values (X_1, Y_1) and (X_2, Y_2), the one-dimensional random variables $U_1 = \lambda_1 X_1 + \lambda_2 Y_1$ and $U_2 = \lambda_1 X_2 + \lambda_2 Y_2$ can be presented. This linear combination of measurements of biomarkers dominates all the other possible linear combinations in the sense that it provides a maximum of the AUC-type measure $Pr(U_1 < U_2)$ for all λ_1 and λ_2. Thus, the optimal $\boldsymbol{\lambda^o} = (\lambda_1^o, \lambda_2^o)^T$ maximizes the AUC-type measure, denoted by A, where

$$A(\lambda_1,\, \lambda_2) = Pr(\lambda_1 X_1 + \lambda_2 Y_1 < \lambda_1 X_2 + \lambda_2 Y_2) = Pr(\lambda_1 X_1 - \lambda_1 X_2 + \lambda_2 Y_1 - \lambda_2 Y_2 < 0)$$

over all possible values of λ_1 and λ_2, i.e., $(\lambda_1^o, \lambda_2^o) = \arg\max_{\lambda_1,\, \lambda_2} A(\lambda_1, \lambda_2)$.

Under the assumption of multivariate normality of the biomarkers' measurements distribution, $(\lambda_1, -\lambda_1, \lambda_2, -\lambda_2)(X_1, X_2, Y_1, Y_2)^T$ follows the normal distribution with mean $\lambda_1 \Delta\mu_X + \lambda_2 \Delta\mu_Y$ and variance $\delta_1 \lambda_1^2 + 2\delta_2 \lambda_1 \lambda_2 + \delta_3 \lambda_2^2$, where

$$\Delta\mu_X = \mu_{X_1} - \mu_{X_2}\,, \Delta\mu_Y = \mu_{Y_1} - \mu_{Y_2},$$

and

$$\delta_1 = \sigma_{11} + \sigma_{22} - 2\sigma_{12}, \delta_2 = \sigma_{13} - \sigma_{14} - \sigma_{23} + \sigma_{24}, \delta_3 = \sigma_{33} + \sigma_{44} - 2\sigma_{34}.$$

Then, the corresponding AUC-type measure has the form of

$$\Phi\left(-\frac{\lambda_1 \Delta\mu_X + \lambda_2 \Delta\mu_Y}{\sqrt{\delta_1 \lambda_1^2 + 2\delta_2 \lambda_1 \lambda_2 + \delta_3 \lambda_2^2}}\right), \tag{1}$$

where Φ is a standard normal cumulative distribution function. The best linear combination can be defined by maximizing the AUC-type measure, and obtaining values of $(\lambda_1^o, \lambda_2^o)$ shown in the following proposition.

Proposition 2.1.1. The best linear combination coefficients $(\lambda_1^o, \lambda_2^o)$ are proportional to

$$(\Delta\mu_X, \Delta\mu_Y)\begin{pmatrix} -\delta_3 & \delta_2 \\ \delta_2 & -\delta_1 \end{pmatrix} = (-\Delta\mu_X\delta_3 + \Delta\mu_Y\delta_2, \Delta\mu_X\delta_2 - \Delta\mu_Y\delta_1).$$

The proof is shown in the Appendix.

Given the best linear combination derived in Proposition 2.1.1, the maximized AUC-type measure has the form of

$$\Phi\left(\frac{\delta_3\Delta\mu_X^2 + \delta_1\Delta\mu_Y^2 - 2\delta_2\Delta\mu_X\Delta\mu_Y}{\sqrt{\delta_1{\delta_3}^2\Delta\mu_X^2 + \delta_3{\delta_1}^2\Delta\mu_Y^2 - \delta_3{\delta_2}^2\Delta\mu_X^2 - \delta_1{\delta_2}^2\Delta\mu_Y^2 - 2\delta_1\delta_2\delta_3\Delta\mu_X\Delta\mu_Y + 2{\delta_2}^3\Delta\mu_X\Delta\mu_Y}}\right). \quad (2)$$

If biomarkers are mutually independent, that is, $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{Y} = (Y_1, Y_2)$ are independent, the best linear combination coefficients are

$$(\lambda_1^o, \lambda_2^o) \propto (\Delta\mu_X, \Delta\mu_Y)\begin{pmatrix} -\delta_3 & 0 \\ 0 & -\delta_1 \end{pmatrix},$$

that is, proportional to the weighted change in the mean vector $(-\delta_3\Delta\mu_X, -\delta_1\Delta\mu_Y)$.

In a special case of independent pre- and post-treatment measurements of biomarkers, which is an analogy to the statement of a case-control study, we have the same result as that proposed by Su and Liu (1993). The result is formalized in the following proposition.

Proposition 2.1.2. If pre- and post-treatment measurements are independent for both markers, that is, X_1 is independent of X_2, and Y_1 is independent of Y_2, the best linear combination coefficients are $(\lambda_1^o, \lambda_2^o) \propto (-\Delta\mu_X\delta_3 + \Delta\mu_Y\delta_2,\ \Delta\mu_X\delta_2 - \Delta\mu_Y\delta_1)$. Thus, $(\lambda_1^o, \lambda_2^o)^T \propto (\Sigma_{post} + \Sigma_{pre})^{-1}(\boldsymbol{\mu_{post}} - \boldsymbol{\mu_{pre}})$, where

$$\boldsymbol{\mu_{pre}} = \left(\mu_{X_1}, \mu_{Y_1}\right)^T, \boldsymbol{\mu_{post}} = \left(\mu_{X_2}, \mu_{Y_2}\right)^T, \Sigma_{pre} = \begin{pmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{pmatrix}, \Sigma_{post} = \begin{pmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{42} & \sigma_{44} \end{pmatrix}.$$

The corresponding proof is outlined in the Appendix.

Thus, we propose to use the maximized AUC-type measure in the context of the best linear combinations to depict the total treatment effects based on pre- and post-treatment measurements of biomarkers. The total treatment effect $Pr(\lambda_1X_1 + \lambda_2Y_1 < \lambda_1X_2 + \lambda_2Y_2)$ has the value of (2).

2.2 Maximum likelihood ratio tests

In this section, we propose the maximum likelihood ratio test for comparing treatments' effects based on best linear combinations of pre- and post-treatment measurements of biomarkers. To this end, we modify the technique proposed in Vexler *et al.* (2008).

Let X_{rki} represent the pre- $(r = 1)$ and post-treatment $(r = 2)$ measurements of a biomarker (X) for the i-th $(i = 1, \dots, n_k)$ patient in the k-th group, $k = 1$ for the new therapy group, and $k = 2$ for the control group, respectively. Likewise, let Y_{jki} represent the pre- $(r = 1)$ and post-treatment $(r = 2)$ measurements of another biomarker (Y) for the i-th $(i = 1, \dots, n_k)$ patient in the k-th group, $k = 1$ for the new therapy group, and $k = 2$ for the control group, respectively. Assume biomarkers' measurements for the new therapy group $\boldsymbol{v_{1i}} = (X_{11i}, X_{21i}, Y_{11i}, Y_{21i})^T$ $(i = 1, \dots, n_1)$ and biomarkers' measurements for the control group $\boldsymbol{v_{2j}} = (X_{12j}, X_{22j}, Y_{12j}, Y_{22j})^T$ $(j = 1, \dots, n_2)$ follow a multivariate normal distribution with the mean vector $\boldsymbol{\mu_k} = (\mu_{1k}, \mu_{2k}, \mu_{3k}, \mu_{4k})^T = E(\boldsymbol{v_{k1}}) = (E(X_{1k1}), E(X_{2k1}), E(Y_{1k1}), E(Y_{2k1}))^T$ and with the covariance matrix $\Sigma_k = E\left(\left(\boldsymbol{v_{k1}} - E(\boldsymbol{v_{k1}})\right)\left(\boldsymbol{v_{k1}} - E(\boldsymbol{v_{k1}})\right)^T\right) = (\sigma_{hlk}), 1 \le h \le 4, 1 \le l \le 4, k = 1, 2$. Let A_1 and A_2 denote the maximized AUC-type measures for the new therapy group and the control group, respectively. In this section, for the comparison of the treatment effects for the new therapy group and the control group based on paired observations, we formally consider testing hypothesis,

$$H_0: A_1 = A_2 \text{ vs. } H_1: A_1 \neq A_2. \tag{3}$$

In Section 2.1, we showed that the maximized AUC-type measures in both groups have the form of (1). Thus, $A_k = \Phi\left(\frac{N_k}{\sqrt{D_k}}\right)$ can be expressed as a function of $\boldsymbol{\mu_k}$ and Σ_k, where

$N_k = (\mu_{1k} - \mu_{2k})^2(\sigma_{33k} - 2\sigma_{34k} + \sigma_{44k}) + (\mu_{3k} - \mu_{4k})^2(\sigma_{11k} - 2\sigma_{12k} + \sigma_{22k}) -$

$2(\mu_{1k} - \mu_{2k})(\mu_{3k} - \mu_{4k})(\sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k})$,

$D_k = 2(\mu_{1k} - \mu_{2k})(\mu_{3k} - \mu_{4k})(\sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k})^3 - (\mu_{1k} - \mu_{2k})^2(\sigma_{33k} - 2\sigma_{34k} + \sigma_{44k})(\sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k})^2 - (\mu_{3k} - \mu_{4k})^2(\sigma_{11k} - 2\sigma_{12k} + \sigma_{22k})(\sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k})^2 + (\mu_{1k} - \mu_{2k})^2(\sigma_{11k} - 2\sigma_{12k} + \sigma_{22k})(\sigma_{33k} - 2\sigma_{34k} + \sigma_{44k})^2 + (\mu_{3k} - \mu_{4k})^2(\sigma_{11k} - 2\sigma_{12k} + \sigma_{22k})^2(\sigma_{33k} - 2\sigma_{34k} + \sigma_{44k}) - 2(\mu_{1k} - \mu_{2k})(\mu_{3k} - \mu_{4k})(\sigma_{11k} - 2\sigma_{12k} + \sigma_{22k})(\sigma_{33k} - 2\sigma_{34k} + \sigma_{44k})(\sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k})$.

Therefore the hypothesis setting (3) is equivalent to

$$H_0\colon \frac{N_1}{\sqrt{D_1}} = \frac{N_2}{\sqrt{D_2}} \text{ vs. } H_1\colon \frac{N_1}{\sqrt{D_1}} \neq \frac{N_2}{\sqrt{D_2}}. \tag{4}$$

Under the null hypothesis, μ_{11} can be represented as a function of the remaining set of parameters, say, $\mu_{11} = h(\cdot) = h(\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2)$, for a certain function h. We show the exact form of the function h in the Appendix. Thus, in a simple case, when all the parameters are known, we can utilize the classical most powerful likelihood ratio method for testing H_0. To this end, the likelihood functions under H_1 and H_0 can be presented correspondingly as

$$L_1(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2) = \prod_{i=1,\dots,n_1} \phi(\boldsymbol{v_{1i}}; \boldsymbol{\mu_1}, \Sigma_1) \prod_{j=1,\dots,n_2} \phi\big(\boldsymbol{v_{2j}}; \boldsymbol{\mu_2}, \Sigma_2\big),$$

$$L_0(\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2) = \prod_{\substack{i=1,\dots,n_1 \\ j=1,\dots,n_2}} \phi(\boldsymbol{v_{1i}}; (h, \mu_{21}, \mu_{31}, \mu_{41})^T, \Sigma_1)\, \phi\big(\boldsymbol{v_{2j}}; \boldsymbol{\mu_2}, \Sigma_2\big),$$

where $\phi(\cdot)$ denotes the multivariate normal density function known as

$$\phi(\boldsymbol{v}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-2}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{v}-\boldsymbol{\mu})}.$$

Therefore the classical likelihood ratio test-statistic is

$$\Lambda = \prod_{i=1,\dots,n_1} \frac{\phi(\boldsymbol{v_{1i}}; \boldsymbol{\mu_1}, \Sigma_1)}{\phi(\boldsymbol{v_{1i}}; (h, \mu_{21}, \mu_{31}, \mu_{41})^T, \Sigma_1)}. \tag{5}$$

When the parameters are unknown, we can apply the maximum likelihood ratio to be the test statistic

$$\Lambda = \frac{\sup_{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2} L_1(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2)}{\sup_{\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2} L_0(\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2)}$$

$$= \frac{\sup_{\boldsymbol{\mu_1}, \Sigma_1} \prod_{i=1,\dots,n_1} \phi(\boldsymbol{v_{1i}}; \boldsymbol{\mu_1}, \Sigma_1) \sup_{\boldsymbol{\mu_2}, \Sigma_2} \prod_{j=1,\dots,n_2} \phi\left(\boldsymbol{v_{2j}}; \boldsymbol{\mu_2}, \Sigma_2\right)}{\sup_{\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2} \prod_{\substack{i=1,\dots,n_1 \\ j=1,\dots,n_2}} \phi(\boldsymbol{v_{1i}}; (h, \mu_{21}, \mu_{31}, \mu_{41})^T, \Sigma_1)\, \phi\left(\boldsymbol{v_{2j}}; \boldsymbol{\mu_2}, \Sigma_2\right)}. \qquad (6)$$

The maximum likelihood estimators under H_1 have closed form solutions. The maximum log-likelihood under H_1 is

$$2(log(2\pi) + 1)(n_1 + n_2) + \frac{n_1}{2} log\left|\hat{\Sigma}_1\right| + \frac{n_2}{2} log\left|\hat{\Sigma}_2\right|, \qquad (7)$$

where

$$\hat{\Sigma}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\boldsymbol{v_{1i}} - \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{v_{1i}}\right)\left(\boldsymbol{v_{1i}} - \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{v_{1i}}\right)^T,$$

$$\hat{\Sigma}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\left(\boldsymbol{v_{2j}} - \frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{v_{2j}}\right)\left(\boldsymbol{v_{2j}} - \frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{v_{2j}}\right)^T.$$

Under H_0, in order to calculate the maximum likelihood, we carried out the numerical approach without specifying the closed forms of the estimators of the unknown parameters.

Thus, we reject the null hypothesis if $\Lambda > \Lambda_\alpha$, where the threshold Λ_α corresponds to Type I error α. Following the Wilks' Theorem (e.g., Lehmann and Romano, 1997), under H_0, the statistic $2\ln\Lambda$ asymptotically has a χ_1^2 distribution. Thus, the threshold Λ_α can be easily obtained from $Pr(\Lambda > \Lambda_\alpha) = \alpha$, as $n_1, n_2 \to \infty$. Moreover, the proposed test is asymptotically locally most powerful, e.g., see Choi et al. (1996).

Remark 1. *Numerical calculations.* Note that, applying statistical software such as R, SPlus, etc., allows us to calculate the minimization of $-\log\big(L_0(\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2)\big)$ without using closed forms of the estimators of the unknown parameters. The basic procedure “optim” in R (2012) can be carried out to minimize the negative log-likelihood under H_0 and the procedure “multiroot” helps finding this minimization. The related R codes are available from the authors upon request.

Remark 2. *Transformed normal approach.* In the case that the normal assumptions are not satisfied, for example, when biological mechanisms induce log-normal distributions of biomarkers (2001), we can fit the data to a Box-Cox power transformation model (1964) to better achieve normality of biomarkers. To be more specific, for the i-th ($i = 1, \dots, n_1$) measurement of k-th ($k = 1, 2$) group $\boldsymbol{v_{ki}} = (X_{1ki}, X_{2ki}, Y_{1ki}, Y_{2ki})$, the Box-Cox power transformed values are defined as $\boldsymbol{v_{ki}}^{(\tau,\vartheta)} = (X_{1ki}^{(\tau_{1k})}, X_{2ki}^{(\tau_{2k})}, Y_{1ki}^{(\vartheta_{1k})}, Y_{2ki}^{(\vartheta_{2k})})$, where

$$X_{lki}^{(\tau_{lk})} = \begin{cases} \frac{{X_{lki}}^{\lambda_{lk}}-1}{\lambda_{lk}}, \tau_{lk} \neq 0 \\ \log(X_{lki}), \tau_{lk} = 0 \end{cases}, \ l = 1, 2, \text{ and } Y_{lki}^{(\vartheta_{lk})} = \begin{cases} \frac{{Y_{lki}}^{\vartheta_{lk}}-1}{\vartheta_{lk}}, \vartheta_{lk} \neq 0 \\ \log(Y_{lki}), \vartheta_{lk} = 0 \end{cases}, \ l = 1, 2.$$

The power coefficients λ_{1k}, λ_{2k}, ϑ_{1k} and ϑ_{2k} can be estimated by maximizing the likelihood

$$\sup_{\boldsymbol{\mu_k}, \Sigma_k, \tau_{1k}, \tau_{2k}, \vartheta_{1k}, \vartheta_{2k}} \prod_{i=1,\dots,n_1} \phi\left(\boldsymbol{v_{ki}}^{(\tau,\vartheta)}; \boldsymbol{\mu_k}, \Sigma_k\right).$$

Then the normality-based best linear combinations of biomarkers and the maximum likelihood ratio test can be used on the transformed data.

3. Simulation study

In this section, Monte Carlo simulations are conducted to examine the power properties of the proposed tests under different scenarios. We also compare AUC-type measures between the proposed optimal combination case and only one biomarker case.

3.1 Power and Type I error

To study the power and the Type I error of the proposed test, 2,000 samples of biomarkers’ measurements for a new therapy group $(X_{11i}, X_{21i}, Y_{11i}, Y_{21i})^T$ $(i = 1, \dots, n_1)$ of sample size of n_1 were generated from multivariate normal distribution with the mean vector $\boldsymbol{\mu_1}$ and the covariance matrix Σ_1, and biomarkers' measurements for a control group $(X_{12j}, X_{22j}, Y_{12j}, Y_{22j})^T$ $(j = 1, \dots, n_2)$ of sample size of n_2 were generated from multivariate normal distribution with the mean vector $\boldsymbol{\mu_2}$ and the covariance matrix Σ_2, where

$$\boldsymbol{\mu_1} = (7.9333, \mu_{21}, 26.2000, 26.5333)^T,$$

$$\boldsymbol{\mu_2} = (7.9333, \mu_{22}, 26.2000, 26.5333)^T.$$

We consider the unequal covariance case, where

$$\Sigma_1 = \begin{pmatrix} 21.2622 & 10.7689 & 18.7467 & 13.7690 \\ 10.7689 & 20.5156 & 10.2934 & 16.4490 \\ 18.7467 & 10.2934 & 42.2935 & 32.5602 \\ 13.7690 & 16.4490 & 32.5602 & 38.1158 \end{pmatrix}, \tag{8}$$

$$\Sigma_2 = \begin{pmatrix} 23.8711 & 21.8047 & 3.7383 & 0.5820 \\ 21.8047 & 26.7344 & 3.4766 & 4.0391 \\ 3.7383 & 3.4766 & 31.5898 & 13.1211 \\ 0.5820 & 4.0391 & 13.1211 & 14.9023 \end{pmatrix}.$$

and the equal covariance case with the common covariance matrix as shown in (8). These parameters were chosen to reflect a real data example with values close to those in the treatment study for children with ADHD and severe mood dysregulation (SMD) introduced in Section 1.

TABLE 1 HERE

The values of μ_{21} and μ_{22} are shown from Table 1 to Table 4, which are chosen such that difference in the maximized AUC-type measures between two groups are set to be 0, 0.1 or 0.2, i.e., $A_1 - A_2 = 0$, 0.1 or 0.2, where A_1 and A_2 denote the maximized AUC-type measures in the context of best linear combinations of biomarkers' values for the new therapy group and the control group, respectively. We consider following scenarios: $A_1 = 0.70$, $A_2 = 0.60$; $A_1 = 0.85$, $A_2 = 0.75$; $A_1 = 0.80$, $A_2 = 0.60$ as well as $A_1 = 0.95$, $A_2 = 0.75$ with different sample sizes with $\Lambda_\alpha = 3.84$, $\alpha = 0.05$.

TABLE 2 HERE

In the same setting of parameters, Table 1 compares the Monte Carlo (MC) powers of the proposed MLR test in the context of the optimally combined two biomarkers to the powers using one biomarker alone in the equal covariance case. Table 2 depicts type I errors of the proposed MLR test with the best linear combination of two biomarkers in the equal covariance matrix case. With the same setting of parameters, Table 3 compares the Monte Carlo powers of the proposed MLR test in the context of the optimally combined two biomarkers to the powers using one biomarker alone in the unequal covariance case. Table 4 depicts type I errors of the proposed MLR test with the best linear combination of two biomarkers in the unequal covariance case.

When the difference in AUC-type measures between two groups and the sample size increases, the MLR tests provide increased powers as anticipated in both equal and unequal covariance matrix cases. Table 1 and Table 3 show that the powers of the proposed test with the best linear combinations of two biomarkers are very high when the sample size is large enough in both equal and unequal covariance cases. The power is close to be 1 when the difference in the AUC-type measures between two groups is 0.2 and the sample size in each group is 300. Compared to the power of the proposed test with optimal combinations, powers with one biomarker alone are much smaller. The type I errors of the MLR tests are well controlled even for relatively small sample sizes, say, 30 in each group.

TABLE 3 HERE

3.2 AUC-type measures

To compare AUC-type measures between the proposed optimal combination case and only one biomarker case, 20,000 samples of biomarkers' measurements $(X_1, X_2, Y_1, Y_2)^T$ of various sample size of n $(n = 30, 50, 100, 300)$ were generated from multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ.

TABLE 4 HERE

We consider following scenarios: $a.$ $\boldsymbol{\mu} = (7.9333, 11.0925, 26.2000, 26.5333)^T$ and the covariance matrix Σ as shown in (8); $b.$ $\boldsymbol{\mu} = (7.9333, 8.7008, 26.2000, 26.5333)^T$ and the covariance matrix Σ as shown in (8) with 16.6389 in the $(1, 2)$ element and $(2, 1)$ element instead. The best linear combination of measurements of biomarker X and biomarker Y is proportional to $(1, -1.5686)$ in scenario a and $(1, -1.4047)$ in scenario b. The AUC-type measure associated with the best linear combination has the form of (2), which is 0.8. The AUC-type measure for X alone corresponds to equation (1) where $\lambda_1 = 1$, and $\lambda_2 = 0$, while the AUC-type measure for Y alone corresponds to the case where $\lambda_1 = 0$, and $\lambda_2 = 1$. Table 5 shows the theoretical AUC-type measures and values based on 20,000 simulations as well as the Monte Carlo (MC) variance of the simulated AUC-type measures. In the scenario a, the AUC type measure for X alone appears to be similar to that for the best linear combinations, suggesting that Y, in fact, adds little to the discriminating capacity of X. In the scenario b, it is observed that the optimal combination provides substantially better discrimination than does X alone or Y alone. When the sample size n is large, the simulated AUC-type measures with small variability are almost exactly the theoretical values as anticipated.

TABLE 5 HERE

4. Applications to data

In this section, we exemplify the proposed method with data from two clinical studies briefly described in the introduction.

4.1 Oral colonization data

A randomized, double-blind, placebo-controlled clinical trial tested oral topical 0.12% chlorhexidine gluconate ($n_1 = 14$) or placebo ($n_2 = 11$), applied twice a day by staff nurses. The paired data were constituted by two measurements of plaque on the denture surface taken from the same subjects at the baseline (day 0) and the endpoint (day 4). The goal was to determine the best regimen of oral hygiene in the TICU based on the mean plaque quantification for the sets of left teeth (upper left first bicuspid, lower left first molar, and lower left central incisor) and right teeth (the upper right first molar, upper right central incisor, and lower right first bicuspid). For the treatment group, the best linear combination of right teeth scores and left teeth scores is proportional to (5.000, 1), leading to the maximized AUC-type measure of 0.7687 The optimized AUC-type measure is higher the AUC-type measure of 0.7677 with the right teeth scores alone and the AUC-type measure of 0.7456 with the right teeth scores alone. For the control group, the best linear combination of right teeth scores and left teeth scores is proportional to (-2.6070, 1), leading to the maximized AUC-type measure of 0.5888. The corresponding p-value of the hypothesis test of (2) is 0.042, indicating the rejection of the null hypothesis of 'lack of treatment effect' at the 0.05 significance level. The decontamination of the oral cavity with chlorhexidine improved the oral hygiene among mechanically ventilated patients in TICU, potentially indicating reduction of potential respiratory pathogens.

4.2 ADHD data

The attention deficit-hyperactivity disorder (ADHD) and severe mood dysregulation (SMD) data were produced in Center for Children and Families at University at Buffalo to examine the feasibility and efficacy of a group-based therapy program for children with ADHD and SMD. A novel group-based therapy program was studied to treat ADHD and mood problems since most ADHD treatments have not designed to help mood problems. Children ages 7 to 12 with ADHD and SMD were randomly assigned to receive either an 11 week experimental group-based therapy program for children and parents (treatment group, $n_1 = 17$), or to community psychosocial treatment (control group, $n_2 = 15$).

FIGURE 1 HERE

Clinicians rate Children's Depression Rating Scale-revised version (CDRS-R) scores and Young Mania Rating Scale (YMRS) scores. The CDRS-R consist of 17 clinician rated items, with 14 items based on the child's self-report or reports from the parents or teachers and 3 items based on the child's nonverbal behavior during the interviews. The YMRS is an 11-item, multiple-choice diagnostic questionnaire which psychiatrists use to measure the severity of manic episodes in patients. The paired data were constituted by two measurements taken from the same subjects at the baseline (week 0) and the endpoint (week 11). The objective is to compare treatment effects with respect to CDRS-R and YMRS between the treatment group and the control group. Figure 1 displays AUC-type measures with linear combinations $\boldsymbol{\lambda_1}$YMRS+$\boldsymbol{\lambda_2}$CDRS-R versus the ratio λ_1/λ_2 for $\lambda_1/\lambda_2 \in (-\infty, \infty)$ for the treatment group. For ease of presentation, the plot displays AUC-type measures versus λ_2/λ_1 when $\lambda_1/\lambda_2 > 1$ or $\lambda_1/\lambda_2 < -1$. As can be observed in this plot, the best linear combination of YMRS scores and CDRS-R scores is proportional to (0.1076, 1), leading to the maximized AUC-type measure of 0.9350. The maximized AUC-type measure is higher the AUC-type measure of 0.8449 with the YMRS scores alone ($\lambda_2/\lambda_1 = 0$) and the AUC-type measure of 0.9347 with the CDRS-R scores alone ($\lambda_1/\lambda_2 = 0$). Similarly for the control group, the best linear combination of YMRS scores and CDRS-R scores is proportional to (0.5903, 1), leading to the maximized AUC-type measure of 0.8507, which is higher than the AUC-type measure of 0.7455 with the YMRS scores alone and the AUC-type measure of 0.8156 with the CDRS-R scores alone. The corresponding p-value of the hypothesis test of (2) is 0.0085, indicating the null hypothesis of ‘lack of treatment effect’ is rejected at the 0.05 significance level. Since larger AUC values indicate better diagnostic quality. We conclude the experimental group-based therapy program is better than the community psychosocial treatment.

5. Conclusions

It is well known that the ROC curve is the most commonly used statistical tool to assess the quality of diagnostic biomarkers. In this paper, we constructed best linear combinations of biomarkers’ measurements based on correlated data maximizing the AUC-type criterion among all possible linear combinations of the biomarker values. In a special case of independent pre- and post-treatment measurements of biomarkers, we showed the same result as that proposed by Su and Liu (1993). Thus, the proposed method can be applied to both independent data as well as paired data. In the context of maximized AUC-type measure, we proposed to use maximum likelihood ratio tests to compare treatment effects based on pre- and post-treatment measurements of multiple biomarkers. Through the Monte Carlo study, the proposed methodology has been confirmed to be very efficient and the proposed test demonstrated adequate power properties corresponding to the hypotheses and sample sizes while keeping the Type I error under control even with moderate sample sizes. The superiority of the best linear combination over one biomarker alone was also verified. The analyses of a randomized trial of chlorhexidine gluconate on oral bacterial pathogens in mechanically ventilated patients as well as a treatment study for children with ADHD and severe mood dysregulation (SMD) demonstrated the fact that the proposed method is relevant to compare treatment groups with correlated multiple outcomes and easy to apply.

Appendix

A.1. Proof of Proposition 2.1.1.

To maximize the AUC-type measures, we calculate first partial derivatives of the function of AUC-type measure with respect to λ_1 and λ_2. Set $\partial A/\partial\lambda_1 = 0$, and $\partial A/\partial\lambda_2 = 0$. The equations are equivalent to

$$(\delta_3\Delta\mu_X - \delta_2\Delta\mu_Y)\lambda_2^2 - (\delta_1\Delta\mu_Y - \delta_2\Delta\mu_X)\lambda_1\lambda_2 = 0,$$

$$(\delta_1\Delta\mu_Y - \delta_2\Delta\mu_X)\lambda_1^2 - (\delta_3\Delta\mu_X - \delta_2\Delta\mu_Y)\lambda_1\lambda_2 = 0.$$

Thus,

$$\lambda_1^o = 0, \text{ or } \lambda_2^o/\lambda_1^o = (\delta_2\Delta\mu_X - \delta_1\Delta\mu_Y)/(\delta_2\Delta\mu_Y - \delta_3\Delta\mu_X), \text{ if } \lambda_1^o \neq 0,$$

$$\lambda_2^o = 0, \text{ or } \lambda_1^o/\lambda_2^o = (\delta_3\Delta\mu_X - \delta_2\Delta\mu_Y)/(\delta_1\Delta\mu_Y - \delta_2\Delta\mu_X), \text{ if } \lambda_2^o \neq 0.$$

It can be confirmed that $\partial^2 A/\partial{\lambda_1}^2 < 0$, and $\partial^2 A/\partial{\lambda_2}^2 < 0$.

Proof of Proposition 2.1.2. In the special case of independent paired data, that is, $\sigma_{12} = \sigma_{34} = \sigma_{14} = \sigma_{23} = 0$. Based on Proposition 2.1.1, we have

$$(\lambda_1^o, \lambda_2^o) \propto (-\Delta\mu_X\delta_3 + \Delta\mu_Y\delta_2, \Delta\mu_X\delta_2 - \Delta\mu_Y\delta_1).$$

By Su and Liu's result,

$$\begin{pmatrix}\lambda_1^o\\ \lambda_2^o\end{pmatrix} \propto (\Sigma_{case} + \Sigma_{control})^{-1}(\mu_{case} - \mu_{control}),$$

that is,

$$\begin{pmatrix}\lambda_1^o\\ \lambda_2^o\end{pmatrix} \propto \left(\Sigma_{post} + \Sigma_{pre}\right)^{-1}\left(\mu_{post} - \mu_{pre}\right) = \begin{pmatrix}\sigma_{11}+\sigma_{22} & \sigma_{13}+\sigma_{24}\\ \sigma_{13}+\sigma_{24} & \sigma_{33}+\sigma_{44}\end{pmatrix}^{-1}\begin{pmatrix}\mu_{X_2} - \mu_{X_1}\\ \mu_{Y_2} - \mu_{Y_1}\end{pmatrix}.$$

Thus, we have

$$\begin{pmatrix}\lambda_1^o\\ \lambda_2^o\end{pmatrix} \propto \begin{pmatrix}\sigma_{33}+\sigma_{44} & -(\sigma_{13}+\sigma_{24})\\ -(\sigma_{13}+\sigma_{24}) & \sigma_{11}+\sigma_{22}\end{pmatrix}\begin{pmatrix}\mu_{X_2} - \mu_{X_1}\\ \mu_{Y_2} - \mu_{Y_1}\end{pmatrix} = \begin{pmatrix}-\Delta\mu_X\delta_3 + \Delta\mu_Y\delta_2\\ \Delta\mu_X\delta_2 - \Delta\mu_Y\delta_1\end{pmatrix}.$$

It is obvious that the proposed method corresponds to Su and Liu's result.

A.2. The form of the function h associating μ_{11} with $\mu_{21}, \mu_{31}, \mu_{41}, \boldsymbol{\mu_2}, \Sigma_1, \Sigma_2$ under H_0.

Based on the equation under the null hypothesis $\frac{N_1}{\sqrt{D_1}} = \frac{N_2}{\sqrt{D_2}}$ and assuming that higher values indicate better performance, that is, $A = Pr($ combined pre-treatment biomarkers' values<combined post-treatment biomarkers' values), the root of the equation under the null hypothesis for μ_{11} is

$$\mu_{11} = \mu_{21} + \frac{\left(\Delta_{\mu_{Y1}}(\sigma_{13} - \sigma_{14} - \sigma_{23} + \sigma_{24}) - \left(\left({\Delta_{\mu_{Y1}}}^2 - M(\sigma_{33} - 2\sigma_{34} + \sigma_{44})\right) S\right)^{\frac{1}{2}}\right)}{(\sigma_{33} - 2\sigma_{34} + \sigma_{44})},$$

where

$$\Delta_{\mu_{Y1}} = \mu_{31} - \mu_{41}, M = N_2^2/D_2,$$

$$S = \sigma_{13}^2 - 2\sigma_{13}\sigma_{14} - 2\sigma_{13}\sigma_{23} + 2\sigma_{13}\sigma_{24} + \sigma_{14}^2 + 2\sigma_{14}\sigma_{23} - 2\sigma_{14}\sigma_{24} + \sigma_{23}^2 - 2\sigma_{23}\sigma_{24} + \sigma_{24}^2 - \sigma_{11}\sigma_{33} + 2\sigma_{11}\sigma_{34} + 2\sigma_{12}\sigma_{33} - 4\sigma_{12}\sigma_{34} - \sigma_{11}\sigma_{44} - \sigma_{22}\sigma_{33} + 2\sigma_{12}\sigma_{44} + 2\sigma_{22}\sigma_{34} - \sigma_{22}\sigma_{44}.$$

Assuming that lower values indicate better performance, that is, $A = Pr($combined pre-treatment biomarkers' values>combined post-treatment biomarkers' values$)$, the root of the equation under the null hypothesis for μ_{11} is

$$\mu_{11} = \mu_{21} + \frac{\left(\Delta_{\mu_{Y1}}(\sigma_{13} - \sigma_{14} - \sigma_{23} + \sigma_{24}) + \left(\left({\Delta_{\mu_{Y1}}}^2 - M(\sigma_{33} - 2\sigma_{34} + \sigma_{44})\right) S\right)^{\frac{1}{2}}\right)}{(\sigma_{33} - 2\sigma_{34} + \sigma_{44})}.$$

Acknowledgments

This research was supported by the NIH grant 1R03DE020851 - 01A1 (the National Institute of Dental and Craniofacial Research).

Refernces

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

Box G.E.P. and Cox D.R. (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B **26,** 211–243.

Choi, S., Hall, W. J., and Schick, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Annals of Statistics* **24,** 841 – 861.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **4**, 837-845.

Eckhard L, Werner A. S and Markus A. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience* **51(5),** 341-352.

Hauck, W., Hyslop, T., and Anderson, S. (2000). Generalized treatment effects for clinical trials. *Statistics in medicine* **19**, 887-899.

Lehmann, E. L. and Romano J. P. (1997). *Testing Statistical Hypotheses*, 2nd edition. New York: John Wiley and Sons.

Liu, C., Liu, A. and Halabi, S. (2011). A min-max combination of biomarkers to improve diagnostic accuracy. *Statistics in Medicine* **30**, 2005-2014.

McClish, D.K. (1987). Comparing the areas under more than two independent ROC curves. *Medical Decision Making* **7 (3)**, 149–155.

Pepe, M.S. and Thompson, M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1–2**, 123–140.

Pepe, M.S., Cai, T. and Longton, G. (2006). Combining Predictors for Classification Using The Area under the Receiver Operating Characteristic Curve. *Biometrics* **62**, 221–229.

R Development Core Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0

⟨http://www.R-project.org⟩.

Su, J.Q. and Liu, J.S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350-1355.

Tian, L. (2008). Confidence intervals for $P(Y_1 > Y_2)$ with normal outcomes in linear models. *Statistis in Medicine* **27**, 4221-4237.

Tian, L., Li, X., and Yan, L. (2012). Testing equality of generalized treatment effects. *Journal of Biopharmaceutical Statistics* **22**, 582-595.

Vexler, A., Liu, A., Schisterman, E.F. and Wu, C. (2006). Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *Journal of Nonparametric Statistics* **18**, 145–158.

Vexler, A., Liu, A., Eliseeva, E. and Schisterman, E.F. (2008). Maximum Likelihood Ratio Tests for Comparing the Discriminatory Ability of Biomarkers Subject to Limit of Detection. *Biometrics* **64**, 895–903.

Vexler, A., Schisterman, E.F., and Liu, A. (2008) Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine* **27**, 280–296.

Vexler, A., Tsai, W.M., Gurevich, G. and Yu, J. (2012). Two sample density-based empirical likelihood ratio tests based on paired data, with application to a treatment study of attention-deficit/hyperactivity disorder and severe mood dysregulation. *Statistics in Medicine* **31**, 1821–1837.

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

Table 1

The Monte Carlo powers of the proposed test with different sample sizes $(\boldsymbol{n_1}, \boldsymbol{n_2})$ at the expected 0.05 significance level in the equal covariance case.

Difference	AUC$_{max}$	μ_{21}, μ_{22}	(30,30)	(100,100)	(300,300)
0.1	$A_1 = 0.70$,	$\mu_{21} = 9.9784$,	0.1131^{o}	0.4299^{o}	0.8867^{o}
	$A_2 = 0.60$	$\mu_{22} = 9.0071$	0.1223*	0.2450*	0.5790*
			0.1342**	0.3320**	0.6420**
0.1	$A_1 = 0.85$,	$\mu_{21} = 11.7739$,	0.2300^{o}	0.5884^{o}	0.9611^{o}
	$A_2 = 0.75$	$\mu_{22} = 10.5066$	0.0988*	0.1968*	0.3551*
			0.1714**	0.3026**	0.6019**
0.2	$A_1 = 0.80$,	$\mu_{21} = 11.0925$,	0.4649^{o}	0.9600^{o}	1.0000^{o}
	$A_2 = 0.60$	$\mu_{22} = 9.0071$	0.1945*	0.3140*	0.5899*
			0.1762**	0.3460**	0.6501**
0.2	$A_1 = 0.95$,	$\mu_{21} = 13.8977$,	0.8111^{o}	0.9994^{o}	1.0000^{o}
	$A_2 = 0.75$	$\mu_{22} = 10.5066$	0.1932*	0.3591*	0.4980*
			0.1591**	0.3082**	0.6419**

Note: "o" denotes the power of the proposed test with respect to the best linear combination of two biomarkers, while "*" denotes the power for one biomarker (X) alone based on values of $\mathbf{X_{rki}}$, and "**" denotes the power for the other biomarker (Y) alone based on values of $\mathbf{Y_{rki}}$.

Table 2

The Monte Carlo Type I errors of the proposed test with the best linear combination of two biomarkers with different sample sizes $(\boldsymbol{n_1}, \boldsymbol{n_2})$ in the equal covariance case.

AUC*max*	$\mu_{21} = \mu_{22}$	(30,30)	(100,100)	(300,300)
$A_1 = A_2 = 0.60$	9.0071	0.0294	0.0431	0.0493
$A_1 = A_2 = 0.75$	10.5066	0.0325	0.0524	0.0498

Table 3

The Monte Carlo powers of the proposed test with different sample sizes $(\boldsymbol{n_1}, \boldsymbol{n_2})$ at the expected 0.05 significance level in the unequal covariance case.

Difference	AUC*max*	μ_{21}, μ_{22}	(30,30)	(100,100)	(300,300)
0.1	$A_1 = 0.70$, $A_2 = 0.60$	$\mu_{21} = 9.2989$, $\mu_{22} = 9.0071$	0.0800^{o}	0.2900^{o}	0.8056^{o}
			0.1518*	0.3620*	0.6898*
			0.1496**	0.3199**	0.6479**
0.1	$A_1 = 0.85$, $A_2 = 0.75$	$\mu_{21} = 10.5920$, $\mu_{22} = 10.5066$	0.1096^{o}	0.3425^{o}	0.9394^{o}
			0.1507*	0.3594*	0.6723*
			0.1794**	0.3394**	0.6667**
0.2	$A_1 = 0.80$, $A_2 = 0.60$	$\mu_{21} = 10.1010$, $\mu_{22} = 9.0071$	0.2727^{o}	0.8964^{o}	1.0000^{o}
			0.2479*	0.4540*	0.7819*
			0.1770**	0.3193**	0.6321**
0.2	$A_1 = 0.95$, $A_2 = 0.75$	$\mu_{21} = 12.1232$, $\mu_{22} = 10.5066$	0.7436^{o}	0.9999^{o}	1.0000^{o}
			0.3037*	0.5778*	0.8480*
			0.1346**	0.2906**	0.6600**

Note: "o" denotes the power of the proposed test with respect to the best linear combination of two biomarkers, while "*" denotes the power for one biomarker (X) alone based on values of $\mathbf{X_{rki}}$, and "**" denotes the power for the other biomarker (Y) alone based on values of $\mathbf{Y_{jki}}$.

Table 4

The Monte Carlo Type I errors of the proposed test with the best linear combination of two biomarkers with different sample sizes $(\boldsymbol{n_1}, \boldsymbol{n_2})$ in the unequal covariance case.

AUC*max*	μ_{21}, μ_{22}	(30,30)	(100,100)	(300,300)
$A_1 = A_2 = 0.60$	$\mu_{21} = 8.6033,$ $\mu_{22} = 9.0071$	0.0126	0.0485	0.0261
$A_1 = A_2 = 0.75$	$\mu_{21} = 9.6790,$ $\mu_{22} = 10.5066$	0.0513	0.0521	0.0466

Table 5

Comparison of AUC-type measure between the proposed optimal combination case and only one biomarker case (X or Y alone)

	$AUC_{optimal}$	AUC_X	AUC_Y
theoretical value	0.8000[a] 0.8000[b]	0.7587[a] 0.6038[b]	0.5349[a] 0.5340[b]
<u>n=30</u>			
estimated AUC-type measure (MC variance)	0.8147 (0.0034)[a] 0.8140 (0.0034)[b]	0.7639 (0.0040)[a] 0.6069 (0.0054)[b]	0.5355 (0.0057)[a] 0.5354 (0.0056)[b]
<u>n=100</u>			
estimated AUC-type measure (MC variance)	0.8040 (0.0011)[a] 0.8042 (0.0011)[b]	0.7598 (0.0012)[a] 0.6042 (0.0016)[b]	0.5340 (0.0016)[a] 0.5339 (0.0016)[b]
<u>n=300</u>			
estimated AUC-type measure (MC variance)	0.8014 (0.0004)[a] 0.8013 (0.0003)[b]	0.7592 (0.0004)[a] 0.6042 (0.0005)[b]	0.5340 (0.0005)[a] 0.5342 (0.0005)[b]

Note: "a" denotes the result for scenario ***a*** and "b" denotes the result for scenario ***b*** with the Monte Carlo (MC) variance shown in parentheses.