Published in final edited form as:

Stat Med. 2017 July 10; 36(15): 2347–2362. doi:10.1002/sim.7274.

Combining biomarkers for classification with covariate adjustment

Soyoung Kim^a and Ying Huang^{b,*}

^aDivision of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A

^bVaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A

Abstract

Combining multiple markers can improve classification accuracy compared with using a single marker. In practice, covariates associated with markers or disease outcome can affect the performance of a biomarker or biomarker combination in the population. The covariate-adjusted receiver operating characteristics (ROC) curve has been proposed as a tool to tease out the covariate effect in the evaluation of a single marker; this curve characterizes the classification accuracy solely due to the marker of interest. However, research on the effect of covariates on the performance of marker combinations and on how to adjust for the covariate effect when combining markers is still lacking. In this article, we examine the effect of covariates on classification performance of linear marker combinations and propose to adjust for covariates in combining markers by maximizing the nonparametric estimate of the area under the covariate-adjusted ROC curve. The proposed method provides a way to estimate the best linear biomarker combination that is robust to risk model assumptions underlying alternative regression-model-based methods. The proposed estimator is shown to be consistent and asymptotically normally distributed. We conduct simulations to evaluate the performance of our estimator in cohort and case/control designs and compare several different weighting strategies during estimation with respect to efficiency. Our estimator is also compared with alternative regression-model-based estimators or estimators that maximize the empirical area under the ROC curve, with respect to bias and efficiency. We apply the proposed method to a biomarker study from an HIV vaccine trial.

Keywords

Area under the ROC curve; Biomarker combination; Classification; Covariate adjustment; Receiver operating characteristic curve

1. Introduction

Biomarkers can play an important role in disease classification and risk prediction. The receiving operating characteristic (ROC) curve has been a widely used technique for

*Correspondence to: yhuang@fhcrc.org.

characterizing the classification accuracy of a given biomarker [1, 2, 3]. The ROC curve for a marker-based classifier is a graph of sensitivity versus $1 - \text{specificity}$ based on a rule that classifies a subject as positive if the classifier is greater than a decision threshold c , which varies between $-\infty$ and ∞ . The area under the ROC curve (AUC) is a common summary index for the ROC curve, which is interpreted as the probability that a randomly chosen subject with the disease of interest has a larger classification score than a subject without the disease of interest.

In practice, covariates such as age and gender are often collected for study subjects. These covariates can affect the distribution of a biomarker such that the same points on the ROC curve based on a marker can correspond to different decision thresholds for testing positive across covariate levels; moreover, a marker's classification accuracy can also differ across covariate levels [3, 4, 5]. Covariate adjustment has been advocated as a way to tease out the covariate effect in evaluating a biomarker such that the classification accuracy solely due to the marker can be characterized. For evaluation of a univariate marker adjusting for covariate effect, Jane and Pepe [4] proposed the covariate-adjusted ROC curve, which is a weighted average of covariate-specific ROC curves across covariate levels.

For subjects in biomedical or clinical studies, multiple markers possibly correlated with a binary disease outcome variable are often available. Combining multiple markers can potentially improve sensitivity and specificity compared to using a single marker, as well as improve the prediction performance [2]. It is therefore important to develop effective statistical methods that can identify marker combinations with good classification performance. Logistic regression is a common way to identify marker combinations; however, this approach relies on specification of a parametric form for the risk model and can have sub-optimal classification performance when the model assumption is violated. Previously, Pepe et al. [6] proposed using the empirical AUC as an objective function to maximize when identifying linear marker combinations. Maximizing a smoothed approximation of the empirical AUC, such as the sigmoid approximation, has been proposed by other authors [7]. Few existing AUC maximization approaches, however, address the need for covariate adjustment when combining markers, which nevertheless is an important question of scientific interest. To see this, note that empirical AUC involves the comparisons between both cases and controls in the same covariate stratum and cases and controls across strata; therefore, the marker combination identified to maximize the empirical AUC may not be the best combination that maximizes classification performance within a covariate stratum. The latter quantity is oftentimes desirable because of i) the need to understand the achievable marker performance among individuals with a particular covariate level of interest, and ii) the ability to transport the identified marker combination with covariate adjustment to new study populations with possibly different covariate distributions. In contrast, the performance of marker combinations that maximize the empirical AUC in one biomarker study may not be generalizable to new populations in another study that differ in covariate distributions from the current study. Liu and Zhou [5] studied biomarker combination and covariate adjustment and investigated biomarker combinations that maximize the area under a covariate-adjusted ROC and a covariate standardized ROC curve. However, the criterion function to be maximized in Liu and Zhou [5] was estimated based on explicit modeling of the biomarker distributions among cases and controls. Therefore, the

performance of the corresponding marker combinations can be susceptible to violation of these assumptions.

In this article, we examine the effect of covariates on classification performance of linear marker combinations and propose a robust method for deriving best linear marker combinations. This method is based on the work of Pepe et al. [6], but is generalized to allow for covariate adjustment. In particular, we propose to adjust for covariates in combining markers by maximizing the nonparametric estimate of the area under the covariate-adjusted ROC curve (AAUC). In section 2, we present a few examples that demonstrate the importance of adjusting for covariates in deriving optimal marker combinations, where the marker combination that optimizes the marginal AUC will not optimize the covariate-specific AUC, even when the best marker combination is the same across covariate levels. We then propose our AAUC-maximization estimators of linear marker combinations, with the AAUC constructed based on several different covariate-stratum weighting schemes, and develop the asymptotic properties of the estimators. In section 3, we conduct extensive simulations to: i) evaluate the efficiency of incorporating covariate adjustment with different types of covariate-specific weight functions in covariate-adjusted AUC; and ii) compare our proposed estimator of marker combinations with alternative estimators that maximize empirical AUC or maximize likelihood. In section 4, we apply our approach to a biomarker study from an HIV vaccine trial. Concluding Remarks are made in section 5.

2. Model

We consider a binary outcome variable Y_i that indicates the disease status for subject i . Without loss of generality, we define $Y = 1$ for the diseased group and $Y = 0$ for the healthy group. Suppose there are d biomarkers measured on each subject. Let $X_i = (X_{i1}, X_{i2}, \dots, X_{id})$ be the biomarker measurement for subject i . We use X_D to indicate the biomarkers for a subject with the disease and use X_H to indicate the biomarkers for a healthy subject. Moreover, let Z_i denote the covariates for subject i . In this paper, we focus on discrete Z such as gender or age subgroup.

Consider the search for the optimal linear combination of X . The overall classification performance of the ROC curve can be summarized by AUC as a single index. When a generalized linear model holds for the risk of Y conditional on marker X , $P(Y|X) = G(\beta^T X)$ for a pre-specified link function G , the best classifier can be shown to be $\beta_0^T X$, which achieves maximum AUC of $P(\beta_0^T X_D > \beta_0^T X_H)$ where $\beta_0 = (\beta_{01}, \dots, \beta_{0d})$ is the true parameter [8]. The marginal AUC can be empirically estimated by

$$\hat{AUC}(\beta) = \frac{\sum_{i \neq j} I(\beta^T X_{Di} > \beta^T X_{Hj})}{n_D n_H}, \quad (1)$$

where n_D and n_H are the numbers of subjects with the disease and healthy subjects, respectively, and I is the indicator function. Pepe and Thompson [9] proposed to identify the best linear marker combination by maximizing the empirical AUC. Note that the marker

combination maximizing the empirical AUC is a special case of the maximum rank correlation (MRC) estimator proposed by Han [10] and the MRC estimator was shown to be consistent and asymptotically normally distributed by Sherman [11].

When it is of interest to examine the accuracy of markers within a subpopulation, the covariate-specific AUC has been recommended [12]. Given a specific covariate level $Z=z$, the covariate-specific AUC for $\beta^T X$ is defined as

$AUC_z(\beta) = P(\beta_0^T X_D > \beta_0^T X_H | Z_D = Z_H = z)$ where Z_D and Z_H are the covariates for subjects with the disease and healthy subjects, respectively. When a generalized linear model for risk of Y conditional on X holds within covariate level z , the best classifier within this covariate level can be similarly found as the maximizer of the covariate-specific AUC.

Next we consider a generalized linear risk model across covariate levels where the risk of Y conditional on X and Z follows $P(Y|X, Z) = \text{G}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 X_1 Z + \beta_5 X_2 Z)$. The best marker combination within covariate level z is $(\beta_1 + \beta_4 z)X_1 + (\beta_2 + \beta_5 z)X_2$, or $X_1 + (\beta_2 + \beta_5 z)/(\beta_1 + \beta_4 z)X_2$ if $\beta_1 + \beta_4 z \neq 0$ following McIntosh and Pepe [8]. For scenarios where the interaction between X and Z (i.e. β_4 & β_5) is zero, or $(\beta_2 + \beta_5)/(\beta_1 + \beta_4) = \beta_2/\beta_1$, the best marker combination will be the same across covariate strata. One might think that covariate adjustment is not necessary in this case to identify the best linear marker combination, i.e., the common linear marker combination that maximizes the covariate-specific classification performance will also have the best population performance. However, as we will show in the following examples, the common best marker combination that optimizes AUC within each covariate stratum may not be the marker combination that maximizes the marginal AUC in the presence of covariate-disease and/or covariate-marker associations.

2.1. Examples of different best marker combinations with and without covariate adjustment

In this section we present two examples where the best linear marker combination that is common across covariate levels is different from the linear marker combination that maximizes the marginal AUC.

Example 1—Suppose biomarkers $X = (X_1, X_2)$ and covariate Z follow a multivariate normal distribution with mean μ and covariance matrix Σ where

$$\mu = \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \\ \mu_z \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{x1}^2 & \rho_{x1x2}\sigma_{x1}\sigma_{x2} & \rho_{x1z}\sigma_{x1}\sigma_z \\ \rho_{x1x2}\sigma_{x1}\sigma_{x2} & \sigma_{x2}^2 & \rho_{x2z}\sigma_{x2}\sigma_z \\ \rho_{x1z}\sigma_{x1}\sigma_z & \rho_{x2z}\sigma_{x2}\sigma_z & \sigma_z^2 \end{pmatrix}.$$

It follows that the distribution of Z conditional on X_1 and X_2 , $Z|X_1, X_2$, is normal with mean $k_0 + k_1 X_1 + k_2 X_2$ where $k_1 = (\rho_{x1z} - \rho_{x2z}\rho_{x1x2})\sigma_z / (1 - \rho_{x1x2}^2)\sigma_{x1}$, $k_2 = (\rho_{x2z} - \rho_{x1z}\rho_{x1x2})\sigma_z / (1 - \rho_{x1x2}^2)\sigma_{x2}$, $k_0 = -k_1\mu_{x1} - k_2\mu_{x2}$, and variance $\sigma_*^2 = \{1 - (\rho_{x1z}^2 - 2\rho_{x1z}\rho_{x2z}\rho_{x1x2} + \rho_{x2z}^2) / (1 - \rho_{x1x2}^2)\}\sigma_z^2$. We further assume a probit model for the risk of Y conditional on X and Z : $P(Y=1|X, Z) = \Phi(\alpha(X_1 + \beta X_2) + \gamma + \lambda Z)$ such

that the best marker combination within each covariate stratum is $X_1 + \beta X_2$. As presented in Appendix A of the Supplementary Material, a closed-form expression of the risk of Y conditional on X can be derived as

$$P(Y=1|X) = \Phi \left\{ \frac{(\alpha + \lambda k_1)(X_1 + \frac{\alpha\beta + \lambda k_2}{\alpha + \lambda k_1} X_2) + \gamma + \lambda k_0}{\sqrt{(1 + \lambda^2 \sigma_*^2)}} \right\}.$$

The marker combination that maximizes the marginal AUC is thus $X_1 + (\alpha\beta + \lambda k_2)/(\alpha + \lambda k_1)X_2$. It will be different from the best marker combination conditional on covariate if

$$\beta \neq \sigma_{x1} (\rho_{x_2 z} - \rho_{x_1 x_2} \rho_{x_1 z}) / \sigma_{x2} (\rho_{x_1 z} - \rho_{x_1 x_2} \rho_{x_2 z}).$$

Example 2—We consider the following distributions for covariate Z , disease outcome Y , and two biomarkers X_1 and X_2 : 1) Z follows a Bernoulli distribution with $P(Z=z) = P_{Z=z}$; 2) $Y|Z=z$ follows a Bernoulli distribution with $P(Y=y|Z=z) = P_{Y=y|Z=z}$; 3) $(X_1, X_2)|(Y=y, Z=z)$ follows a bivariate normal distribution with mean $\mu_{zy} = (\mu_{zy,1}, \mu_{zy,2})$, variance 1 for each marker, and correlation ρ between two markers. This configuration induces the logistic model for the risk of Y conditional on X and Z : $\text{logit}\{P(Y=1|X, Z)\} = \{\alpha_0 + (\alpha_1 - \alpha_0)Z\} [X_1 + \{\beta_0 + (\beta_1 - \beta_0)Z\}X_2] + (\gamma_1 - \gamma_0)Z + \gamma_0, z=0,1$ where

$$\beta_z = \frac{\mu_{z1,2} - \mu_{z0,2} + \rho(\mu_{z0,1} - \mu_{z1,1})}{\mu_{z1,1} - \mu_{z0,1} + \rho(\mu_{z0,2} - \mu_{z1,2})}, \alpha_z = \frac{\mu_{z1,1} - \mu_{z0,1} + \rho(\mu_{z0,2} - \mu_{z1,2})}{1 - \rho^2},$$

$$\gamma_z = \frac{1}{2(1 - \rho^2)} \{ \mu_{z0,1}^2 + \mu_{z0,2}^2 - 2\rho\mu_{z0,1}\mu_{z0,2} - (\mu_{z1,1}^2 + \mu_{z1,2}^2 - 2\rho\mu_{z1,1}\mu_{z1,2}) \},$$

for $z=0,1$. Through Bayes' theorem, we have:

$$\begin{aligned} \log \frac{P(Y=1|X)}{P(Y=0|X)} &= \log \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)} \\ &= \log \frac{\sum_z P(X|Y=1, Z=z)P(Y=1, Z=z)}{\sum_z P(X|Y=0, Z=z)P(Y=0, Z=z)} \\ &= \log \frac{\sum_z P(X|Y=1, Z=z)P(Y=1|Z=z)P(Z=z)}{\sum_z P(X|Y=0, Z=z)P(Y=0|Z=z)P(Z=z)} \\ &= \log \frac{W_1 P(X|Z=0, Y=1) + W_2 P(X|Z=1, Y=1)}{W_3 P(X|Z=0, Y=0) + W_4 P(X|Z=1, Y=0)}, \end{aligned} \quad (2)$$

where $W_1 = P(Z=0, Y=1)$, $W_2 = P(Z=1, Y=1)$, $W_3 = P(Z=0, Y=0)$, $W_4 = P(Z=1, Y=0)$. Suppose $\{\mu_{01,2} - \mu_{00,2} + \rho(\mu_{00,1} - \mu_{01,1})\} / \{\mu_{01,1} - \mu_{00,1} + \rho(\mu_{00,2} - \mu_{01,2})\} = \{\mu_{11,2} - \mu_{10,2} + \rho(\mu_{10,1} - \mu_{11,1})\} / \{\mu_{11,1} - \mu_{10,1} + \rho(\mu_{10,2} - \mu_{11,2})\}$ such that β is constant across covariate levels (i.e. $\beta_0 = \beta_1 = \beta$). In order for $X_1 + \beta X_2$ to be the marker combination maximizing the marginal AUC, there must exist a monotone transformation of (2) that equals $X_1 + \beta X_2$. As shown in Appendix B of the Supplementary Material, this is true for certain parameter settings. For example, it is satisfied for $\beta=1$, where (2) can be written as $\alpha(X_1 + \beta X_2) + C$ with C a constant. When $\beta=1$, this is satisfied, for example, if 1) $P(X_1, X_2|$

$Y = 1, Z = 0) = P(X_1, X_2 | Y = 1, Z = 1)$ and $P(X_1, X_2 | Y = 0, Z = 0) = P(X_1, X_2 | Y = 0, Z = 1)$ or $2) \mu_{00,k} = t_1 \mu_{01,k} = t_2 \mu_{10,k} = t_3 \mu_{11,k}$ for $k = 1, 2$ where $t_1 = 0, t_2 = 0, t_3 = 0$ are constants together with some constraints on $\mu_{zy,k} (z = 0, 1, y = 0, 1, k = 1, 2)$ and ρ as presented in Appendix B of the Supplementary Material. However, in general, there is no guarantee that $X_1 + \beta X_2$ will be the marker combination maximizing the marginal AUC. In Appendix B of the Supplementary Material, we present two counter examples where the common linear marker combination maximizing the covariate-specific AUC is different from the best linear marker combination maximizing the marginal AUC. In both counter examples $P(X = 1 | Y = 0, Z)$ $P(X = 1 | Y = 1, Z)$; Z and Y are independent in one example and correlated in the other.

2.2. AAUC estimator of marker combinations

To characterize the performance of a single marker accounting for a covariate effect, the covariate-adjusted AUC, $\int P(X_D > X_H | Z) dP(Z | Y = 1)$ has been proposed by [4], which is essentially a weighted average of covariate-specific AUC where weight for stratum Z equals density of Z among cases. We generalize this to a covariate-adjusted AUC curve based on a linear marker combination: $AAUC(\beta) = \sum_z W(z) P(\beta_0^T X_D > \beta_0^T X_H | Z_D = Z_H = z)$, where $W(z)$ is the weight given to the covariate level z with $\sum_z W(z) = 1$. For given β , one can estimate the covariate-adjusted AUC nonparametrically with

$$\widehat{AAUC}(\beta) = \sum_z W(z) \left\{ \frac{\sum_{i \neq j} I(\beta^T X_{Di} > \beta^T X_{Hj}) I(Z_{Di} = Z_{Hj} = z)}{\sum_{i \neq j} I(Z_{Di} = Z_{Hj} = z)} \right\}. \quad (3)$$

We propose to estimate β , the parameter for the best marker combination adjusting for covariate effect, through the maximization of the nonparametric estimate of AAUC. That is, we solve for

$$\hat{\beta} = \argmax_{\beta} \{ \widehat{AAUC}(\beta) \}. \quad (4)$$

In general, the weight function $W(z)$ can be chosen to reflect the importance of having good classification performance at various covariate levels z when identifying a marker combination. For scenarios where the marker combination maximizing AUC is the same across covariate strata, it will also maximize the AAUC regardless of which weight scheme we use. A wise choice of weights, however, might lead to better efficiency in estimating the common best linear marker combination. In this paper, we consider three different types of weights for constructing the AAUC. The first type of weight (Type 1) is proportional to the total number of pairs within each covariate stratum $P(Z_D = z)P(Z_H = z)/\sum_z P(Z_D = z)P(Z_H = z)$. The second type of weight (Type 2) is $P(Z = z)$, the density of each covariate stratum in the population. The third type of weight (Type 3) is the density of each covariate stratum among cases $P(Z | Y = 1)$, which has been adopted in [4]. Note that the AAUC based on Type 3 weight has a useful interpretation as the probability of correctly ordering a randomly chosen case and a control observation with the same covariate value. While other choices of $W(z)$ may not have a similar probability interpretation, they still have meaningful

interpretations in terms of a weighted average of covariate-specific AUC. For example, suppose it is of interest to apply the marker combination to a new population where thresholds of the marker combination are allowed to vary across covariate strata. One can choose the weights to reflect the relative importance of classification accuracy desired to achieve in each covariate stratum.

Next we develop an asymptotic theory for AAUC-maximization estimator of marker combinations for a general weight Scheme.

2.3. Asymptotic properties

In this section, we present asymptotic properties for our proposed estimator $\hat{\beta}$ for true β_0 . Let $B = (X, Y, Z)$ denote an observation from a distribution P on a set $S \subseteq \mathbb{R}^d \otimes \mathbb{R} \otimes \mathbb{R}$ where $X = (X_1, X_2, \dots, X_d)$ is a vector of d markers, Y is a binary outcome, and Z is a discrete covariate. We extend the risk model introduced by [10] to consider covariate adjustment. Suppose the risk model Y given X and Z is given by

$$P(Y|X, Z) = G\{\beta'X + \sum_{z=1}^K \gamma_z I(Z=z)\},$$

where G is a monotone increasing function, $\beta = (\beta_1, \beta_2, \dots, \beta_d)'$ are unknown parameters for main effect of X , and γ_z are unknown parameters for main effect of $Z = z$, $z = 1, \dots, K$.

For convenience, let $\beta = (1, \theta)'$ where $\theta = (\beta_2/\beta_1, \dots, \beta_d/\beta_1)'$ is the parameter dictating the covariate-specific optimal marker combination. Suppose a random sample of size n is obtained from P , based on which we obtain the AAUC estimator of marker combination $\hat{\beta} = (1, \hat{\theta})$ based on (4). To show consistency of $\hat{\theta}$, we make the following assumptions:

A1: The true parameter value θ_0 is an interior point of Θ , which is a compact subset of $\mathbb{R}^{d-1}$.

A2: Let S_X denote the support of the covariate vector X .

- i. S_X is not contained in any proper linear subspace of $\mathbb{R}^d$.
- ii. The first component of X has an everywhere positive density, conditional on other components.

A3: $W(z)$ is constant with $\sum_{z=1}^K W(z) = 1$ and bounded as n goes to ∞ .

The consistency of $\hat{\theta}$ is presented in Theorem 2.1 as follows.

Theorem 2.1—Under the assumptions A1–A3, $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \rightarrow \infty$.

The detail for the proof is provided in Appendix C of the Supplementary Material.

Next we summarize asymptotic normality of $\hat{\theta}$. Let χ denote the last $d-1$ components of the covariate X and $g_0(\cdot)$ the marginal density of $\beta'_0 X$. Let the expectation of χ conditional

on $\beta'_0 X$ be $\chi_0 = E(\chi | \beta'_0 X)$. Denote $A(Y, \beta'_0 X, Z) = Y - G \left\{ \beta'_0 X + \sum_{z=1}^K \gamma_z I(Z=z) \right\}$ and $A'(Y, t, z) = \partial_t A(Y, t, z)$. For normality of $\hat{\theta}$, we make the following additional assumptions:

A4: $A'(Y, \beta'_0 X, Z)$ exists and $E|X|^2 < \infty$.

A5: $W^2(Z)$ is bounded.

A6: Σ and Δ exist and Σ is negative definite,

where

$$\Sigma = E \left[\sum_{z=1}^K I(Z=z) W(z)^2 (\chi - \chi_0)' (\chi - \chi_0) \{A(Y, \beta'_0 X, Z=z)\}^2 \{g_0(\beta'_0 X)\}^2 \right],$$

$$2\Delta = -E \left[\sum_{z=1}^K I(Z=z) W(z) A'(Y, \beta'_0 X, Z=z) g_0(\beta'_0 X) \{(\chi - x)' (\chi - x) | \beta'_0 X\} \right].$$

The asymptotic normality of $\hat{\theta}$ is presented in Theorem 2.2 as follows:

Theorem 2.2—Under the assumptions A1–A6, $n^{1/2}(\hat{\theta} - \theta_0) \rightarrow^D N(0, \Delta^{-1} \Sigma \Delta^{-1})$ as $n \rightarrow \infty$,

The derivation of the asymptotic variance $\Delta^{-1} \Sigma \Delta^{-1}$ follows the Taylor expansion, Theorem 3 from Sherman [11], and the Euclidean properties of Nolan and Pollard [13]. The proof of Theorem 2.2 is given in Appendix D of the Supplementary Material.

In practice, we use a nonparametric bootstrap resampling method for estimating the standard error of $\hat{\theta}$ and for construction of its confidence interval, as recommended in other AUC-maximization studies, e.g. one by Ma and Huang [7].

3. Simulations

In this section we examine finite-sample performance of our proposed AAUC estimator for deriving the best linear combination adjusting for covariate effects. Consider the setting with a binary covariate Z , a binary outcome Y , and two continuous biomarkers X_1 and X_2 . Our proposed AAUC estimator of β (for a best linear marker combination $X_1 + \beta X_2$) is computed as below.

1. Corresponding to a candidate β value, we calculate covariate-specific AUC nonparametrically within each covariate level (i.e. $\widehat{AUC}_{z=0}(\beta)$ and $\widehat{AUC}_{z=1}(\beta)$).
2. We calculate the covariate-adjusted AUC which is a weighted average of covariate-specific AUC weighted by specific weight function such as Type 1, Type2 and Type3 (i.e. $\widehat{AAUC}(\beta) = W_0 \times \widehat{AUC}_{z=0}(\beta) + (1 - W_0) \widehat{AUC}_{z=1}(\beta)$, where W_0 is the weight when $Z = 0$).

3. We repeat steps (1) and (2) over a grid of β values from -2 to 2 with grid size 0.01 , and identify the β value corresponding to the maximum of $\widehat{AAUC}$.

In our simulations, we investigate both scenarios where risk of disease follows a linear logistic regression model and scenarios where the linear logistic model assumption is violated. Next we describe the setup for the first type of scenario and leave the description of the second type to Section 3.2.

In the first type of settings, the joint distributions of Z , Y , X_1 , and X_2 are generated as in example 2) of Section 2.1. **In particular, Z is generated from a Bernoulli distribution with probability $P(Z=1) = P_{Z=1}$, $Y|Z$ is generated from Bernoulli distribution with $P_{Y=1|Z}$ and then $(X_1, X_2)|Z=z, Y=y$ are generated from a bivariate normal distributions with mean $(\mu_{zy,1}, \mu_{zy,2})$ and variance matrix**

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

This implies that $(X_1, X_2)|Z$ is a mixture normal distribution. The risk of Y given X and Z follows the logistic model

$$\text{logit} \{P(Y=1|X, Z)\} = \{\alpha_0 + (\alpha_1 - \alpha_0) Z\} (X_1 + \{\beta_0 + (\beta_1 - \beta_0) Z\} X_2) + (\gamma_1 - \gamma_0) Z + \gamma_0,$$

where the value of α , β , γ can be found in example 2). Note that this model corresponds to a linear logistic regression model that includes main effects for covariate and each biomarker, and interaction terms between covariate and each biomarker.

Based on this model, the best marker combinations within $Z=0$ and $Z=1$ are $X_1 + \beta_0 X_2$ and $X_1 + \beta_1 X_2$, where

$$\beta_i = \frac{\mu_{i1,2} - \mu_{i0,2} + \rho(\mu_{i0,1} - \mu_{i1,1})}{\mu_{i1,1} - \mu_{i0,1} + \rho(\mu_{i0,2} - \mu_{i1,2})}, (i=0, 1). \quad (5)$$

In our simulations, we consider scenarios where the best linear marker combinations are same across covariate strata, i.e., $\beta_0 = \beta_1$ but distribution of X among cases or controls are different between covariate strata, i.e. $P(X/Y, Z=0) \neq P(X/Y, Z=1)$, although the application of our estimator is not limited to these scenarios, as discussed later in the Concluding Remarks. This corresponds to the logistic regression model with the following form:

$$\text{logit} \{P(Y=1|X, Z)\} = \{\alpha_0 + (\alpha_1 - \alpha_0) Z\} (X_1 + \beta X_2) + (\gamma_1 - \gamma_0) Z + \gamma_0.$$

When $\alpha_0 = \alpha_1$, the model only includes main effects of markers and covariate:

$$\text{logit} \{P(Y=1|X, Z)\} = \alpha_0 (X_1 + \beta X_2) + (\gamma_1 - \gamma_0) Z + \gamma_0. \quad (6)$$

When $\alpha_0 = \alpha_1$, the model includes in addition the interaction term between markers and covariates:

$$\text{logit} \{P(Y=1|X, Z)\} = \alpha_0 (X_1 + \beta X_2) + (\alpha_1 - \alpha_0) (X_1 + \beta X_2) Z + (\gamma_1 - \gamma_0) Z + \gamma_0. \quad (7)$$

Under the model in (6), we set $\mu_{00} = (0,0)$, $\mu_{01} = (1,1)$, $\mu_{10} = (0.5, 0.5)$, and $\mu_{11} = (1.5, 1.5)$ for true $\beta_0 = \beta_1 = 1$ and $\mu_{00} = (0,0)$, $\mu_{01} = (1,1.25)$, $\mu_{10} = (-1, -3)$, and $\mu_{11} = (0, -1.75)$ for true $\beta_0 = \beta_1 = 2$. Under the model in (7), for true $\beta_0 = \beta_1 = 2$, means of X are set to be 1) $\mu_{00} = (0, 0)$, $\mu_{01} = (0.8, 1)$, $\mu_{10} = (-1, -3)$, and $\mu_{11} = (-0.04, -1.8)$, and 2) $\mu_{00} = (0,0)$, $\mu_{01} = (0.8, 1)$, $\mu_{10} = (-3, -2)$, and $\mu_{11} = (-5, -4.5)$. For scenarios where Y and Z are not correlated, we set $(P_{Y=1|Z=0}, P_{Y=1|Z=1}) = (0.2, 0.2)$. For scenarios where Y and Z are correlated, we set $(P_{Y=1|Z=0}, P_{Y=1|Z=1}) = (0.14, 0.26)$ when $P_{Z=1} = 0.5$ and $(P_{Y=1|Z=0}, P_{Y=1|Z=1}) = (0.14, 0.34)$ when $P_{Z=1} = 0.3$. These scenarios are designed to have disease prevalence 20% in the population. For each simulation setting, 2,000 Monte Carlo simulations are conducted.

3.1. Comparison of different weighting strategies for AAUC estimator

We first examine the performance of AAUC estimators under the three types of weights as specified in Section 2.2, using both cohort samples and frequency-matched case-control samples. For the cohort study, the estimates of Type 1, Type 2, and Type 3 weights are used to compute the AAUC-estimator of marker combinations. In particular, the estimated Type 1 weight is given by

$$\begin{aligned} \hat{W}_1(Z=z) &= \frac{\hat{I}(z=1) \{ \hat{P}_{Z=1}^2 \hat{P}_{Y=1|Z=1} (1 - \hat{P}_{Y=1|Z=1}) + \hat{I}(z=0) \{ (1 - \hat{P}_{Z=1})^2 \hat{P}_{Y=1|Z=0} (1 - \hat{P}_{Y=1|Z=0}) \} \}}{\hat{P}_{Z=1}^2 \hat{P}_{Y=1|Z=1} (1 - \hat{P}_{Y=1|Z=1}) + (1 - \hat{P}_{Z=1})^2 \hat{P}_{Y=1|Z=0} (1 - \hat{P}_{Y=1|Z=0})} \\ &= \frac{\{ \sum I(Y=1 \& Z=z) \} \times \{ \sum I(Y=0 \& Z=z) \}}{\sum_{z=0}^1 \{ \sum I(Y=1 \& Z=z) \} \times \{ \sum I(Y=0 \& Z=z) \}} \end{aligned}$$

where $\hat{P}_{Z=1}$, $\hat{P}_{Y=1|Z=1}$, and $\hat{P}_{Y=1|Z=0}$ are the empirically estimated probabilities for $Z=1$, $Y=1$ given $Z=1$, and $Y=1$ given $Z=0$ using the cohort sample. Type 2 and Type 3 are

estimated by $\hat{W}_2(Z=z) = \hat{P}_{Z=1}$ and $\hat{W}_3(Z=z) = \hat{P}_{Z=z|Y=1}$. The three types of weights are generated similarly for case-control sample, where we replace population parameters $P(Z)$, $P(Y|Z=1)$ and $P(Y|Z=0)$ by their analogues in the sample distribution.

Table 1 presents results for the AAUC-estimator of marker combinations under risk model (6), for various probabilities of Z and probabilities of Y given Z , based on cohort samples of size 600 or 2400. In particular, we report the average of β , its empirical standard deviation (SD), the average of the nonparametric bootstrap standard errors (SE), and the median of the nonparametric bootstrap standard errors ($\overline{SE}$) with 200 resamples without replacement in bootstrap. The results indicate that for all types of weights, the estimates for β are approximately unbiased; both the average and median bootstrap standard errors are in

general close to the empirical standard deviations; the coverage of the 95% percentile bootstrap confidence interval is close to the nominal level. For the cohort study, the expected Type 1, Type 2, and Type 3 weights are same under the setting with $P_{Z=1} = 0.5$ and $(P_{Y=1|Z=0}, P_{Y=1|Z=1}) = (0.2, 0.2)$. Thus, the overall performances of the Type 1, Type 2, and Type 3 weights are similar under this setting. In general, Type 3 weight is only slightly more efficient compared to Type 1 and Type 2 weights.

The results using frequency-match case-control samples under the same set of settings are also presented in Table 1. We conducted the frequency-matched case-control sampling as follows: first, a random sample of cases is simulated; then, samples from two times as many controls were obtained randomly within each covariate stratum. Nonparametric bootstrap sampling is performed with 200 resamples stratified on case/control status and on covariate stratum within controls. The overall performance of the AAUC estimator of marker combinations is similar to that in the cohort study: all estimates for β have minimum bias; the average bootstrap standard error is close to the empirical standard deviation of $\hat{\beta}$ and the coverage of the 95% bootstrap percentile CI is close to the nominal level. Here, for the frequency-matched case-control study, the distribution of the covariate among all samples is the same as its distribution among cases. As a result, the Type 2 and Type 3 weights are exactly same and equal $\hat{P}(Z=z|Y=1)$. The Type 1 weight, on the other hand, equals $\hat{P}(Z=z|Y=1) / \sum_{z=0}^1 \hat{P}(Z=z|Y=1)$. The expected Type 1 weight is the same as the Type 2 and Type 3 weights when $P_{Z=1} = 0.5$ and $P_{Y|Z=0} = P_{Y|Z=1} = 0.2$ or $P_{Z=1} = 0.3$ and $P_{Y=1|Z=0} = 0.14$, $P_{Y=1|Z=1} = 0.34$. Under these settings, the standard deviations for $\hat{\beta}$ are comparable between the three weight schemes. In general, Type 2 or Type 3 weights are only slightly more efficient compared to Type 1 weight. The variances of β estimates for case-control design are smaller than those for cohort design with the same sample size, suggesting that a case-control design is more efficient compared to a cohort study design for finding the best marker combinations.

3.2. Comparison of AAUC estimators with other estimators

In this section, we compare the performance of the proposed AAUC-maximization estimator of marker combinations with other existing estimators of marker combinations. In particular, we consider the marker combination maximizing the marginal AUC, and a logistic regression based estimator, where **we fit the model** $P(Y=1|X, Z) = \psi_0 + \psi_1 X_1 + \psi_2 X_2 + \psi_3 Z$ **and compute the best marker combination with** $\hat{\beta} = \hat{\psi}_2 / \hat{\psi}_1$. We call this estimator the LGT estimator. In addition, we investigate the parametric estimator proposed by Liu and Zhou [5] (LZ), which assumes the two markers follow a bivariate normal distribution with the following mean and variance matrix:

$$\mu = \begin{pmatrix} \mu_{x1} = \zeta_1 Z + \gamma_1 Y \\ \mu_{x2} = \zeta_2 Z + \gamma_2 Y \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{x1}^2 & \rho_{x1x2} \sigma_{x1} \sigma_{x2} \\ \rho_{x1x2} \sigma_{x1} \sigma_{x2} & \sigma_{x2}^2 \end{pmatrix}.$$

Under this assumption, covariate-specific AUC is common for all the covariates and maximizing AAUC is the same as maximizing the covariate-specific AUC. LZ thus

estimates $\beta = \frac{\gamma_2 \sigma_{x_1}^2 - \gamma_1 \rho_{x_1 x_2} \sigma_{x_1} \sigma_{x_2}}{\gamma_1 \sigma_{x_1}^2 - \gamma_2 \rho_{x_1 x_2} \sigma_{x_1} \sigma_{x_2}}$. For our proposed AAUC estimator, we use the Type 3 weight, given that all three types of weights have comparable performance as shown in Table 1.

We consider both settings where the linear logistic regression model assumption is true and settings where the assumption is violated. For each scenario, the best marker combination is a linear combination that is the same across covariate strata. Cohort samples (training sets) of size 500 or 2000 are simulated in each setting to estimate the best linear marker combination. Test sets of size 2000 are simulated from the target population to estimate the performance of the estimated marker combinations.

I. Scenarios where the risk of disease follows a linear logistic model with main effects only:

We first consider a setting where the model in (6) holds, i.e. the linear logistic regression model with main effects only. Under this model, covariate-specific AUC based on a particular marker combination is the same across covariate strata. The model assumptions for both the LGT and the LZ estimators are satisfied. We compare different estimators with respect to β estimator as well as the common covariate-specific AUC, denoted as $AUC_z(\hat{\beta})$, by applying the estimated marker combination to the population. **Note that the best marker combination $\hat{\beta}$ for the LGT and the LZ are ratio estimators. The non-negligible probability mass around 0 for the denominator can make the distribution of the point estimates for β highly skewed. For these two estimators, we therefore present the following transformation of $\hat{\beta}$ when true β is 2:**

$$\hat{\beta}^* = \begin{cases} \hat{\beta} & \text{if } -1 \leq \hat{\beta} \leq 5 \\ 6 - 1/(\hat{\beta} - 4) & \text{if } \hat{\beta} > 5 \\ -2 - 1/\hat{\beta} & \text{if } \hat{\beta} < -1 \end{cases},$$

such that $\hat{\beta}^*$ is bounded between -2 and 6 . Table 2 presents the average of the β estimates; their empirical standard deviations (SD); the average of $AUC_z(\hat{\beta})$; and their standard deviations for AAUC, AUC, LGT, and LZ estimators. The results show that estimates of marker combinations based on AAUC, LGT, and LZ are asymptotically unbiased. In contrast, bias is observed for all marker combinations based on maximizing the empirical AUC estimates. When sample size is not very large ($n=500$), LGT and LZ have small bias, but bias vanishes as sample size increases to 2000. With their model assumptions satisfied, both the LGT and LZ estimators are slightly more efficient than the AAUC estimators for large samples, as expected. However, the AAUC estimator appears to be more efficient than the LGT and LZ estimators for medium sample sizes ($n=500$), which are more likely to reflect the actual sample sizes in biomedical studies. The covariate-specific AUC values are similar between the AAUC, LGT, and LZ estimators for each sample size, whereas the classification accuracy based on marker combinations that maximize the empirical AUC is much smaller.

II. Scenarios where the risk of disease follows a linear logistic model that includes a marker $\times$ covariate interaction:

We next consider scenarios under a linear logistic risk model (7), which includes interactions between X_1 , X_2 and Z . Under this model, the covariate-specific AUC based on the same marker combination is different across covariate strata. The model assumption for both LGT and LZ are violated. **We again present bounded estimates $\hat{\beta}^*$ for both LGT and LZ as in Scenarios 1.**

Table 3 summarizes the performance of the estimated marker combination and the corresponding covariate-specific AUC. The AAUC estimator of the marker combination has minimum bias, whereas bias can be observed for the AUC estimator of marker combinations. The bias for the LGT and LZ estimators of marker combinations decreases when the sample size gets large. These estimators tend to be less efficient than the AAUC estimator of the marker combination under a sample size of 500; for a sample size of 2000, they may or may not be more efficient than the AAUC estimator, depending on the actual setting.

The covariate-specific AUC based on the estimated marker combination is comparable between the AAUC, LGT, and LZ estimators. In contrast, the covariate-specific AUCs based on the marker combination that maximizes the empirical AUC are much smaller compared to other methods.

The simulation results for the frequency-matched case-control study under models (6) and (7) are provided in Appendix F (Table 1) of the Supplementary Material. The pattern we observe for comparing different estimators is similar to those observed in cohort study.

III. Scenarios where the linear logistic regression model assumption is violated:

In this section we consider two scenarios where the risk model conditional on biomarker and covariate does not follow a linear logistic regression model, but the best marker combination is still a linear combination. First we consider a logistic regression model that includes a cubic effect of the linear biomarker combination:

$$\text{logit} \{P(Y=1|X, Z)\} = -2.5 + X_1 - 0.3X_2 + Z\{2.5(X_1 - 0.3X_2) + 2(X_1 - 0.3X_2)^3 + 1\}. \quad (8)$$

Here the disease risk is a monotone increasing function of $X_1 - 0.3X_2$ for each Z value, thus $X_1 - 0.3X_2$ is the best marker combination within each covariate stratum. We assume Z follows a Bernoulli distribution with mean 0.5 and, $(X_1, X_2)|Z$ follows a bivariate normal distribution with mean 0 and different covariances at each covariate level. When $Z=0$, X_1 and X_2 both have variance 0.25 and have a correlation 0.5; when $Z=1$, X_1 and X_2 are independent and have variances 0.25 and 1, respectively. This setting induces disease prevalences of 0.18 and 0.31 for $Z=0$ and $Z=1$, and an overall prevalence of 0.25.

We also consider another setting where $\text{logit} \{P(Y=1|X, Z)\} = 0.75Q_1 + 0.25Q_2$, where

$$\begin{aligned} \text{logit}\{Q_1\} &= -3 + (X_1 + 2X_2) + Z \left\{ 2(X_1 + 2X_2)^3 - 2 \right\}, \\ \text{logit}\{Q_2\} &= \sqrt{(X_1 + 2X_2)}/3, \end{aligned} \quad (9)$$

where Z follows a Bernoulli distribution with mean 0.5, and $(X_1, X_2)|Z$ follow truncated bivariate normal distributions with both variables truncated to be greater or equal to zero. The bivariate normal distribution before truncation has mean (0,0), variance 0.25 each, and correlation 0.5 for $Z = 0$, and has mean (-1,0), variance 0.25 and 1.25 each, and correlation 0 for $Z = 1$. The best marker combination in this setting is $X_1 + 2X_2$. This setting has disease prevalences 0.20 and 0.36 for $Z = 0$ and $Z = 1$, and an overall prevalence of 0.28. Under both models (8) and (9), the assumptions for LGT and LZ estimators are violated. **Under model (8) where true β is -0.3, we present the following transformed estimates of $\hat{\beta}$ for LGT and LZ:**

$$\hat{\beta}^* = \begin{cases} \hat{\beta} & \text{if } -3.3 \leq \hat{\beta} \leq 2.7 \\ 3.7 - 1/(\hat{\beta} - 1.7) & \text{if } \hat{\beta} > 2.7 \\ -4.3 - 1/(\hat{\beta} + 2.3) & \text{if } \hat{\beta} < -3.3 \end{cases}.$$

Under model (9) where true β is 2, we present the bounded estimates $\hat{\beta}^*$ for LGT and LZ as in Scenarios I and II.

Table 4 summarizes the results. Under model (8), the AAUC estimator of marker combinations is approximately unbiased, whereas bias is observed for the AUC, LGT, and LZ estimators. Classification performances for AAUC estimator is in general better than AUC, LGT, and LZ estimators. Covariate-adjustment appears to be particularly important in this setting, with the AAUC estimator tremendously outperforming the AUC estimator.

Under model (9), the proposed AAUC estimators again have better performance compared to all other estimators. **In particular, the AAUC estimator of marker combinations is approximately unbiased, while the AUC, LGT, and LZ estimators all have biases.** The classification performance of the AAUC estimators within each stratum is comparable or better than other estimators.

4. Data Analysis

We applied the proposed estimator of marker combinations to immune response data from the RV144 Thailand HIV vaccine trial. The RV144 trial was a community-based, randomized, placebo-controlled efficacy trial and enrolled 16,402 healthy men and women aged from 18 to 30 years in Rayong and Chon Buri provinces in Thailand. It was the first and only HIV vaccine trial to date to demonstrate positive vaccine efficacy in preventing HIV infection. [14].

In order to study the association between vaccine-induced immune-response variables and the risk of HIV infection, an immune response study was subsequently conducted among vaccine recipients in the RV144 trial, measuring peak immunogenicity data at week 26 (two

weeks after the final vaccination) [15, 16]. A case-control study was conducted to select cases and controls among vaccinees who were free of HIV-1 infection at the time of immune response measurement. Cases ($Y = 1$) were defined as vaccinees who were diagnosed of HIV-1 before 42 months; controls ($Y = 0$) who did not acquire infection over a follow-up period of 42 months were matched to cases in 5:1 ratio among strata defined by gender, the number of vaccinations received, and per-protocol status. The case-control sample consists of 41 vaccinated cases and 205 vaccinated controls [15].

Six primary immune assays (X) were measured in the immune correlates study, including 1) the binding of plasma IgA antibody to Env, 2) the avidity of IgG antibodies for Env, 3) antibody-dependent cellular cytotoxicity, 4) HIV-1 neutralizing antibodies, 5) the binding of IgG antibodies to variable regions 1 and 2 (V1V2) of the gp120 Env glycoprotein, and 6) the level of Env-specific CD4+ T cells. As gender is a potential risk factor for HIV infection risk, it is of interest to identify the best marker combinations adjusting for gender effect. Empirical estimators of gender-specific AUC for each individual correlate are presented in Appendix F (Table 2) of the Supplementary Material, where IgA antibodies binding to Env has the highest AUC. Our analyses thus focus on identifying the best way to combine the IgA response (X_1) with each of the other five primary immune assays (X_2). We present a gender-adjusted AAUC estimator for marker combinations using the Type 3 weights. For comparison, we also present the AUC, LGT, and LZ estimators of marker combinations. For each estimator of marker combinations, we estimated gender-specific AUC using random 3-fold cross-validation to correct for over-fitting bias. In particular, the data are randomly split into three folds stratified on disease status and gender, 2 folds for training and the remaining 1 fold for testing. Based on each random training subset, the best marker combinations are estimated using each estimator; the estimated rule is then applied to the random test subset to estimate the gender-specific AUC. The procedure is repeated 500 times with the average gender-specific AUC computed. Table 5 reports the estimate of marker combination β based on full data, its standard error based on 200 bootstrap resamples, the corresponding gender-specific AUCs estimated by cross-validation, and the weighted average of the gender-specific AUCs using Type 3 weight. In general, the largest differences in estimated marker combinations are observed between LGT or LZ estimators and the two AUC-related estimators (ie. AAUC and AUC). Small differences are observed between the AAUC and AUC estimators of marker combinations. In general, AAUC-based marker combinations achieve the largest weighted average of gender-specific classification performance. The only exception is in combining IgA and cellular toxicity response, where the AAUC estimator has the best AUC among females and a slightly smaller weighted average of gender-specific AUCs compared to the AUC estimator. The combination of IgA and IgG V1V2 based on AAUC estimator achieves the best cross-validated AAUC.

5. Concluding remarks

In this paper, we addressed the problem of adjusting for covariate effects in identifying the best marker combinations for disease classification and risk prediction. Our research builds upon and extends existing research that robustly estimates marker combinations by maximizing the empirical area under the ROC curve (AUC). Through constructed examples and simulation studies, we demonstrated, for the first time, that marker combinations that

maximize the AUC in each covariate stratum may not necessarily maximize the marginal AUC. Covariate adjustment is therefore essential when classification performance solely due to biomarkers is of interest, even when there is a common best marker combination across covariate strata.

In the early phase of biomarker development, it is desirable to identify a common biomarker combination that might be useful across covariate-strata; at the same time, one might want to allow for the threshold of the biomarker combination score to differ across covariate strata to account for covariate effects. This problem is similar in nature to the problem of identifying a common odds ratio across covariate-strata in logistic regression. To develop such a marker combination, we need a summary measure that characterizes the marker's impact if applied to a target population. The covariate-adjusted AUC provides a summary of the average performance of the biomarker combination score based on covariate-specific weights chosen to reflect the importance of having good classification accuracy in each covariate stratum. It is of interest to identify a marker combination that maximizes this quantity. In this paper, we proposed new estimators to adjust for covariate effects in combining markers by maximizing the nonparametric estimate of the area under the covariate-adjusted ROC curve (AAUC). When the best linear marker combination is the same across covariate strata, our proposed estimator aims to identify this common marker combination adjusting for "confounding" covariate effects. When the covariate is a "confounder" of the marker combination, i.e. the best marker combination is the same across covariate strata, the best marker combination that maximizes the covariate-adjusted AUC is the same as the best marker combination that maximizes the AUC within each covariate stratum. However, maximizing the covariate-adjusted AUC allows us to more efficiently estimate the marker combination than estimating the marker combination using data from only one particular covariate stratum. That being said, the application of our estimator does not require the assumption of a common best marker combination across covariate strata because the covariate-adjusted AUC we aim to maximize is itself a meaningful summary measure of a biomarker's performance in the population.

For various scenarios in which the best marker combination is the same linear combination across covariate strata, but the marker distributions differ across covariate strata, we compared the proposed AAUC estimator of marker combinations with an estimator that maximizes the empirical AUC, a logistic regression based estimator (LGT), and a parametric estimator (LZ) proposed by Liu and Zhou [5]. The AUC-based estimator that does not adjust for covariate effects can be severely biased with poor classification performance. When their model assumptions are satisfied, both the LGT and LZ estimators are asymptotically unbiased and can be more efficient than the AAUC estimator of marker combinations for large sample sizes, whereas the AAUC estimator can be more efficient for small to medium sample sizes, which we note are more frequently observed in practice. Moreover, the AAUC estimator is more robust than the LGT and LZ estimators; the latter two can be severely biased under misspecification of regression model assumptions. Therefore, the proposed AAUC maximization method provides a robust alternative to complement existing methods for identifying optimal marker combinations adjusting for covariate effects and has a strong potential to contribute to biomarker research in clinical practice.

In this paper we focused on a discrete covariate Z , but our methods could be extended to handle continuous Z using nonparametric smoothing techniques. For example, to compute the AAUC for the covariate-adjusted AUC curve as proposed in [4], one can compute

$$\widehat{AAUC}(\beta) = \frac{\sum_{i=1}^{n_D} \sum_{j \neq i} I(\beta^T X_{D_i} > \beta^T X_{H_j}) K\{(Z_{D_i} - Z_{H_j})/h\}}{\sum_{j \neq i} K\{(Z_{D_i} - Z_{H_j})/h\}}$$

where K can be a symmetric kernel function such as the Gaussian kernel and the Epanechnikov kernel and h is the bandwidth.

Finally, it is worth noting that combining biomarkers while adjusting for covariates is a related but different problem from the development of risk prediction rules combining biomarkers and covariates. The latter is also a problem of common interest in biomarker research, which focuses on the joint effect of biomarkers and covariates. In the early phase of biomarker development, researchers typically start by developing a combination score based on lab assays, which then need to be validated in different cohorts. Having a biomarker score adjusted for covariates allows one to understand the pure impact of the biomarker, which may later be generalized to different populations with different covariate distributions. Later, for the purpose of developing a best risk prediction rule to be applied to a population, we would want to combine the identified biomarkers with covariates such as individuals' demographics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work was supported by NIH awards R01 GM106177-01 and R01 GM54438. The authors thank the participants, investigators, and sponsors of the RV144 trials, specifically, the U.S. the U.S. Military HIV Research Program (MHRP); U.S. Army Medical Research and Materiel Command; NIAID; U.S. and Thai Components, Armed Forces Research Institute of Medical Science; Ministry of Public Health, Thailand; Mahidol University; Sanofi Pasteur; and Global Solutions for Infectious Diseases. The views expressed are those of the authors and should not be construed to represent the positions of the U.S. Army or the Department of Defense.

References

1. Baker S. The central role of receiver operating characteristic (roc) curves in evaluating tests for the early detection of cancer. *J Nat Cancer Inst.* 2003; 95:511–5. [PubMed: 12671018]
2. Pepe M, Etzioni R, Feng Z, Potter J, Thompson M, Thornquist M, Winget M, Yatsui Y. Phases of biomarker development for early detection of cancer. *Nat Cancer Inst.* 2001; 93:105461.
3. Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press; Oxford: 2003.
4. Janes H, Pepe M. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika.* 2009; 96:371–382. [PubMed: 22822245]
5. Liu D, Zhou XH. Roc analysis in biomarker combination with covariate adjustment. *Acad Radiol.* 2013; 20:874–82. [PubMed: 23747153]
6. Pepe M, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics.* 2006; 62:221–229. [PubMed: 16542249]

7. Ma S, Huang J. Combining multiple markers for classification using roc. *Biometrics*. 2007; 63:751–757. [PubMed: 17825007]
8. McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics*. 2002; 58:657–664. [PubMed: 12230001]
9. Pepe M, Thompson M. Combining diagnostic test results to increase accuracy. *Biostatistics*. 2000; 1:123–140. [PubMed: 12933515]
10. Han AK. Non-parametric analysis of a generalized regression model. *Journal of Econometrics*. 1987; 35:303–316.
11. Sherman RP. The limiting distribution of the maximum rank correlation estimator. *Econometrika*. 1993; 61:123–137.
12. Liu D, Zhou X. Semiparametric estimation of the covariate-specific roc curve in presence of ignorable verification bias. *Biometrics*. 2011; 67:906–916. [PubMed: 21361890]
13. Nolan D, Pollard D. U-processes: Rates of convergence. *The Annals of Statistics*. 1987; 15:780–799.
14. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, et al. Vaccination with alvac and aidsvac to prevent hiv-1 infection in thailand. *The new england journal of medicine*. 2009; 361:2209–2220. [PubMed: 19843557]
15. Haynes B, Gilbert P, McElrath J, et al. Immune-correlates analysis of an hiv-1 vaccine efficacy trial. *N Engl J Med*. 2012; 366:1275–1286. [PubMed: 22475592]
16. Rolland M, Gilbert P. Evaluating immune correlates in hiv type 1 vaccine efficacy trials: what rv144 may provide. *AIDS Res Hum Retroviruses*. 2012; 28:400–4. [PubMed: 21902593]

Table 1

Simulation results for comparison of weights in case-control and cohort samples.

Study design	n	p _{Z=1}	(P _{Y=1 Z=0} , P _{Y=1 Z=1})	Type 1					Type 2					Type 3				
				$\hat{\beta}$	SD	SE	$\overline{SE}$	CI	$\hat{\beta}$	SD	SE	$\overline{SE}$	CI	$\hat{\beta}$	SD	SE	$\overline{SE}$	CI
CH	600	0.5	(0.2,0.2)	1.05	0.39	0.38	0.38	0.96	1.06	0.39	0.38	0.38	0.96	1.05	0.39	0.38	0.38	0.96
				(0.14,0.26)	1.05	0.40	0.38	0.38	0.95	1.06	0.41	0.39	0.39	0.95	1.06	0.40	0.39	0.38
		0.3	(0.2,0.2)	1.05	0.41	0.39	0.39	0.96	1.05	0.39	0.38	0.38	0.96	1.05	0.39	0.38	0.38	0.96
				(0.14,0.34)	1.06	0.42	0.40	0.40	0.95	1.06	0.42	0.39	0.39	0.95	1.06	0.40	0.39	0.39
	2400	0.5	(0.2,0.2)	1.01	0.20	0.22	0.21	0.95	1.01	0.20	0.22	0.21	0.95	1.01	0.20	0.22	0.21	0.95
				(0.14,0.26)	1.02	0.22	0.24	0.22	0.95	1.01	0.21	0.23	0.22	0.95	1.01	0.20	0.22	0.21
		0.3	(0.2,0.2)	1.01	0.21	0.22	0.22	0.95	1.01	0.20	0.22	0.21	0.95	1.01	0.20	0.22	0.21	0.95
				(0.14,0.34)	1.01	0.22	0.24	0.23	0.95	1.01	0.22	0.23	0.22	0.95	1.02	0.21	0.23	0.22
CC	600	0.5	(0.2,0.2)	1.04	0.35	0.35	0.34	0.97	1.04	0.35	0.35	0.34	0.97	1.04	0.35	0.35	0.34	0.97
				(0.14,0.26)	1.03	0.35	0.35	0.35	0.96	1.04	0.35	0.35	0.35	0.96	1.04	0.35	0.35	0.35
		0.3	(0.2,0.2)	1.04	0.36	0.35	0.35	0.95	1.05	0.35	0.35	0.35	0.96	1.05	0.35	0.35	0.35	0.96
				(0.14,0.34)	1.04	0.35	0.35	0.35	0.95	1.05	0.35	0.35	0.35	0.95	1.05	0.35	0.35	0.35
	2400	0.5	(0.2,0.2)	1.01	0.17	0.19	0.18	0.96	1.01	0.17	0.19	0.18	0.96	1.01	0.17	0.19	0.18	0.96
				(0.14,0.26)	1.01	0.17	0.19	0.18	0.95	1.01	0.17	0.19	0.18	0.96	1.01	0.17	0.19	0.18
		0.3	(0.2,0.2)	1.01	0.17	0.19	0.19	0.96	1.01	0.16	0.19	0.18	0.96	1.01	0.16	0.19	0.18	0.96
				(0.14,0.34)	1.01	0.17	0.19	0.18	0.95	1.01	0.17	0.19	0.18	0.95	1.01	0.17	0.19	0.18

SD refers to the sample standard deviation; SE refers to the mean bootstrap standard error; $\overline{SE}$ refers to the median bootstrap standard error; CI refers to the 95% confidence interval; CH refers to the cohort study sample; CC refers to the case-control sample

Table 2Simulation results for comparison of the AAUC, AUC, and logistic regression estimators under model (6) with true $\beta = 2$

$P_{z=1}$	$P_{Y=1 Z}$	Estimators	$n=500$				$n=2000$			
			$\hat{\beta}$		$AUC_z(\hat{\beta})$		$\hat{\beta}$		$AUC_z(\hat{\beta})$	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.5	(0.2,0.2)	AAUC	2.18	0.86	0.824	0.0017	2.07	0.48	0.825	0.0005
		AUC	0.16	0.17	0.779	0.0161	0.13	0.07	0.778	0.0084
		LGT	2.24	0.99	0.824	0.0014	2.05	0.42	0.825	0.0004
		LZ	2.25	0.97	0.824	0.0025	2.05	0.42	0.825	0.0004
	(0.14,0.26)	AAUC	2.19	0.89	0.824	0.0019	2.07	0.50	0.825	0.0005
		AUC	-0.09	0.10	0.740	0.0207	-0.10	0.05	0.742	0.0102
		LGT	2.25	1.02	0.824	0.0014	2.04	0.43	0.825	0.0004
		LZ	2.25	0.99	0.824	0.0014	2.05	0.43	0.825	0.0004
0.3	(0.2,0.2)	AAUC	2.20	0.87	0.824	0.0018	2.09	0.49	0.825	0.0005
		AUC	0.32	0.36	0.789	0.0187	0.25	0.11	0.790	0.0091
		LGT	2.24	0.99	0.824	0.0014	2.05	0.42	0.825	0.0004
		LZ	2.25	0.97	0.824	0.0014	2.05	0.42	0.825	0.0004
	(0.14,0.34)	AAUC	2.23	0.90	0.824	0.0018	2.10	0.52	0.825	0.0005
		AUC	-0.17	0.11	0.722	0.0269	-0.18	0.05	0.723	0.0127
		LGT	2.25	1.04	0.824	0.0026	2.05	0.44	0.825	0.0004
		LZ	2.25	1.00	0.824	0.0026	2.05	0.44	0.825	0.0004

Table 3

Simulation results for comparison of the AAUC, AUC, logistic regression, and LZ estimators under model (7) with true $\beta=2$

$P_{z=1}$	$P_{Y=1 Z}$	Estimators	n=500						n=2000					
			$\hat{\beta}$		$AUC_0(\hat{\beta})$		$AUC_1(\hat{\beta})$		$\hat{\beta}$		$AUC_0(\hat{\beta})$		$AUC_1(\hat{\beta})$	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.5	(0.2,0.2)	AAUC	2.20	0.91	0.772	0.0017	0.814	0.0019	2.08	0.54	0.772	0.0005	0.815	0.0006
		AUC	0.20	0.23	0.733	0.0154	0.772	0.0170	0.16	0.08	0.733	0.0078	0.772	0.0086
		LGT	2.29	1.10	0.772	0.0031	0.814	0.0034	2.06	0.48	0.773	0.0005	0.815	0.0005
		LZ	2.30	1.09	0.772	0.0025	0.814	0.0027	2.07	0.48	0.773	0.0005	0.815	0.0005
	(0.14,0.26)	AAUC	2.20	0.93	0.771	0.0020	0.814	0.0021	2.08	0.54	0.772	0.0005	0.815	0.0006
		AUC	-0.10	0.11	0.696	0.0201	0.731	0.0228	-0.10	0.05	0.697	0.0098	0.732	0.0111
		LGT	2.28	1.12	0.772	0.0036	0.814	0.0040	2.06	0.48	0.773	0.0005	0.815	0.0005
		LZ	2.29	1.09	0.772	0.0031	0.814	0.0033	2.06	0.47	0.773	0.0004	0.815	0.0005
0.5	(0.2,0.2)	AAUC	2.21	0.94	0.771	0.0019	0.968	0.0012	2.10	0.57	0.772	0.0006	0.969	0.0004
		AUC	-1.57	0.93	0.659	0.0369	0.840	0.0565	-1.59	0.16	0.654	0.0150	0.838	0.0248
		LGT	2.12	2.02	0.762	0.0241	0.960	0.0237	2.32	1.20	0.771	0.0042	0.968	0.0038
		LZ	2.07	2.05	0.761	0.0253	0.959	0.0249	2.34	1.26	0.771	0.0047	0.968	0.0042
	(0.14,0.26)	AAUC	2.20	0.92	0.771	0.0020	0.968	0.0012	2.09	0.54	0.772	0.0006	0.969	0.0003
		AUC	-3.33	1.51	0.714	0.0249	0.918	0.0281	-2.95	1.03	0.714	0.0158	0.920	0.0168
		LGT	2.39	1.56	0.770	0.0101	0.967	0.0096	2.17	0.76	0.772	0.0009	0.969	0.0005
		LZ	2.36	1.66	0.768	0.0131	0.966	0.0125	2.19	0.83	0.772	0.0010	0.969	0.0006

Simulation results for comparison of the AAUC, AUC, logistic regression, and LZ estimators under models (8) and (9). Results are based on 2000 Monte-Carlo simulations.

Table 4

	n	Estimator	$\hat{\beta}$		$AUC_0(\hat{\beta})$		$AUC_1(\hat{\beta})$	
			Mean	SD	Mean	SD	Mean	SD
Model (8)	300	AAUC	-0.302	0.064	0.6156	0.00375	0.8969	0.00175
		AUC	-1.069	0.361	0.5487	0.02989	0.8273	0.02971
		LGT	-0.412	0.095	0.6082	0.00727	0.8915	0.00788
		LZ	-0.429	0.096	0.6069	0.00749	0.8902	0.00841
	500	AAUC	-0.305	0.057	0.6155	0.00335	0.8973	0.00154
		AUC	-1.080	0.282	0.5464	0.02392	0.8248	0.02310
		LGT	-0.414	0.070	0.6084	0.00524	0.8922	0.00595
		LZ	-0.432	0.070	0.6070	0.00543	0.8908	0.00641
	2000	AAUC	-0.301	0.030	0.6160	0.00183	0.8984	0.00059
		AUC	-1.038	0.135	0.5485	0.01269	0.8260	0.01173
Model (9)	1000	LGT	-0.406	0.034	0.6091	0.00231	0.8937	0.00269
		LZ	-0.425	0.034	0.6079	0.00239	0.8923	0.00307
	2000	AAUC	2.008	0.799	0.6135	0.00118	0.8530	0.00003
		AUC	10.144	1.641	0.6135	0.00010	0.8528	0.00002
	4000	LGT	1.333	3.640	0.5863	0.02477	0.8522	0.00058
		LZ	2.148	3.607	0.5926	0.02456	0.8524	0.00058
	8000	AAUC	2.044	0.770	0.6136	0.00108	0.8530	0.00004
		AUC	10.308	1.652	0.6135	0.00010	0.8528	0.00002
	16000	LGT	1.433	3.796	0.5863	0.02445	0.8522	0.00057
		LZ	2.640	3.709	0.5947	0.02391	0.8524	0.00056

Data analysis results for combining two biomarkers: $X_1 = \text{IgA antibody binding to Env}$. Reported are estimate marker combinations and corresponding gender-specific AUC values estimated by cross-validation.

Table 5

X_2	Estimator	$\hat{\beta}$	$\widehat{SE}(\hat{\beta})^*$	$AUC_{female}(\hat{\beta})$	$AUC_{male}(\hat{\beta})$	$AAUC(\hat{\beta})$
Level of IgG antibodies to HIV-1 Env V1V2	AAUC	-1.376	0.355	0.6209	0.6530	0.6415
	AUC	-1.288	0.343	0.6216	0.6521	0.6412
	LGT	-0.775	0.385	0.6168	0.6378	0.6303
	LZ	-0.772	0.443	0.6174	0.6336	0.6276
Avidity of IgG antibodies for Env	AAUC	-0.361	0.173	0.6140	0.6205	0.6182
	AUC	-0.356	0.173	0.6121	0.6196	0.6169
	LGT	-0.353	0.399	0.6009	0.6035	0.6026
	LZ	-0.366	0.674	0.6015	0.6042	0.6032
Cellular toxicity	AAUC	-0.180	0.171	0.6291	0.5978	0.6090
	AUC	-0.184	0.170	0.6275	0.6019	0.6110
	LGT	0.223	0.422	0.6111	0.5793	0.5907
	LZ	0.142	1.936	0.6078	0.5831	0.5923
CD4+ T cells	AAUC	-0.322	0.210	0.6126	0.6237	0.6198
	AUC	-0.099	0.249	0.6099	0.6171	0.6145
	LGT	-0.061	0.404	0.6042	0.5987	0.6007
	LZ	-0.146	0.571	0.5979	0.5996	0.5990
Neutralizing antibodies	AAUC	-0.274	0.192	0.6164	0.6100	0.6123
	AUC	-0.012	0.212	0.6156	0.6089	0.6113
	LGT	0.066	0.467	0.6101	0.5934	0.5994
	LZ	-0.017	1.020	0.6002	0.5902	0.5939

$\widehat{SE}(\hat{\beta})^*$: standard error of $\hat{\beta}$ estimated by 200 bootstrap resamples.