7-26-2017

Combining Biomarkers by Maximizing the True Positive Rate for a Fixed False Positive Rate

Allison Meisner
University of Washington, Seattle, meisnera@uw.edu

Marco Carone
Department of Biostatistics, University of Washington, mcarone@uw.edu

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

Kathleen F. Kerr
University of Washington, katiek@u.washington.edu

Combining Biomarkers by Maximizing the True Positive Rate for a Fixed False Positive Rate

A. Meisner*[1], M. Carone†[1], M. S. Pepe‡[2], and K. F. Kerr§[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.
[2]Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.

Abstract

Biomarkers abound in many areas of clinical research, and often investigators are interested in combining them for diagnosis, prognosis and screening. In many applications, the true positive rate for a biomarker combination at a prespecified, clinically acceptable false positive rate is the most relevant measure of predictive capacity. We propose a distribution-free method for constructing biomarker combinations by maximizing the true positive rate while constraining the false positive rate. Theoretical results demonstrate good operating characteristics for the resulting combination. In simulations, the biomarker combination provided by our method demonstrated improved operating characteristics in a variety of scenarios when compared with more traditional methods for constructing combinations.

Keywords: Biomarker; Combination; Sensitivity; True positive rate.

1 Introduction

As the number of available biomarkers has grown, so has the interest in combining them for the purposes of diagnosis, prognosis, and screening. In the past decade, much work has been done to construct biomarker combinations by targeting measures of performance, including those related to the receiver operating characteristic, or ROC, curve. This is in contrast to more traditional methods that construct biomarker combinations by optimizing global fit criteria, such as the maximum likelihood approach. While methods to construct both linear and nonlinear combinations have been proposed, linear biomarker combinations are more commonly used than nonlinear combinations, primarily due to their greater interpretability and ease of construction (Wang & Chang, 2011; Hsu & Hsueh, 2013).

Although the area under the ROC curve, the AUC, is arguably the most popular way to summarize the ROC curve, there is often interest in identifying biomarker combinations with maximum true positive rate, the proportion of correctly classified diseased individuals, while setting the false positive rate, the proportion of incorrectly classified nondiseased individuals, at some clinically acceptable level. It is common practice among applied researchers to construct linear biomarker combinations using logistic regression, and then calculate the true positive rate for the prespecified false positive rate, e.g., Moore *et al.* (2008). While much work has been done to construct biomarker combinations by maximizing the AUC or the partial AUC, none of these methods directly target the true positive rate for a specified false positive rate.

We propose a distribution-free method for constructing linear biomarker combinations by maximizing the true positive rate while constraining the false positive rate. We demonstrate desirable theoretical properties

*meisnera@uw.edu
†mcarone@uw.edu
‡mspepe@uw.edu
§katiek@uw.edu

of the resulting combination, and provide empirical evidence of good small-sample performance through simulations. To illustrate the use of our method, we consider data from a prospective study of diabetes mellitus in 532 adult women with Pima Indian heritage (Smith et al., 1988). Several variables were measured for each participant, and criteria from the World Health Organization were used to identify women with diabetes. A primary goal of the study was to predict the onset of diabetes within five years.

2 Background

2.1 ROC curve and related measures

The ROC curve provides a means to evaluate the ability of a biomarker or, equivalently, a biomarker combination Z to identify individuals who have or will experience a binary outcome D. For example, in the diagnostic setting, D may denote the presence or absence of disease and Z may be used to identify individuals with the disease. The ROC curve provides information about how well the biomarker discriminates between individuals who have or will experience the outcome, that is, the cases, and individuals who do not have or will not experience the outcome, that is, the controls (Pepe, 2003). Mathematically, if larger values of Z are more indicative of having or experiencing the outcome, for each threshold δ we can define the true positive rate as pr$(Z > \delta \mid D = 1)$ and the false positive rate as pr$(Z > \delta \mid D = 0)$ (Pepe, 2003). For a given δ, the true positive rate is also referred to as the sensitivity, and one minus the specificity equals the false positive rate (Pepe, 2003). The ROC curve is a plot of the true positive rate versus the false positive rate as δ ranges over all possible values; as such, it is non-decreasing and takes values in the unit square (Pepe, 2003). A perfect biomarker has an ROC curve that reaches the upper left corner of the unit square, and a useless biomarker has an ROC curve on the 45-degree line (Pepe, 2003).

The most common summary of the ROC curve is the AUC, the area under the ROC curve. The AUC ranges between 0.5 for a useless biomarker and 1 for a perfect biomarker (Pepe, 2003). The AUC has a probabilistic interpretation: it is the probability that the biomarker value for a randomly chosen case is larger than that for a randomly chosen control, assuming that higher biomarker values are more indicative of having or experiencing the outcome (Pepe, 2003). Both the ROC curve and the AUC are invariant to monotone increasing transformations of the biomarker Z (Pepe, 2003).

The AUC summarizes the entire ROC curve, but in many situations, it may be more appropriate to only consider certain false positive rate values. For example, screening tests require a very low false positive rate, while diagnostic tests for fatal diseases may allow for a slightly higher false positive rate if the corresponding true positive rate is very high (Hsu & Hsueh, 2013). This consideration led to the development of the partial AUC, the area under the ROC curve over some range (t_0, t_1) of false positive rate values (Pepe, 2003). Rather than considering a range of false positive rate values, there may be interest in fixing the false positive rate at a single value, determining the corresponding threshold δ, and evaluating the true positive rate for that threshold. As opposed to the AUC and the partial AUC, this method returns a single classifier, or decision rule, which may be appealing to researchers seeking a tool for clinical decision-making.

2.2 Biomarker combinations

Many methods to combine biomarkers have been proposed, and they can generally be divided into two categories. The first includes indirect methods that seek to optimize a measure other than the performance measure of interest, while the second category includes direct methods that optimize the target performance measure. We focus on the latter.

Targeting the entire ROC curve, that is, constructing a combination that produces an ROC curve that dominates the ROC curve for all other linear combinations at all points, is very challenging and can generally only be done under special circumstances. Su & Liu (1993) demonstrated that, when the vector X of biomarkers has a multivariate normal distribution conditional on D with proportional covariance matrices, it is possible to identify the linear combination that maximizes the true positive rate uniformly over the entire range of false positive rates (Su & Liu, 1993). If the D-specific covariance matrices are equal, this optimal linear combination dominates not just every other linear combination, but also every nonlinear combination. This follows from the fact that in this case, the linear logistic model stipulating that $\text{logit}\{\text{pr}(D = 1|X = x)\} = \theta^\top x$ holds for some p-dimensional θ, where p is the dimension of X (McIntosh & Pepe, 2002). If the covariance matrices are proportional but not equal, the likelihood ratio is a nonlinear function of the biomarkers, as shown in the Supplementary Material for $p = 2$, and the optimal biomarker combination with respect to the ROC curve is nonlinear (McIntosh & Pepe, 2002).

In general, there is no linear combination that dominates all others in terms of the true positive rate over the entire range of false positive rates (Su & Liu, 1993; Anderson & Bahadur, 1962). Thus, methods to optimize the AUC have been proposed. When the biomarkers are conditionally multivariate normal with nonproportional covariance matrices, Su & Liu (1993) gave an explicit form for the best linear combination with respect to the AUC. Others have targeted the AUC without any assumption on the distribution of the biomarkers; many of these methods rely on smooth approximations to the empirical AUC, which involves indicator functions (Ma & Huang, 2007; Fong *et al.*, 2016; Lin *et al.*, 2011).

Acknowledging that often only a range of false positive rate values is of interest clinically, methods have been proposed to target the partial AUC for some false positive rate range (t_0, t_1). Some methods make parametric assumptions about the joint distribution of the biomarkers (Yu & Park, 2015; Hsu & Hsueh, 2013) while others do not (Wang & Chang, 2011; Komori & Eguchi, 2010). The latter group of methods generally use a smooth approximation to the partial AUC, similar to some of the methods that aim to maximize the AUC (Wang & Chang, 2011; Komori & Eguchi, 2010). One challenge faced in partial AUC maximization is that for narrow intervals, that is, when t_0 is close to t_1, the partial AUC is often very close to 0, which can make optimization difficult (Hsu & Hsueh, 2013).

In recent years, the AUC has been heavily criticized because it does not directly measure the clinical impact of using the biomarker or biomarker combination: while the AUC can be interpreted probabilistically in terms of case-control pairs, patients do not present to clinicians in randomly selected case-control pairs (Pepe & Janes, 2013). Moreover, the AUC includes, and may in fact be dominated by, regions of the ROC curve that are not clinically relevant (Pepe & Janes, 2013). Measures such as the partial AUC were proposed to address this shortcoming, but the partial AUC does not directly correspond to a decision rule, making clinical implementation challenging. Thus, there is growing interest in evaluating biomarkers and biomarker combinations by considering the true positive rate at a fixed, clinically acceptable false positive rate.

Some work in constructing biomarker combinations by maximizing the true positive rate has been done for conditionally multivariate normal biomarkers. In this setting, procedures for constructing a linear combination that maximizes the true positive rate for a fixed false positive rate have been considered (Anderson & Bahadur, 1962; Gao *et al.*, 2008). Methods have also been proposed to construct linear combinations by maximizing the true positive rate for a range of false positive rate values (Liu *et al.*, 2005). The major disadvantage of this approach is that the range of false positive rate values over which the fitted combination is optimal may depend on the combination itself; that is, the range of false positive rate values may be determined by the combination and so may not be fixed in advance (Liu *et al.*, 2005). Baker (2000) proposed a flexible nonparametric method for combining biomarkers by optimizing the ROC curve over a narrow target region of false positive rate values, but this method is not well-suited to situations in which more than a few biomarkers are to be combined.

An important benefit of constructing linear biomarker combinations by targeting the performance measure of interest is that the performance of the combination will be at least as good as the performance of the individual biomarkers (Pepe et al., 2006). Indeed, several authors have recommended matching the objective function to the performance measure by constructing biomarker combinations by optimizing the relevant measure of performance (Hwang *et al.*, 2013; Liu *et al.*, 2005; Wang & Chang, 2011; Ricamato & Tortorella, 2011). To that end, we propose a distribution-free method to construct biomarker combinations by maximizing the true positive rate for a given false positive rate.

3 Methodology

3.1 Description

We will assume a non-trivial disease prevalence throughout, $\text{pr}(D = 1) \in (0, 1)$. Cases will be denoted by the subscript 1, and controls will be denoted by the subscript 0. Let X_{1i} denote the vector of biomarkers for the i^{th} case, and let X_{0j} denote the vector of biomarkers for the j^{th} control.

We propose constructing a linear biomarker combination of the form $\theta^\top X$ for a p-dimensional X by maximizing the true positive rate when the false positive rate is below some prespecified, clinically acceptable value t. We define the true and false positive rates for a given X as a function of θ and δ:

$$\text{TPR}(\theta, \delta) = \text{pr}(\theta^\top X > \delta | D = 1), \quad \text{FPR}(\theta, \delta) = \text{pr}(\theta^\top X > \delta | D = 0).$$

Since the true and false positive rates for a given combination θ and threshold δ are invariant to scaling of the parameters (θ, δ), we must restrict (θ, δ) to ensure identifiability. Specifically, we constrain $||\boldsymbol{\theta}|| = 1$ as in Fong *et al.* (2016). For any fixed $t \in (0, 1)$, we can consider

$$(\theta_t, \delta_t) \in \underset{(\theta,\delta)\in\Omega_t}{\arg\max}\, \text{TPR}(\theta, \delta),$$

where $\Omega_t = \{\theta \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\theta|| = 1, \text{FPR}(\theta, \delta) \leq t\}$. This provides the optimal combination θ_t and threshold δ_t. We define (θ_t, δ_t) to be an element of $\arg\max_{(\theta,\delta)\in\Omega_t} \text{TPR}(\theta, \delta)$, where $\arg\max_{(\theta,\delta)\in\Omega_t} \text{TPR}(\theta, \delta)$ may be a set.

Of course, in practice, the true and false positive rates are unknown, so θ_t and δ_t cannot be computed. We can replace these unknowns by their empirical estimates,

$$\text{T}\hat{\text{P}}\text{R}_{n_1}(\theta, \delta) = \frac{1}{n_1} \sum_{i=1}^{n_1} 1(\theta^\top X_{1i} > \delta), \quad \text{F}\hat{\text{P}}\text{R}_{n_0}(\theta, \delta) = \frac{1}{n_0} \sum_{j=1}^{n_0} 1(\theta^\top X_{0j} > \delta),$$

where n_1 is the number of cases and n_0 is the number of controls, giving the total sample size $n = n_1 + n_0$. We can then define

$$(\hat{\theta}_t, \hat{\delta}_t) \in \underset{(\theta,\delta)\in\hat{\Omega}_{t,n_0}}{\arg\max}\; \text{T}\hat{\text{P}}\text{R}_{n_1}(\theta, \delta)$$

where $\hat{\Omega}_{t,n_0} = \{\theta \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\theta|| = 1, \text{F}\hat{\text{P}}\text{R}_{n_0}(\theta, \delta) \leq t\}$. It is possible to conduct a grid search over (θ, δ) to perform this constrained optimization, though this becomes computationally demanding when combining more than two biomarkers.

Furthermore, since the objective function involves indicator functions, it is not a smooth function of the parameters (θ, δ). Derivative-based methods therefore cannot be readily used. However, smooth approximations to indicator functions exist and have been used for AUC maximization (Ma & Huang, 2007; Fong *et al.*, 2016; Lin *et al.*, 2011). One such smooth approximation is $1(w > 0) \approx \Phi(w/h)$, where Φ is the standard normal distribution function, and h is a tuning parameter representing the trade-off between approximation accuracy and estimation feasibility such that h tends to zero as the sample size grows (Lin *et al.*, 2011). We can use this smooth approximation to implement the method described above, writing the smooth approximations to the empirical true and false positive rates as

$$\text{T}\tilde{\text{P}}\text{R}_{n_1}(\theta, \delta) = \frac{1}{n_1} \sum_{i=1}^{n_1} \Phi\left(\frac{\theta^\top X_{1i} - \delta}{h}\right), \quad \text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta, \delta) = \frac{1}{n_0} \sum_{j=1}^{n_0} \Phi\left(\frac{\theta^\top X_{0j} - \delta}{h}\right).$$

Thus, we propose to compute

$$(\tilde{\theta}_t, \tilde{\delta}_t) \in \underset{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}}{\arg\max}\; \text{T}\tilde{\text{P}}\text{R}_{n_1}(\theta, \delta), \tag{1}$$

where $\tilde{\Omega}_{t,n_0} = \{\theta \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\theta|| = 1, \widetilde{\text{FPR}}_{n_0}(\theta, \delta) \leq t\}$. We can obtain $(\tilde{\theta}_t, \tilde{\delta}_t)$ by using gradient-based methods that incorporate the constraints imposed by $\tilde{\Omega}_{t,n_0}$, such as Lagrange multipliers, for example. Estimation can be accomplished with existing software, such as the `Rsolnp` package in `R`. Details related to implementation, including the choice of tuning parameter h, are discussed below.

3.2 Asymptotic properties

We present a theorem establishing that, under certain conditions, the combination obtained by optimizing the smooth approximation to the empirical true positive rate while constraining the smooth approximation to the empirical false positive rate has desirable operating characteristics. In particular, its false positive rate is bounded almost surely by the acceptable level t in large samples. In addition, its true positive rate converges almost surely to the supremum of the true positive rate over the set where the false positive rate is constrained.

Rather than enforcing $(\tilde{\theta}_t, \tilde{\delta}_t)$ to be a strict maximizer, in the theoretical study below, we allow it to be a near-maximizer of $\widetilde{\text{TPR}}_{n_1}(\theta, \delta)$ within $\tilde{\Omega}_{t,n_0}$ in the sense that

$$\widetilde{\text{TPR}}_{n_1}(\tilde{\theta}_t, \tilde{\delta}_t) \geq \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\text{TPR}}_{n_1}(\theta, \delta) - a_n \ ,$$

where a_n is a decreasing sequence of positive real numbers tending to zero. This provides some flexibility to accommodate situations in which a strict maximizer either does not exist or is numerically difficult to identify. In practice, a_n would be chosen to be as small as pragmatically feasible.

Before stating our key theorem, we give the following conditions.

Condition 1: Observations are randomly sampled conditional on disease status D, and the group sizes tend to infinity proportionally, in the sense that $n = n_1 + n_0 \to \infty$ and $n_1/n_0 \to \rho \in (0, 1)$.

Condition 2: For each $d \in \{0, 1\}$, observations X_{di}, $i = 1, 2, \ldots, n_d$, are independent and identically distributed p-dimensional random vectors with distribution function F_d.

Condition 3: For each $d \in \{0, 1\}$, no proper linear subspace $S \subset \mathbb{R}^p$ is such that $\text{pr}(X \in S \mid D = d) = 1$.

Condition 4: For each $d \in \{0, 1\}$, the distribution and quantile functions of $\theta^\top X$ given $D = d$ are globally Lipschitz continuous uniformly over $\theta \in \mathbb{R}^p$ such that $||\theta|| = 1$.

Condition 5: The map $(\theta, \delta) \mapsto \text{TPR}(\theta, \delta)$ is globally Lipschitz continuous over $\Omega = \{\theta \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\theta|| = 1\}$.

Theorem 1. *Under conditions (1)–(5), for every fixed* $t \in (0, 1)$*, we have that (a)* $\limsup_n \text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) \leq t$ *almost surely; and (b)* $|\text{TPR}(\tilde{\theta}_t, \tilde{\delta}_t) - \sup_{(\theta,\delta)\in\Omega_t} \text{TPR}(\theta, \delta)|$ *tends to zero almost surely.*

The proof of Theorem 1 is given in the Appendix. The proof relies on two lemmas, which are given in the Appendix. Lemma A1 demonstrates almost sure convergence to zero of the difference between the supremum of a function over a fixed set and the supremum of the function over a stochastic set that converges to the fixed set in an appropriate sense. Lemma A2 establishes the almost sure uniform convergence to zero of the difference between the false positive rate and the smooth approximation to the empirical false positive rate and the difference between the true positive rate and the smooth approximation to the empirical true positive rate. The proof of Theorem 1 then demonstrates that Lemma A1 holds for the relevant function and sets, relying in part on the conclusions of Lemma A2. The conclusions of Lemmas A1 and A2 are then used to demonstrate the claims of Theorem 1.

3.3 Implementation details

In order to implement these methods, certain considerations must first be addressed, including the choice of tuning parameter h and starting values $(\tilde{\theta}, \tilde{\delta})$ for the optimization routine. In using similar methods to maximize the AUC, Lin *et al.* (2011) proposed using $h = \tilde{\sigma} n^{-1/3}$, where $\tilde{\sigma}$ is the sample standard error of $\tilde{\theta}^\top X$. In simulations, we considered both $h = \tilde{\sigma} n^{-1/3}$ and $h = \tilde{\sigma} n^{-1/2}$ and found that using the latter had little impact on the convergence of the optimization routine. Thus, we use $h = \tilde{\sigma} n^{-1/2}$. We must also identify initial values $(\tilde{\theta}, \tilde{\delta})$ for our procedure. As done in Fong *et al.* (2016), we use normalized estimates from robust logistic regression, which is described in greater detail below. Based on this initial value $\tilde{\theta}$, we choose $\tilde{\delta}$ such that $\widetilde{\text{FPR}}_{n_0}(\tilde{\theta}, \tilde{\delta}) = t$.

Finally, we have also found that when $\widetilde{\text{FPR}}_{n_0}$ is bounded by t, the performance of the optimization routine can be poor. Thus, we introduce another tuning parameter, α, which allows for a small amount of relaxation in the constraint on the smooth approximation to the empirical false positive rate, imposing instead $\widetilde{\text{FPR}}_{n_0}(\theta, \delta) \leq t + \alpha$. Since the effective sample size for the smooth approximation to the empirical false positive rate is n_0, we chose to scale α with n_0, and have found $\alpha = 1/(2n_0)$ to work well in simulations. Other values of α may give combinations with better performance and could be considered.

Our method does not require limiting the number of biomarkers considered, although the risk of overfitting is expected to grow as the number of biomarkers increases relative to the sample size. In addition, our method does not impose constraints on the distribution of the biomarkers that can be included, except for weak conditions that allow us to establish its large-sample properties. An R package including code to implement our method, `maxTPR`, will be publicly available.

4 Simulations

Fong *et al.* (2016) suggest that the presence of outliers may lead to diminished performance of likelihood-based methods, while AUC-based methods may be less affected since the AUC is a rank-based measure. This feature would be expected to extend to the true and false positive rates, which are also rank-based measures. We consider simulations with and without outliers in the data-generating distribution, and simulate data under a model similar to that used by Fong *et al.* (2016). We consider two biomarkers X_1 and X_2 constructed as

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (1 - \Delta) \times Z_0 + \Delta \times Z_1$$

and D is then simulated as a Bernoulli random variable with success probability $f\left\{\beta_0 + 4X_1 - 3X_2 - 0.8(X_1 - X_2)^3\right\}$, where Δ is a Bernoulli random variable with success probability $\pi = 0.05$ when outliers are simulated and $\pi = 0$ otherwise, Z_0 and Z_1 are independent bivariate normal random variables with mean zero and respective covariance matrices

$$0.2 \times \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \; 2 \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We consider two f functions: $f_1(v) = \text{expit}(v) = e^v/(1 + e^v)$ and a piecewise logistic function,

$$f_2(v) = 1(v < 0) \times \frac{1}{1 + e^{-v/3}} + 1(v \geq 0) \times \frac{1}{1 + e^{-3v}}.$$

We vary β_0 to reflect varying prevalences, with a prevalence of approximately 50–60% for $\beta_0 = 0$, 16–18% for $\beta_0 < 0$, and 77–82% for $\beta_0 > 0$. We considered $t = 0.05$, 0.1, and 0.2. A plot illustrating the data-generating distribution with $f = f_1$ and $\beta_0 = 0$, with and without outliers, is given in the Supplementary Material.

The proposed method was used to estimate the combination and threshold using training data with 200, 400, or 800 observations. We evaluated the fitted combination in a large test set with 10^6 observations from the same population giving rise to the training data. We compared the fitted combination from the proposed method to those based on robust logistic regression and standard logistic regression. The robust logistic regression method used here is that of Bianco & Yohai (1996), an estimation method that is designed to have some robustness against so-called anomalous data, including, for example, the normal mixture model defined above. In short, each of the three methods is used to fit a linear combination of the biomarkers. Both standard and robust logistic regression use the logit link to model the data, while the proposed method does not depend on the specification of a link function. Standard and robust logistic regression differ in how they fit the logistic model; in particular, standard logistic regression maximizes a likelihood, while robust logistic regression minimizes a loss function designed to limit the influence of individual observations.

We evaluated the true positive rate in the test data for a false positive rate of t in the test data. In other words, for each combination, the threshold used to calculate the true positive rate in the test data was chosen such that the false positive rate in the test data was equal to t. We evaluated the false positive rate in the test data using the thresholds estimated in the training data. For standard and robust logistic regression, this threshold is the $(1 - t)$th quantile of the fitted biomarker combination among controls in the training data. For the proposed method, two thresholds are considered: the threshold estimated directly by the proposed method, as defined in Equation (1), and the $(1 - t)$th quantile of the fitted biomarker combination among controls in the training data. While the true and false positive rates in the test data are empirical estimates, the test set is so large that the estimates will be very close to the true and false positive rates. The simulations were repeated 1000 times.

Table 1 summarizes the results for the logit link, f_1, with moderate prevalence. The performance of the proposed method is generally similar to robust logistic regression and is similar to or better than standard logistic regression in terms of the true positive rate, though the false positive rate for the proposed method tends to be slightly higher than t when both t and the training dataset are small. There are some benefits in terms of the precision of the true positive rate for standard logistic regression and, when outliers are not present, robust logistic regression, relative to the proposed method. Improvements are generally seen for the proposed method when the threshold δ is reestimated in the training data based on the fitted combination, as opposed to estimated directly.

Table 2 presents the results for the piecewise logistic function, f_2, with moderate prevalence. When there are no outliers, the performance of the proposed method in terms of the true positive rate is generally comparable to standard and robust logistic regression, though there tends to be less variability in performance for standard and robust logistic regression. When there are outliers, the proposed method tends to perform better than both standard and robust logistic regression in terms of the true positive rate. Whether or not there are outliers, the false positive rate for the proposed method tends to be slightly higher than t when both t and the training dataset are small. In most cases, improvements are seen for the proposed method when the threshold δ is reestimated.

The results for low and high prevalence are presented in the Supplementary Material. The results are generally similar to those presented in Tables 1 and 2, though there are some differences. When the prevalence is low and outliers are present, the differences between the methods are smaller than in Tables 1 and 2. When the prevalence is low and there are no outliers, the differences in terms of the false positive rate are smaller than in Tables 1 and 2. When the prevalence is high, the differences in terms of the false positive rate are slightly larger than in Tables 1 and 2. Furthermore, when t is small, the prevalence is high, and the sample size is small, all of the methods have difficulty maintaining the acceptable false positive rate, as might be expected. For f_1, when the prevalence is high, the differences in the true positive rate are smaller than was seen in Table 1 when outliers are present and are slightly larger when outliers are not present.

For some data-generating distributions, the gains offered by the proposed method over robust logistic regression are quite substantial. For example, we considered a scenario with $f = f_2$, true combination $\beta_0 + 4X_1 - 3X_2 - 0.6(X_1 - X_2)^3$, a training set size of 800, $t = 0.2$, outliers in the data-generating distribution, and $\beta_0 = 1.5$, giving a prevalence of approximately 93%. The fitted combinations were evaluated as described above. Across 1000 simulations, the mean (standard deviation) true positive rate, as a percentage, was 55.3 (8.4) for standard logistic regression, 61.9 (14.2) for robust logistic regression, and 70.1 (15.4) for the proposed method. Likewise, the mean (standard deviation) false positive rate, as a percentage, was 21.3 (5.1) for standard logistic regression, 21.4 (5.1) for robust logistic regression, 23.0 (5.9) for the proposed method with the threshold estimated directly, and 22.5 (5.3) for the proposed method with the threshold

Table 1: Mean true and false positive rates and standard deviation (in parentheses) for $f(v) = f_1(v) \equiv \text{expit}(v) = e^v/(1+e^v)$ and β_0 = 0 across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n	True positive rate			False positive rate			
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	12.2 (2.1)	13.6 (2.6)	13.4 (2.7)	5.7 (2.2)	5.9 (2.3)	6.8 (2.5)	6.4 (2.4)
	400	12.1 (1.7)	14.1 (2.3)	13.9 (2.4)	5.4 (1.6)	5.4 (1.6)	6.0 (1.7)	5.9 (1.7)
	800	11.8 (1.2)	14.4 (2.2)	14.4 (2.3)	5.1 (1.1)	5.2 (1.1)	5.5 (1.2)	5.5 (1.2)
No	200	18.3 (0.6)	18.3 (0.6)	17.8 (1.8)	5.5 (2.2)	5.5 (2.2)	6.8 (2.5)	6.2 (2.4)
	400	18.5 (0.3)	18.5 (0.3)	18.1 (1.6)	5.3 (1.5)	5.3 (1.5)	5.9 (1.7)	5.7 (1.6)
	800	18.6 (0.2)	18.6 (0.2)	18.4 (1.2)	5.2 (1.1)	5.2 (1.1)	5.6 (1.2)	5.5 (1.2)
				$t = 0.10$				
Yes	200	22.5 (3.8)	24.6 (4.3)	24.6 (4.2)	10.9 (3.1)	11.1 (3.0)	12.0 (3.2)	11.7 (3.2)
	400	21.8 (2.8)	25.1 (4.0)	25.2 (4.0)	10.4 (2.0)	10.5 (2.1)	11.1 (2.1)	11.0 (2.1)
	800	21.4 (2.0)	25.7 (3.6)	25.8 (3.6)	10.1 (1.5)	10.1 (1.5)	10.5 (1.5)	10.5 (1.5)
No	200	29.4 (0.8)	29.5 (0.8)	28.9 (2.2)	10.5 (3.1)	10.5 (3.1)	11.8 (3.3)	11.4 (3.2)
	400	29.8 (0.4)	29.8 (0.4)	29.5 (1.3)	10.4 (2.1)	10.4 (2.1)	11.1 (2.3)	10.9 (2.2)
	800	29.9 (0.2)	29.9 (0.2)	29.7 (1.5)	10.2 (1.5)	10.2 (1.5)	10.6 (1.7)	10.6 (1.5)
				$t = 0.20$				
Yes	200	38.0 (5.1)	40.8 (5.8)	41.0 (5.7)	20.9 (4.0)	21.1 (4.0)	22.0 (4.0)	21.8 (4.0)
	400	37.4 (3.9)	41.7 (5.3)	41.9 (5.2)	20.5 (2.8)	20.6 (2.9)	21.2 (2.9)	21.1 (2.9)
	800	36.9 (2.9)	42.4 (4.6)	43.0 (4.4)	20.2 (2.0)	20.4 (2.0)	20.7 (1.9)	20.7 (2.0)
No	200	46.1 (0.9)	46.1 (0.9)	45.7 (1.5)	20.7 (4.1)	20.8 (4.1)	22.1 (4.2)	21.7 (4.2)
	400	46.4 (0.5)	46.4 (0.5)	46.2 (0.8)	20.3 (2.8)	20.3 (2.8)	21.1 (2.8)	21.0 (2.8)
	800	46.5 (0.2)	46.5 (0.3)	46.4 (0.6)	20.1 (2.0)	20.1 (2.0)	20.6 (2.0)	20.5 (2.0)

Table 2: Mean true and false positive rates and standard deviation (in parentheses) for $f(v) = f_2(v) \equiv 1(v < 0) \times (1 + e^{-v/3})^{-1} + 1(v \geq 0) \times (1 + e^{-3v})^{-1}$ and β_0 = 0 across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n	True positive rate			False positive rate			
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	20.2 (7.3)	26.4 (9.1)	27.7 (9.2)	5.9 (2.6)	6.0 (2.6)	6.9 (2.8)	6.5 (2.8)
	400	19.0 (5.9)	27.6 (8.5)	29.3 (8.2)	5.5 (1.8)	5.5 (1.7)	6.0 (1.8)	5.8 (1.8)
	800	17.9 (4.1)	29.4 (7.5)	30.8 (7.3)	5.3 (1.3)	5.3 (1.2)	5.6 (1.3)	5.5 (1.3)
No	200	37.9 (1.7)	37.8 (1.9)	37.5 (3.1)	5.8 (2.7)	5.7 (2.7)	7.3 (2.9)	6.5 (2.9)
	400	38.6 (0.9)	38.5 (1.0)	38.3 (2.1)	5.3 (1.8)	5.3 (1.8)	6.1 (1.8)	5.8 (1.8)
	800	38.9 (0.4)	38.9 (0.5)	38.6 (2.2)	5.2 (1.3)	5.2 (1.3)	5.6 (1.3)	5.5 (1.3)
				$t = 0.10$				
Yes	200	31.1 (8.9)	37.4 (10.8)	39.3 (11.0)	11.0 (3.5)	11.3 (3.6)	12.0 (3.7)	12.0 (3.6)
	400	30.3 (7.1)	39.9 (9.8)	41.5 (9.6)	10.5 (2.5)	10.7 (2.4)	11.0 (2.5)	11.0 (2.5)
	800	28.9 (5.0)	41.1 (8.9)	43.1 (8.6)	10.1 (1.7)	10.3 (1.7)	10.5 (1.8)	10.6 (1.8)
No	200	48.2 (1.8)	48.0 (1.9)	48.2 (2.0)	10.9 (3.5)	10.9 (3.5)	12.3 (3.5)	11.7 (3.6)
	400	48.8 (0.9)	48.7 (1.0)	48.7 (1.1)	10.4 (2.4)	10.4 (2.4)	11.2 (2.4)	10.9 (2.5)
	800	49.2 (0.4)	49.1 (0.5)	49.0 (0.6)	10.2 (1.7)	10.2 (1.7)	10.7 (1.7)	10.7 (1.8)
				$t = 0.20$				
Yes	200	45.0 (8.1)	50.4 (9.8)	51.9 (9.7)	21.2 (4.6)	21.5 (4.7)	22.1 (4.8)	22.0 (4.8)
	400	44.4 (6.3)	52.8 (8.6)	54.0 (8.5)	20.4 (3.2)	20.8 (3.3)	21.2 (3.3)	21.2 (3.4)
	800	44.1 (4.8)	54.8 (7.3)	56.5 (6.6)	20.2 (2.3)	20.3 (2.3)	20.6 (2.2)	20.7 (2.3)
No	200	59.5 (1.3)	59.4 (1.4)	59.3 (1.8)	21.1 (4.6)	21.1 (4.6)	22.6 (4.6)	22.1 (4.7)
	400	60.0 (0.6)	59.9 (0.7)	59.8 (0.9)	20.5 (3.4)	20.6 (3.4)	21.3 (3.3)	21.2 (3.4)
	800	60.2 (0.4)	60.1 (0.4)	60.1 (0.5)	20.3 (2.2)	20.3 (2.2)	20.7 (2.3)	20.7 (2.3)

reestimated based on quantiles of the fitted combination.

In addition to the data-generating distribution described above, we considered conditionally bivariate normal biomarkers with non-proportional covariance matrices. We simulated $D \sim$ Bernoulli(0.7) and

$$\left(\begin{array}{c} X_1 \\ X_2 \end{array} \middle| D = 1 \right) \sim N \left\{ \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right), 0.25 \times \left(\begin{array}{cc} 1 & 0.9 \\ 0.9 & 1 \end{array} \right) \right\},$$
$$\left(\begin{array}{c} X_1 \\ X_2 \end{array} \middle| D = 0 \right) \sim N \left\{ \left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1.25 & 0 \\ 0 & 1.25 \end{array} \right) \right\},$$

with $\mu_1 = 2^{1/2}\Phi^{-1}(\text{AUC}_{X_1})$ and $\mu_2 = 2^{1/2}\Phi^{-1}(\text{AUC}_{X_2})$, where $\text{AUC}_{X_1} = 0.6$ is the marginal AUC for X_1 and $\text{AUC}_{X_2} = 0.8$ is the marginal AUC for X_2. This data-generating distribution corresponds to a situation in which the biomarkers are highly correlated in cases, but essentially constitute noise in controls. Under this data-generating distribution, the optimal combination in terms of the ROC curve is of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2$. We considered a maximum acceptable false positive rate, t, of 0.10 and a training set size of 800. The fitted combinations were evaluated as described above. Across 1000 simulations, the mean (standard deviation) true positive rate, as a percentage, was 32.1 (3.6) for standard logistic regression, 24.6 (6.6) for robust logistic regression, and 32.0 (8.4) for the proposed method. Likewise, the mean (standard deviation) false positive rate, as a percentage, was 10.4 (2.0) for standard logistic regression, 10.4 (2.0) for robust logistic regression, 10.8 (2.9) for the proposed method with the threshold estimated directly, and 11.0 (2.0) for the proposed method with the threshold reestimated based on quantiles of the fitted combination. Thus, in this scenario, the proposed method was comparable to standard logistic regression in terms of the true positive rate but offered substantial improvements over robust logistic regression while maintaining control of the false positive rate near t.

In most simulation settings, convergence of the proposed method was achieved in more than 96% of simulations. For f_1 with $\beta_0 = 1.75$, convergence failed in up to 7.3% of simulations. Thus, caution may be warranted in more extreme scenarios, such as when the prevalence is very high, particularly if the sample size and/or t are small. In addition, when simulating with outliers, the true biomarker combination was occasionally so large that it returned a non-value for the outcome D; for example, with $f_1(v) = \text{expit}(v)$, this occurs in R when $v > 800$. These observations had to be removed from the simulated dataset, though this affected an extremely small fraction of observations.

5 Application to diabetes data

We apply the method we have developed to the study of diabetes in women with Pima Indian heritage. We consider seven predictors measured in this study: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, and age. The diabetes pedigree function is a measure of family history of diabetes (Smith et al., 1988). We used 332 observations as training data and reserved the remaining 200 observations for testing. The training and test datasets had 109 and 68 diabetes cases, respectively. We scaled the variables to have equal variance. The distribution of predictors is depicted in the Supplementary Material.

The combinations were fitted using the training data and evaluated using the test data. We fixed the acceptable false positive rate at $t = 0.10$. We used standard logistic regression, robust logistic regression, and the proposed method to construct the combinations, giving the results in Table 3, where the fitted combinations from standard and robust logistic regression have been normalized to aid in comparison.

Using thresholds based on $\text{FPR} = 0.10$ in the test data, the estimated true positive rate in the test data was 0.544 for both standard and robust logistic regression, and 0.559 for the proposed method. When the thresholds estimated in the training data were used, the estimated false positive rate in the test data was 0.182 for both standard and robust logistic regression, 0.258 for the proposed method using the threshold estimated directly, and 0.265 for the proposed method using the threshold reestimated in the training data based on the fitted combination. The estimated false positive rate in the test data exceeded the target value for all the

Table 3: Fitted combinations of the scaled predictors in the diabetes study. GLM denotes standard logistic regression, rGLM denotes robust logistic regression, and sTPR denotes the proposed method with $t = 0.10$.

Predictor	GLM	rGLM	sTPR
Number of pregnancies	0.321	0.320	0.403
Plasma glucose	0.793	0.792	0.627
Blood pressure	-0.077	-0.073	-0.026
Skin fold thickness	0.089	0.090	-0.146
Body mass index	0.399	0.400	0.609
Diabetes pedigree	0.280	0.281	0.191
Age	0.133	0.134	0.123

methods considered, indicating potentially important differences in the controls between the training and test data.

6 Discussion

The proposed method could be adapted to minimize the false positive rate while controlling the true positive rate to be above some acceptable level. Since the true positive rate and false positive rate are invariant to disease prevalence, the proposed method can be used with case-control data. In the presence of matching, however, it becomes necessary to consider the covariate-adjusted ROC curve and corresponding covariate-adjusted summaries, and thus the methods presented here are not immediately applicable (Janes & Pepe, 2008).

As our smooth approximation function is non-convex, the choice of starting values should be considered further. Extensions of convex methods, such as the ramp function method proposed by Fong *et al.* (2016) for the AUC, could also be considered. Further research into methods for evaluating the true and false positive rates of biomarker combinations after estimation, for example, sample-splitting, bootstrapping, or k-fold cross-validation, is needed.

Acknowledgments

The authors wish to gratefully acknowledge support from the National Institutes of Health. M. Carone also wishes to recognize generous support from the University of Washington Department of Biostatistics Career Development Fund.

Supplementary material

Supplementary material available includes the proof of the optimal combination of conditional multivariate normal biomarkers with proportional covariance matrices, an illustration of the data-generating distribution used to simulate data with and without outliers in Section 4, additional simulation results, the densities of the predictors in the diabetes study, and proofs of the lemmas used in the proof of Theorem 1.

Appendix

The proof of Theorem 1 relies on Lemmas A1 and A2, which are stated below and proved in the Supplementary Material.

Lemma A1. *Say that a bounded function $f : \mathbb{R}^d \to \mathbb{R}$ and possibly random sets $\Omega_0, \Omega_1, \Omega_2, \ldots \subseteq \mathbb{R}^d$ are given, and let $\{a_n\}_{n\geq 1}$ be a decreasing sequence of positive real numbers tending to zero. For each $n \geq 1$, suppose that $\omega_{0,n} \in \Omega_0$ and $\omega_n \in \Omega_n$ are near-maximizers of f over Ω_0 and Ω_n, respectively, in the sense that $f(\omega_{0,n}) \geq \sup_{\omega\in\Omega_0} f(\omega) - a_n$ and $f(\omega_n) \geq \sup_{\omega\in\Omega_n} f(\omega) - a_n$. Further, define*

$$d_n = \sup_{\omega\in\Omega_n} \inf_{\tilde{\omega}\in\Omega_0} d(\omega, \tilde{\omega}), \ \ e_n = \sup_{\omega\in\Omega_0} \inf_{\tilde{\omega}\in\Omega_n} d(\omega, \tilde{\omega}),$$

where d is the Euclidean distance in $\mathbb{R}^d$. If d_n and e_n tend to zero almost surely, and f is globally Lipschitz continuous, then $|f(\omega_{0,n}) - f(\omega_n)|$ tends to zero almost surely. In particular, this implies that

$$\left| \sup_{\omega\in\Omega_0} f(\omega) - \sup_{\omega\in\Omega_n} f(\omega) \right| \longrightarrow 0$$

almost surely.

Lemma A2. *Under conditions (1)–(5), we have that*

$$\sup_{(\theta,\delta)\in\Omega} |\widetilde{\text{FPR}}_{n_0}(\theta, \delta) - \text{FPR}(\theta, \delta)| \longrightarrow 0, \ \sup_{(\theta,\delta)\in\Omega} |\widetilde{\text{TPR}}_{n_1}(\theta, \delta) - \text{TPR}(\theta, \delta)| \longrightarrow 0$$

almost surely as n tends to $+\infty$, where $\Omega = \{(\theta, \delta) \in \mathbb{R}^p \times \mathbb{R} : ||\theta|| = 1\}$.

Proof of Theorem 1. First, we show that $\limsup_n \text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) \leq t$ almost surely. We can write

$$\begin{aligned} \text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) \ &= \ \widetilde{\text{FPR}}_{n_0}(\tilde{\theta}_t, \tilde{\delta}_t) + \{\text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) - \widetilde{\text{FPR}}_{n_0}(\tilde{\theta}_t, \tilde{\delta}_t)\} \\ &\leq \ \widetilde{\text{FPR}}_{n_0}(\tilde{\theta}_t, \tilde{\delta}_t) + |\text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) - \widetilde{\text{FPR}}_{n_0}(\tilde{\theta}_t, \tilde{\delta}_t)| \\ &\leq \ \widetilde{\text{FPR}}_{n_0}(\tilde{\theta}_t, \tilde{\delta}_t) + \sup_{(\theta,\delta)\in\Omega} |\text{FPR}(\theta, \delta) - \widetilde{\text{FPR}}_{n_0}(\theta, \delta)| \ \leq \ t + \sup_{(\theta,\delta)\in\Omega} |\text{FPR}(\theta, \delta) - \widetilde{\text{FPR}}_{n_0}(\theta, \delta)| \,. \end{aligned}$$

As such, it follows that

$$\text{pr}\{\limsup_n \text{FPR}(\tilde{\theta}_t, \tilde{\delta}_t) \leq t\} \ \geq \ \text{pr}\{\limsup_n \sup_{(\theta,\delta)\in\Omega} |\text{FPR}(\theta, \delta) - \widetilde{\text{FPR}}_{n_0}(\theta, \delta)| = 0\} \ = \ 1$$

in view of Lemma A2, thereby establishing the first part of the theorem.

Let $t \in (0, 1)$ be fixed. We now establish that

$$\left| \text{TPR}(\tilde{\theta}_t, \tilde{\delta}_t) - \sup_{(\theta,\delta)\in\Omega_t} \text{TPR}(\theta, \delta) \right| \longrightarrow 0$$

almost surely. For convenience, denote (θ, δ) by ω. Consider the function f defined pointwise as $f(\omega) = \text{TPR}(\theta, \delta)$, and set $\Omega_0 = \Omega_t$ and $\Omega_n = \tilde{\Omega}_{t,n_0}$ for each $n \geq 1$. We verify that the conditions of Lemma A1 hold for these particular choices. We have that $f(\omega) = \text{TPR}(\theta, \delta)$ is a bounded function. We must show d_{n_0} and e_{n_0} tend to zero almost surely, where

$$d_{n_0} = \sup_{\omega\in\tilde{\Omega}_{t,n_0}} \inf_{\tilde{\omega}\in\Omega_t} d(\omega, \tilde{\omega}), \quad e_{n_0} = \sup_{\omega\in\Omega_t} \inf_{\tilde{\omega}\in\tilde{\Omega}_{t,n_0}} d(\omega, \tilde{\omega}),$$

and d is the Euclidean distance in $\mathbb{R}^{p+1}$. We consider d_{n_0} first. Denote by G_θ the conditional distribution function of $\theta^\top X$ given $D = 0$. By assumption, the corresponding conditional quantile function, denoted by G_θ^{-1}, is uniformly Lipschitz continuous over $\{\theta \in \mathbb{R}^p : \|\theta\| = 1\}$, say with constant $C > 0$ independent of θ. Suppose that, for some $\kappa > 0$, $\sup_{\omega\in\tilde{\Omega}_{t,n_0}} |\widetilde{\text{FPR}}_{n_0}(\omega) - \text{FPR}(\omega)| \leq \kappa$. Because it is true that

$$\kappa \geq \sup_{\omega\in\tilde{\Omega}_{t,n_0}} |\widetilde{\text{FPR}}_{n_0}(\omega) - \text{FPR}(\omega)| \ \geq \ \left| \sup_{\omega\in\tilde{\Omega}_{t,n_0}} \widetilde{\text{FPR}}_{n_0}(\omega) - \sup_{\omega\in\tilde{\Omega}_{t,n_0}} \text{FPR}(\omega) \right|,$$

then $\sup_{\omega\in\tilde{\Omega}_{t,n_0}} \mathrm{FPR}(\omega) \le \kappa + t$, giving $\widetilde{\mathrm{FPR}}_{n_0}(\omega) \le t$ and $\mathrm{FPR}(\omega) \le \kappa + t$ for each $\omega \in \tilde{\Omega}_{t,n_0}$.

For any given $\omega = (\theta,\delta) \in \tilde{\Omega}_{t,n_0}$, write $t_*(\omega) = G_\theta(\delta)$, giving $t_*(\omega) = \mathrm{FPR}(\omega) \le \kappa + t$. If $t_*(\omega) \le t$, note also that $\omega \in \Omega_t$ and set $\omega_* = \omega$. Otherwise, find δ_* such that $1 - G_\theta(\delta_*) = t$, namely by taking $\delta_* = G_\theta^{-1}(1-t)$. Defining $\omega_* = (\theta,\delta_*) \in \Omega_t$, observe that

$$d(\omega,\omega_*) \;=\; |\delta - \delta_*| \;=\; |G_\theta^{-1}(1-t_*(\omega)) - G_\theta^{-1}(1-t)| \;\le\; C|t - t_*(\omega)| \;\le\; C\kappa\,.$$

Thus, for each $\omega \in \tilde{\Omega}_{t,n_0}$, it is true that $\inf_{\tilde{\omega}\in\Omega_t} d(\omega,\tilde{\omega}) \le C\kappa$ and therefore $d_{n_0} \le C\kappa$. As such, if $d_{n_0} > \epsilon$ for some $\epsilon > 0$, then $\sup_{\omega\in\tilde{\Omega}_{t,n_0}} |\widetilde{\mathrm{FPR}}_{n_0}(\omega) - \mathrm{FPR}(\omega)| > \kappa_\epsilon$ for $\kappa_\epsilon = \epsilon/C$. This implies that

$$\mathrm{pr}\left(\sup_{m\ge n_0} d_m > \epsilon\right) \;\le\; \mathrm{pr}\left(\sup_{m\ge n_0}\sup_{\omega\in\tilde{\Omega}_{t,m}} |\widetilde{\mathrm{FPR}}_m(\omega) - \mathrm{FPR}(\omega)| > \kappa_\epsilon\right) \;\longrightarrow\; 0$$

by Lemma A2. Thus, d_n tends to zero almost surely since, for each $\epsilon > 0$,

$$\mathrm{pr}\left(\limsup_m \{d_m \ge \epsilon\}\right) \;\le\; \mathrm{pr}\left(\limsup_m d_m \ge \epsilon\right) \;=\; 0\,.$$

Using similar arguments, we may show that e_n also tends to zero almost surely.

The fact that d_n and e_n tend to zero almost surely implies, in view of Lemma A1, that we have that $|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \mathrm{TPR}(\theta,\delta)|$ tends to zero almost surely. Combining this with an application of Lemma A2, we have that

$$\begin{aligned}
&\left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)\right| \\
&\quad\le \left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \mathrm{TPR}(\theta,\delta)\right| + \left|\sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)\right| \\
&\quad\le \left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \mathrm{TPR}(\theta,\delta)\right| + \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} |\mathrm{TPR}(\theta,\delta) - \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)| \;\longrightarrow\; 0
\end{aligned}$$

almost surely. Since $|\mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t) - \widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t)| \le \sup_{(\theta,\delta)\in\Omega} |\mathrm{TPR}(\theta,\delta) - \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)|$ and, by Lemma A2, $\sup_{(\theta,\delta)\in\Omega} |\mathrm{TPR}(\theta,\delta) - \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)|$ tends to zero almost surely, $|\mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t) - \widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t)|$ tends to zero almost surely. In addition, since $(\tilde{\theta}_t,\tilde{\delta}_t)$ is a near-maximizer of $\widetilde{\mathrm{TPR}}_{n_1}$, $\sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta) \le \widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t) + a_n$, giving

$$\begin{aligned}
&\left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t)\right| \\
&\quad\le \left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)\right| + \left|\sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta) - \mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t)\right| \\
&\quad\le \left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)\right| + \left|\sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta) - \widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t)\right| \\
&\qquad + \left|\widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t) - \mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t)\right| \\
&\quad\le \left|\sup_{(\theta,\delta)\in\Omega_t} \mathrm{TPR}(\theta,\delta) - \sup_{(\theta,\delta)\in\tilde{\Omega}_{t,n_0}} \widetilde{\mathrm{TPR}}_{n_1}(\theta,\delta)\right| + a_n + \left|\widetilde{\mathrm{TPR}}_{n_1}(\tilde{\theta}_t,\tilde{\delta}_t) - \mathrm{TPR}(\tilde{\theta}_t,\tilde{\delta}_t)\right| \;\longrightarrow\; 0
\end{aligned}$$

almost surely, completing the proof. □

References

Anderson, T. W. & Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Stat.* **33,** 420–431.

Baker, S. G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56,** 1082–1087.

Bianco, A. M. & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods*, H. Rieder, ed., pp 17–34. Springer New York.

Fong, Y., Yin, S. & Huang, Y. (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. *Stat. Med.* **35,** 3792–3809.

Gao, F., Xiong, C., Yan, Y., Yu, K. & Zhang, Z. (2008). Estimating optimum linear combination of multiple correlated diagnostic tests at a fixed specificity with receiver operating characteristic curves. *J. Data. Sci.* **6,** 105–123.

Hsu, M.-J. & Hsueh, H.-M. (2013). The linear combinations of biomarkers which maximize the partial area under the ROC curves. *Comput. Stat.* **28,** 647–666.

Hwang, K.-B., Ha, B.-Y., Ju, S. & Kim, S. (2013). Partial AUC maximization for essential gene prediction using genetic algorithms. *BMB Rep.* **46,** 41–46.

Janes, H. & Pepe, M. S. (2008). Adjusting for covariates in studies of diagnostic, screening or prognostic markers: an old concept in a new setting. *Am. J. Epidemiol.* **168,** 89–97.

Komori, O. & Eguchi, S. (2010). A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics*.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*, pp 155–178. Springer-Verlag New York.

Lin, H., Zhou, L., Peng, H. & Zhou, X.-H. (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Can. J. Stat.* **39,** 324–343.

Liu, A., Schisterman, E. F. & Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Stat. Med.* **24,** 37–47.

Ma, S. & Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63,** 751–757.

McIntosh, M. W. & Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58,** 657–664.

Moore, R. G., Brown, A. K., Miller, M. C., Skates, S., Allard, W. J., Verch, T., Steinhoff, M., Messerlian, G., DiSilvestro, P., Granai, C. O. & Bast, R. C. (2008). The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. *Gynecol. Oncol.* **108,** 402–408.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, pp 66–95. Oxford University Press.

Pepe, M. S., Cai, T. & Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62,** 221–229.

Pepe, M. & Janes, H. (2013). Methods of evaluating prediction performance of biomarkers and tests. In *Risk Assessment and Evaluation of Predictions*, M.-L.T. Lee, M. Gail, R. Pfeiffer, G. Satten, T. Cai & A. Gandy, eds., pp 107–142. Springer-Verlag New York.

Ricamato, M. T. & Tortorella, F. (2011). Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recogn.* **44,** 2669–2677.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc. Annu. Symp. Comput. Appl. Med. Care*, 261–265.

Su, J. Q. & Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *J. Am. Statist. Assoc.* **88,** 1350–1355.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*, pp 69–91. Springer-Verlag New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, pp 265–290. Cambridge University Press.

Wang, Z. & Chang, Y.-C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12,** 369–385.

Winter, B. B. (1979). Convergence rate of perturbed empirical distribution functions. *J. Appl. Probab.* **16,** 163–173.

Yu, W. & Park, T. (2015). Two simple algorithms on linear combination of multiple biomarkers to maximize partial area under the ROC curve. *Comput. Stat. Data Anal.* **88,** 15–27.

Supplementary Material for 'Combining Biomarkers by Maximizing the True Positive Rate for a Fixed False Positive Rate'

A. Meisner[*,1], M. Carone[†,1], M. S. Pepe[‡,2], and K. F. Kerr[§,1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.
[2]Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.

S1 Optimal combination under proportional covariance matrices

Proposition 1. *If the biomarkers* (X_1, X_2) *are conditionally multivariate normal with proportional covariance matrices given* D*, that is,*

$$(X_1, X_2 \mid D = 0) \sim N(\mu_0, \Sigma), \;\; (X_1, X_2 \mid D = 1) \sim N(\mu_1, \sigma^2\Sigma),$$

then the optimal biomarker combination in the sense of the ROC curve is of the form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2$$

for some vector $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \in \mathbb{R}^5$.

Proof of Proposition 1. It is known that the optimal combination of (X_1, X_2) in terms of the ROC curve is the likelihood ratio, $f(X_1, X_2 \mid D = 1)/f(X_1, X_2 \mid D = 0)$, or any monotone increasing function thereof (McIntosh & Pepe, 2002). Let $M = (X_1, X_2)$. Without loss of generality, let $\mu_0 = 0$ and $\mu_1 = \mu = (\mu_{X_1}, \mu_{X_2})$. Then

$$\begin{aligned}\frac{f(M \mid D=1)}{f(M \mid D=0)} &= \frac{|\sigma^2\Sigma|^{-1/2}\text{exp}\left\{-\frac{1}{2}(M-\mu)^\top(\sigma^2\Sigma)^{-1}(M-\mu)\right\}}{|\Sigma|^{-1/2}\text{exp}\left\{-\frac{1}{2}M^\top\Sigma^{-1}M\right\}} \\ &= \frac{\text{exp}\left\{-\frac{1}{2}(M-\mu)^\top(\sigma^2\Sigma)^{-1}(M-\mu)\right\}}{\sigma^2\text{exp}\left\{-\frac{1}{2}M^\top\Sigma^{-1}M\right\}} \\ &= \frac{1}{\sigma^2}\text{exp}\left\{-\frac{(M-\mu)^\top\Sigma^{-1}(M-\mu)}{2\sigma^2} + \frac{M^\top\Sigma^{-1}M}{2}\right\}.\end{aligned}$$

Denote the entries of Σ^{-1} by

$$\Sigma^{-1} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

[*] meisnera@uw.edu
[†] mcarone@uw.edu
[‡] mspepe@uw.edu
[§] katiek@uw.edu

Then, we can write that

$$
\begin{aligned}
&-\frac{1}{2\sigma^2}(M-\mu)^\top\Sigma^{-1}(M-\mu)+\frac{1}{2}M^\top\Sigma^{-1}M \\
&\quad= \frac{1}{2}\Bigg[\frac{1}{\sigma^2}\big\{-S_{11}(X_1^2-2X_1\mu_{X_1}+\mu_{X_1}^2)-S_{21}(X_1X_2-X_2\mu_{X_1}-X_1\mu_{X_2}+\mu_{X_1}\mu_{X_2}) \\
&\qquad -S_{12}(X_1X_2-X_1\mu_{X_2}-X_2\mu_{X_1}+\mu_{X_1}\mu_{X_2})-S_{22}(X_2^2-2X_2\mu_{X_2}+\mu_{X_2}^2)\big\} \\
&\qquad + S_{11}X_1^2+S_{21}X_1X_2+S_{12}X_1X_2+S_{22}X_2^2\Bigg] \\
&\quad= \frac{1}{2}\left\{\left(S_{11}-\frac{S_{11}}{\sigma^2}\right)X_1^2+\left(S_{22}-\frac{S_{22}}{\sigma^2}\right)X_2^2+\left(S_{12}+S_{21}-\frac{S_{12}}{\sigma^2}-\frac{S_{21}}{\sigma^2}\right)X_1X_2\right. \\
&\qquad +\left(\frac{2S_{11}\mu_{X_1}+S_{21}\mu_{X_2}+S_{12}\mu_{X_2}}{\sigma^2}\right)X_1+\left(\frac{S_{21}\mu_{X_1}+S_{12}\mu_{X_1}+2S_{22}\mu_{X_2}}{\sigma^2}\right)X_2 \\
&\qquad \left.+\frac{-S_{11}\mu_{X_1}^2-S_{21}\mu_{X_1}\mu_{X_2}-S_{12}\mu_{X_1}\mu_{X_2}-S_{22}\mu_{X_2}^2}{\sigma^2}\right\} \\
&\quad= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2+\beta_4X_1^2+\beta_5X_2^2
\end{aligned}
$$

as claimed, where

$$
\begin{aligned}
\beta_0 &= \frac{-S_{11}\mu_{X_1}^2-S_{21}\mu_{X_1}\mu_{X_2}-S_{12}\mu_{X_1}\mu_{X_2}-S_{22}\mu_{X_2}^2}{\sigma^2} \\
\beta_1 &= \left(\frac{2S_{11}\mu_{X_1}+S_{21}\mu_{X_2}+S_{12}\mu_{X_2}}{\sigma^2}\right) \\
\beta_2 &= \left(\frac{S_{21}\mu_{X_1}+S_{12}\mu_{X_1}+2S_{22}\mu_{X_2}}{\sigma^2}\right) \\
\beta_3 &= \left(S_{12}+S_{21}-\frac{S_{12}}{\sigma^2}-\frac{S_{21}}{\sigma^2}\right) \\
\beta_4 &= \left(S_{11}-\frac{S_{11}}{\sigma^2}\right) \\
\beta_5 &= \left(S_{22}-\frac{S_{22}}{\sigma^2}\right).
\end{aligned}
$$

□

S2 Illustration of data-generating distribution

This plot illustrates the data-generating distribution used to simulate data with and without outliers in Section 4. Specifically, we have that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (1-\Delta) \times Z_0 + \Delta \times Z_1$$

and D is then simulated as a Bernoulli random variable with success probability $f\left\{\beta_0 + 4X_1 - 3X_2 - 0.8(X_1 - X_2)^3\right\}$, where Δ is a Bernoulli random variable with success probability $\pi = 0.05$ when outliers are simulated and $\pi = 0$ otherwise, Z_0 and Z_1 are independent bivariate normal random variables with mean zero and respective covariance matrices

$$0.2 \times \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, \; 2 \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

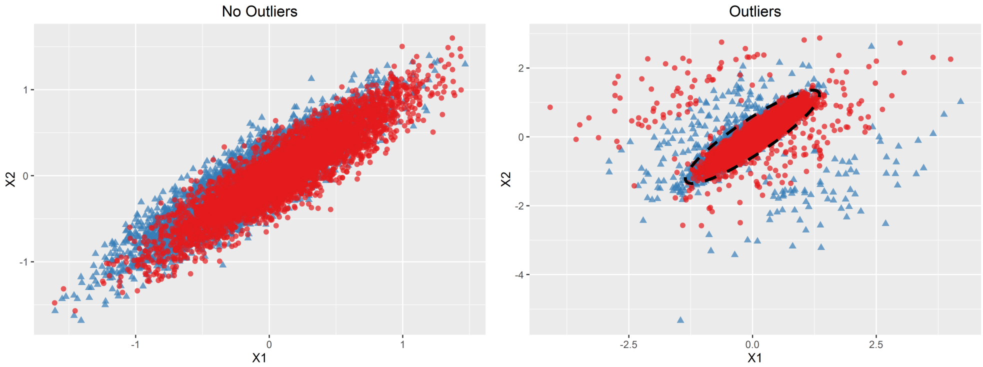

Figure S1: Datasets with $f(v) = f_1(v) \equiv \text{expit}(v)$, $\beta_0 = 0$, without (left plot) and with outliers (right plot). Cases are represented by red circles, and controls are represented by blue triangles. The plot with outliers also includes an ellipse (dashed black line) indicating the 99% confidence region for the distribution of (X_1, X_2) without outliers.

S3 Additional simulation results

The tables below present additional results related to the simulations described in §4; that is, data with and without outliers in the setting of high and low disease prevalences. In each table, we report (i) the mean and standard deviation of the true positive rate in the test data using the threshold corresponding to a false positive rate of t in the test data and (ii) the mean and standard deviation of the false positive rate in the test data corresponding to the thresholds estimated in the training data.

Table S1: Mean true positive rate and false positive rate and corresponding standard deviation (in parentheses) for $f(v) = f_1(v) \equiv \text{expit}(v) = e^v/(1 + e^v)$ and $\beta_0 = -1.75$ across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n	True positive rate			False positive rate			
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	13.0 (2.8)	13.4 (3.4)	13.5 (3.4)	5.3 (1.7)	5.4 (1.7)	5.8 (1.9)	5.7 (1.8)
	400	12.7 (1.9)	13.4 (2.7)	13.6 (2.9)	5.2 (1.2)	5.2 (1.2)	5.4 (1.3)	5.4 (1.2)
	800	12.5 (1.3)	13.2 (2.1)	13.6 (2.5)	5.1 (0.8)	5.2 (0.8)	5.3 (0.9)	5.2 (0.9)
No	200	18.1 (1.0)	18.1 (1.1)	17.5 (2.2)	5.5 (1.8)	5.5 (1.8)	6.1 (1.9)	5.9 (1.8)
	400	18.5 (0.6)	18.5 (0.6)	18.2 (1.6)	5.1 (1.2)	5.2 (1.2)	5.5 (1.3)	5.4 (1.3)
	800	18.7 (0.3)	18.7 (0.3)	18.5 (1.1)	5.1 (0.9)	5.1 (0.9)	5.3 (0.9)	5.3 (0.9)
				$t = 0.10$				
Yes	200	22.1 (4.5)	22.7 (5.3)	23.1 (5.3)	10.4 (2.4)	10.5 (2.4)	10.8 (2.5)	10.8 (2.4)
	400	21.9 (3.6)	22.8 (4.7)	23.4 (4.8)	10.1 (1.7)	10.2 (1.7)	10.4 (1.8)	10.4 (1.8)
	800	21.4 (2.3)	22.3 (3.4)	23.3 (4.3)	10.1 (1.2)	10.1 (1.2)	10.2 (1.4)	10.3 (1.2)
No	200	29.5 (1.3)	29.4 (1.3)	28.8 (2.5)	10.3 (2.3)	10.4 (2.3)	11.1 (2.3)	10.9 (2.3)
	400	29.8 (0.7)	29.8 (0.7)	29.5 (1.5)	10.2 (1.7)	10.2 (1.7)	10.7 (1.7)	10.6 (1.7)
	800	30.1 (0.4)	30.1 (0.4)	29.8 (1.1)	10.1 (1.1)	10.1 (1.1)	10.4 (1.1)	10.3 (1.1)
				$t = 0.20$				
Yes	200	36.4 (6.6)	37.2 (7.8)	38.1 (7.4)	20.5 (3.2)	20.6 (3.1)	20.9 (3.5)	21.0 (3.2)
	400	36.2 (4.7)	37.3 (6.2)	38.5 (6.4)	20.1 (2.2)	20.2 (2.3)	20.4 (2.2)	20.4 (2.2)
	800	35.7 (3.0)	37.0 (4.6)	38.8 (5.7)	20.2 (1.5)	20.2 (1.5)	20.3 (1.7)	20.4 (1.5)
No	200	46.1 (1.7)	46.1 (1.7)	45.5 (2.6)	20.4 (3.1)	20.5 (3.2)	21.1 (3.1)	21.0 (3.2)
	400	46.7 (0.8)	46.7 (0.8)	46.4 (1.3)	20.1 (2.1)	20.2 (2.1)	20.5 (2.1)	20.5 (2.1)
	800	47.0 (0.4)	47.0 (0.4)	46.8 (0.7)	20.0 (1.6)	20.0 (1.6)	20.3 (1.5)	20.2 (1.6)

Table S2: Mean true positive rate and false positive rate and corresponding standard deviation (in parentheses) for $f(v) = f_1(v) \equiv \text{expit}(v) = e^v/(1+e^v)$ and β_0 = 1.75 across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n		True positive rate			False positive rate		
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	8.4 (1.2)	8.4 (1.4)	8.2 (1.8)	7.3 (4.0)	6.9 (3.9)	9.2 (5.6)	7.7 (4.6)
	400	8.6 (0.9)	8.5 (1.1)	8.3 (1.6)	6.3 (2.7)	6.3 (2.7)	7.2 (3.6)	6.7 (2.9)
	800	8.7 (0.6)	8.6 (0.7)	8.5 (1.5)	5.8 (1.8)	5.8 (1.8)	6.2 (2.5)	6.1 (2.0)
No	200	18.7 (1.0)	18.7 (1.0)	17.2 (3.5)	6.3 (4.1)	6.1 (4.0)	9.3 (5.9)	7.4 (4.5)
	400	19.0 (0.5)	19.0 (0.6)	17.9 (2.9)	5.7 (2.7)	5.6 (2.7)	7.1 (3.7)	6.4 (3.0)
	800	19.2 (0.3)	19.2 (0.3)	18.3 (2.9)	5.3 (1.9)	5.3 (1.9)	6.1 (2.5)	5.9 (2.0)
				$t = 0.10$				
Yes	200	18.6 (3.9)	19.1 (4.7)	19.4 (4.8)	12.4 (5.1)	12.4 (5.0)	15.0 (6.2)	13.5 (5.4)
	400	18.6 (2.5)	19.2 (3.5)	19.8 (3.8)	11.1 (3.4)	11.1 (3.5)	12.6 (4.2)	12.0 (3.7)
	800	18.4 (1.4)	19.2 (2.6)	19.8 (3.4)	10.8 (2.6)	10.8 (2.6)	11.4 (3.1)	11.3 (2.7)
No	200	29.9 (1.3)	29.9 (1.3)	28.7 (3.6)	11.7 (5.2)	11.5 (5.2)	14.7 (6.7)	13.1 (5.6)
	400	30.4 (0.6)	30.3 (0.7)	29.4 (3.4)	10.7 (3.6)	10.6 (3.6)	12.4 (4.5)	11.7 (3.8)
	800	30.6 (0.3)	30.6 (0.3)	30.2 (2.0)	10.4 (2.5)	10.4 (2.5)	11.4 (2.8)	11.1 (2.5)
				$t = 0.20$				
Yes	200	34.2 (6.4)	34.9 (7.7)	35.9 (7.1)	22.5 (6.5)	22.7 (6.3)	25.0 (7.4)	24.0 (6.7)
	400	34.2 (4.3)	35.0 (5.6)	36.3 (5.9)	21.4 (4.7)	21.5 (4.7)	22.9 (5.2)	22.4 (4.8)
	800	33.9 (2.8)	35.0 (4.4)	36.2 (5.0)	20.6 (3.3)	20.7 (3.3)	21.5 (3.5)	21.3 (3.4)
No	200	46.4 (1.6)	46.4 (1.6)	45.6 (3.3)	22.2 (7.0)	22.0 (7.0)	25.6 (7.8)	23.9 (7.1)
	400	47.0 (0.8)	47.0 (0.8)	46.5 (2.2)	20.8 (5.0)	20.7 (4.9)	22.8 (5.1)	22.0 (5.0)
	800	47.2 (0.4)	47.2 (0.4)	46.9 (1.9)	20.6 (3.4)	20.6 (3.4)	21.6 (3.8)	21.4 (3.5)

Table S3: Mean true positive rate and false positive rate and corresponding standard deviation (in parentheses) for $f(v) = f_2(v) \equiv 1(v < 0) \times (1/(1 + e^{-v/3})) + 1(v \geq 0) \times (1/(1 + e^{-3v}))$ and $\beta_0 = -5.25$ across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n	True positive rate			False positive rate			
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	7.1 (1.1)	7.1 (1.1)	7.1 (1.1)	5.7 (1.8)	5.7 (1.8)	6.0 (2.0)	5.9 (1.9)
	400	7.4 (1.0)	7.3 (0.9)	7.3 (1.0)	5.3 (1.2)	5.4 (1.2)	5.5 (1.3)	5.5 (1.2)
	800	7.6 (0.8)	7.5 (0.8)	7.5 (0.9)	5.1 (0.8)	5.2 (0.8)	5.2 (1.0)	5.2 (0.9)
No	200	7.3 (1.4)	7.3 (1.4)	7.2 (1.4)	5.5 (1.7)	5.6 (1.7)	6.0 (1.9)	5.9 (1.8)
	400	7.8 (0.9)	7.8 (1.0)	7.7 (1.1)	5.2 (1.2)	5.2 (1.1)	5.5 (1.4)	5.4 (1.2)
	800	8.1 (0.4)	8.1 (0.4)	8.0 (0.7)	5.1 (0.9)	5.1 (0.9)	5.2 (1.1)	5.2 (0.9)
				$t = 0.10$				
Yes	200	12.4 (2.0)	12.3 (2.0)	12.4 (2.0)	10.6 (2.3)	10.6 (2.3)	10.7 (2.9)	10.9 (2.4)
	400	12.6 (1.7)	12.4 (1.7)	12.6 (1.8)	10.4 (1.7)	10.4 (1.7)	10.5 (2.1)	10.6 (1.7)
	800	12.8 (1.5)	12.5 (1.5)	12.7 (1.6)	10.2 (1.2)	10.2 (1.1)	10.3 (1.5)	10.3 (1.2)
No	200	13.9 (2.2)	13.9 (2.2)	13.6 (2.3)	10.7 (2.3)	10.8 (2.3)	11.2 (2.7)	11.2 (2.4)
	400	14.5 (1.5)	14.5 (1.5)	14.4 (1.6)	10.2 (1.6)	10.2 (1.6)	10.5 (1.9)	10.5 (1.6)
	800	15.0 (0.8)	15.0 (0.8)	14.9 (1.0)	10.2 (1.2)	10.2 (1.2)	10.4 (1.3)	10.4 (1.2)
				$t = 0.20$				
Yes	200	22.4 (3.6)	22.2 (3.7)	22.5 (3.7)	20.9 (3.1)	20.9 (3.2)	21.1 (4.0)	21.3 (3.2)
	400	22.6 (3.3)	22.3 (3.3)	22.7 (3.3)	20.6 (2.2)	20.6 (2.2)	20.7 (2.8)	20.9 (2.2)
	800	22.8 (2.7)	22.3 (2.8)	22.8 (2.8)	20.2 (1.5)	20.2 (1.6)	20.2 (2.1)	20.4 (1.6)
No	200	25.8 (3.5)	25.7 (3.5)	25.5 (3.6)	20.9 (3.1)	20.9 (3.1)	21.4 (3.7)	21.4 (3.1)
	400	26.9 (2.3)	26.9 (2.3)	26.8 (2.3)	20.5 (2.1)	20.5 (2.1)	20.8 (2.3)	20.8 (2.1)
	800	27.7 (1.1)	27.7 (1.1)	27.5 (1.3)	20.3 (1.6)	20.3 (1.6)	20.5 (1.7)	20.5 (1.6)

Table S4: Mean true positive rate and false positive rate and corresponding standard deviation (in parentheses) for $f(v) = f_2(v) \equiv 1(v < 0) \times (1/(1 + e^{-v/3})) + 1(v \geq 0) \times (1/(1 + e^{-3v}))$ and β_0 = 0.6 across 1000 simulations. n is the size of the training dataset, t is the acceptable false positive rate, GLM denotes standard logistic regression, rGLM denotes robust logistic regression, sTPR denotes the proposed method with the threshold estimated directly, and sTPR(re) denotes the proposed method with the threshold reestimated based on quantiles of the fitted combination. All numbers are percentages.

Outliers	n	True positive rate			False positive rate			
		GLM	rGLM	sTPR	GLM	rGLM	sTPR	sTPR(re)
				$t = 0.05$				
Yes	200	23.0 (8.6)	30.5 (10.9)	31.9 (10.8)	6.4 (3.3)	6.3 (3.4)	8.2 (3.8)	6.8 (3.7)
Yes	400	21.5 (6.9)	31.8 (10.5)	33.5 (10.1)	5.8 (2.3)	5.8 (2.4)	6.7 (2.7)	6.2 (2.6)
Yes	800	20.0 (4.4)	34.6 (9.2)	35.8 (8.5)	5.4 (1.6)	5.3 (1.6)	5.9 (1.8)	5.7 (1.7)
No	200	49.7 (1.5)	49.5 (1.7)	48.6 (4.4)	6.0 (3.5)	5.9 (3.5)	8.6 (4.1)	6.8 (3.7)
No	400	50.3 (0.7)	50.1 (0.8)	49.7 (2.3)	5.5 (2.5)	5.4 (2.5)	6.8 (2.6)	6.1 (2.6)
No	800	50.5 (0.4)	50.5 (0.5)	50.1 (2.0)	5.2 (1.6)	5.2 (1.6)	6.0 (1.7)	5.6 (1.7)
				$t = 0.10$				
Yes	200	37.3 (11.0)	45.7 (13.7)	48.4 (13.2)	11.5 (4.5)	11.6 (4.5)	13.2 (4.6)	12.4 (4.6)
Yes	400	35.2 (8.5)	47.5 (12.9)	50.6 (12.1)	10.8 (3.1)	10.9 (3.2)	11.7 (3.3)	11.4 (3.3)
Yes	800	34.5 (6.6)	51.3 (10.7)	53.6 (10.2)	10.4 (2.2)	10.4 (2.2)	10.8 (2.4)	10.8 (2.3)
No	200	61.3 (1.4)	61.1 (1.6)	60.7 (3.2)	10.9 (4.5)	10.9 (4.5)	13.4 (4.7)	12.1 (4.7)
No	400	61.8 (0.7)	61.6 (0.8)	61.4 (1.2)	10.6 (3.2)	10.6 (3.2)	12.0 (3.3)	11.4 (3.3)
No	800	62.0 (0.4)	62.0 (0.4)	61.8 (0.8)	10.3 (2.3)	10.3 (2.3)	11.1 (2.3)	10.9 (2.4)
				$t = 0.20$				
Yes	200	53.2 (10.6)	60.9 (13.0)	64.2 (12.3)	21.2 (5.9)	21.8 (6.0)	23.1 (6.1)	22.8 (6.0)
Yes	400	52.0 (8.5)	63.5 (11.8)	65.4 (11.3)	20.7 (4.1)	21.1 (4.2)	21.9 (3.9)	21.7 (4.1)
Yes	800	51.1 (6.0)	66.3 (9.7)	68.6 (8.2)	20.4 (3.0)	20.6 (3.0)	21.1 (2.8)	21.1 (3.0)
No	200	73.3 (1.1)	73.1 (1.3)	73.0 (1.5)	21.4 (6.4)	21.4 (6.4)	23.5 (6.1)	22.5 (6.3)
No	400	73.6 (0.6)	73.5 (0.7)	73.5 (0.8)	20.7 (4.4)	20.7 (4.4)	22.0 (4.3)	21.6 (4.4)
No	800	73.8 (0.3)	73.8 (0.4)	73.8 (0.4)	20.4 (3.0)	20.4 (3.0)	21.2 (3.0)	21.0 (3.0)

S4 Densities of predictors in diabetes data

The distributions of the predictors measured in the diabetes study analyzed in Section 5 are depicted in Figure S2.

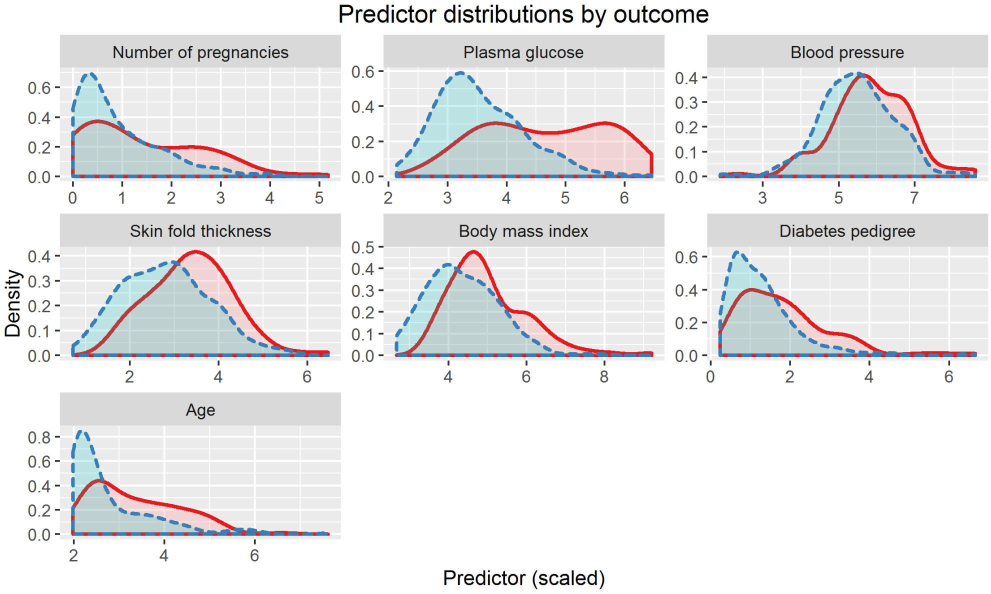

Figure S2: Stratified distributions of the scaled predictors measured in the diabetes study for the observations in the training data. The predictors are number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, and age. The predictor values are shown on the x-axis of each plot. The red solid line represents the distribution among diabetes cases and the blue dotted line represents the distribution among controls.

S5 Proof of Lemma A1

Proof of Lemma A1. Say that both d_n and e_n tend to zero almost surely, and denote by $K > 0$ the Lipschitz constant of f. Suppose that for some $\epsilon > 0$ we have that

$$\text{pr}\left\{ \limsup_n |f(\omega_n) - f(\omega_{0,n})| > \epsilon \right\} > 0 \; .$$

We will show that this leads to a contradiction, and thus that it must be true that $\text{pr}\{\limsup_n |f(\omega_n) - f(\omega_{0,n})| > \epsilon\} = 0$ for each $\epsilon > 0$, thus establishing the desired result.

On a set of probability one, there exists an $n_\epsilon \geq 1$ such that, for each $n \geq n_\epsilon$, there exists $\omega_n^* \in \Omega_0$ and $\omega_{0,n}^* \in \Omega_n$ satisfying $d(\omega_n^*, \omega_n) < \epsilon/(2K)$ and $d(\omega_{0,n}^*, \omega_{0,n}) < \epsilon/(2K)$. Then, on this same set, for $n \geq n_\epsilon$, $|f(\omega_n^*) - f(\omega_n)| \leq \epsilon/2$ and $|f(\omega_{0,n}^*) - f(\omega_{0,n})| \leq \epsilon/2$, so that $f(\omega_{0,n}) \leq f(\omega_{0,n}^*) + \epsilon/2$ and $f(\omega_n) \leq f(\omega_n^*) + \epsilon/2$ in particular. Since $\omega_{0,n}^* \in \Omega_n$ and $\omega_n^* \in \Omega_0$, it must also be true that $f(\omega_{0,n}^*) \leq f(\omega_n) + a_n$ and $f(\omega_n^*) \leq f(\omega_{0,n}) + a_n$. This then implies that $|f(\omega_{0,n}) - f(\omega_n)| \leq \epsilon/2 + a_n$ for all $n \geq n_\epsilon$ on a set of probability one. Since a_n tends to zero deterministically, this yields the sought contradiction.

To establish the last portion of the Lemma, we simply use the first part along with the fact that

$$\left| \sup_{\omega \in \Omega_n} f(\omega) - \sup_{\omega \in \Omega_0} f(\omega) \right| \;\leq\; |f(\omega_{0,n}) - f(\omega_n)| + 2a_n \; .$$

□

S6 Proof of Lemma A2

Proof of Lemma A2. We prove the claim for the false positive rate; the proof for the true positive rate is analogous. We can write

$$\begin{aligned} \sup_{(\theta,\delta)\in\Omega} |\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta) - \text{FPR}(\theta,\delta)| \;\leq\; & \sup_{(\theta,\delta)\in\Omega} |\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta) - E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\}| \\ & + \sup_{(\theta,\delta)\in\Omega} |E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\} - \text{FPR}(\theta,\delta)| \, . \end{aligned}$$

First, we consider $\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta) - E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\}$. We can write this as

$$\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta) - E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\} = \frac{1}{n_0}\sum_{j=1}^{n_0} \Phi\left(\frac{\theta^\top X_{0j} - \delta}{h}\right) - \int \Phi\left(\frac{\theta^\top x - \delta}{h}\right) dF_0(x) \; .$$

The class of functions $\mathcal{G}_1 = \{(\theta,\delta) \mapsto \theta^\top x - \delta : \theta \in \mathbb{R}^p, \delta \in \mathbb{R}, x \in \mathbb{R}^p\}$ is a Vapnik–Chervonenkis (VC) class. Since $u \mapsto \Phi(u/h)$ is monotone for each $h > 0$, the class of functions $\mathcal{G}_2 = \{(\theta,\delta) \mapsto \Phi\{(\theta^\top x - \delta)/h\} : \theta \in \mathbb{R}^p, \delta \in \mathbb{R}, x \in \mathbb{R}^p, h > 0\}$ is also VC (Kosorok, 2008; van der Vaart, 1998; van der Vaart & Wellner, 2000). Since the constant 1 is an applicable envelope function for this class, $\mathcal{G}_2$ is F_0–Glivenko-Cantelli, giving that (Kosorok, 2008; van der Vaart & Wellner, 2000)

$$\sup_{(\theta,\delta)\in\Omega} |\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta) - E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\}| \longrightarrow 0$$

almost surely.

Next, we consider $E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\} - \text{FPR}(\theta,\delta)$. We can write this as

$$E\{\text{F}\tilde{\text{P}}\text{R}_{n_0}(\theta,\delta)\} - \text{FPR}(\theta,\delta) = \int \Phi\left(\frac{\theta^\top x - \delta}{h}\right) dF_0(x) - \text{pr}(\theta^\top X > \delta \mid D = 0).$$

For a general random variable V with distribution function F that is Lipschitz continuous, say with constant $M > 0$, we can write

$$E\left\{\Phi\left(\frac{s-V}{h}\right)\right\} = \int \Phi\left(\frac{s-v}{h}\right) dF(v) = h\int \Phi(u) f(s-hu)du$$

with $u = (s-v)/h$. Using integration by parts and Lemma 2.1 from Winter (1979), this becomes

$$h\int \Phi(u)f(s-hu)du = \int \phi(u)F(s-hu)du\ ,$$

and so, we find that

$$\begin{aligned}\left|E\left\{\Phi\left(\frac{s-V}{h}\right)\right\} - F(s)\right| &= \left|\int \phi(u)F(s-hu)du - F(s)\right| \\ &\leq \int |F(s-hu)-F(s)|\,\phi(u)du \\ &\leq M\int |hu|\phi(u)du = Mh\left(\frac{2}{\pi}\right)^{1/2} .\end{aligned}$$

Since h tends to zero as n tends to infinity, this implies that

$$\sup_s \left|E\left\{\Phi\left(\frac{s-V}{h}\right)\right\} - F(s)\right| = o(1)\ .$$

We now return to $\theta^\top X$ and consider the case $p = 2$, so that $\theta^\top X = \theta_1 X_1 + \theta_2 X_2$. Let $Y_1 = \theta_1 X_1 + \theta_2 X_2$ and $Y_2 = \theta_2 X_2$. Then, we have that $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|\theta_1\theta_2|^{-1}$, where $x_1 = x_1(y_1, y_2) = (y_1 - y_2)/\theta_1$ and $x_2 = x_2(y_1, y_2) = y_2/\theta_2$. We find that

$$\int \Phi\left(\frac{s-\theta^\top x}{h}\right) dF_X(x) = \int \Phi\left(\frac{s-y_1}{h}\right) dF_Y(y) = \int \Phi\left(\frac{s-y_1}{h}\right) dF_{Y_1}(y_1)$$

for any $s \in \mathbb{R}$. Since pr$(\theta^\top X \leq \delta \mid D = 0) =$ pr$(Y_1 \leq \delta \mid D = 0)$, we can write

$$\begin{aligned}&\sup_{(\theta,\delta)\in\Omega}\left|\int \Phi\left(\frac{\theta^\top x-\delta}{h}\right) dF_0(x) - \mathrm{pr}(\theta^\top X > \delta \mid D = 0)\right| \\ &\quad = \sup_{\delta\in\mathbb{R}}\left|\int \Phi\left(\frac{y_1-\delta}{h}\right) dF_{Y_1|D=0}(y_1) - \mathrm{pr}(Y_1 > \delta \mid D = 0)\right| \\ &\quad = \sup_{\delta\in\mathbb{R}}\left|\int \Phi\left(\frac{\delta-y_1}{h}\right) dF_{Y_1|D=0}(y_1) - \mathrm{pr}(Y_1 \leq \delta \mid D = 0)\right|\ ,\end{aligned}$$

implying, in view of condition (4) and the results above, that

$$\sup_{(\theta,\delta)\in\Omega}\left|\int \Phi\left(\frac{\theta^\top x-\delta}{h}\right) dF_0(x) - \mathrm{pr}(\theta^\top X > \delta \mid D = 0)\right| = o(1)\ .$$

The result for $p > 2$ can be proved analogously.

Combining these results, we conclude that $\sup_{(\theta,\delta)\in\Omega}|\mathrm{F\tilde{P}R}_{n_0}(\theta,\delta) - \mathrm{FPR}(\theta,\delta)|$ tends to zero almost surely, as claimed. □

References

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*, pp 155–178. Springer-Verlag New York.

McIntosh, M. W. & Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58,** 657–664.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, pp 265–290. Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (2000). *Weak Convergence and Empirical Processes*, pp 166–168. Springer Series in Statistics.

Winter, B. B. (1979). Convergence rate of perturbed empirical distribution functions. *J. Appl. Probab.* **16,** 163–173.