UW Biostatistics Working Paper Series

7-26-2017

Developing Biomarker Combinations in Multicenter Studies via Direct Maximization and Penalization

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

Chirag R. Parikh
Program of Applied Translational Research, Department of Medicine, Yale School of Medicine, New Haven, CT; Department of Internal Medicine, Vetrans Affairs Medical Center, West Haven, CT, chirag.parikh@yale.edu

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

Developing Biomarker Combinations in Multicenter Studies via Direct Maximization and Penalization

Allison Meisner[1], Chirag R. Parikh[2,3], and Kathleen F. Kerr[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

[2]Program of Applied Translational Research, Department of Medicine, Yale School of Medicine, New Haven, Connecticut, U.S.A.

[3]Department of Internal Medicine, Veterans Affairs Medical Center, West Haven, Connecticut, U.S.A.

Abstract

When biomarker studies involve patients at multiple centers and the goal is to develop biomarker combinations for diagnosis, prognosis, or screening, we consider evaluating the predictive capacity of a given combination with the center-adjusted AUC (aAUC), a summary of conditional performance. Rather than using a general method to construct the biomarker combination, such as logistic regression, we propose estimating the combination by directly maximizing the aAUC. Furthermore, it may be desirable to have a biomarker combination with similar predictive capacity across centers. To that end, we allow for penalization of the variability in center-specific performance. We demonstrate good asymptotic properties of the resulting combinations. Simulations provide small-sample evidence that maximizing the aAUC can lead to combinations with greater predictive capacity than combinations constructed via logistic regression. We further illustrate the utility of constructing combinations by maximizing the aAUC while penalizing variability. We apply these methods to data from a study of acute kidney injury after cardiac surgery.

Keywords: Adjusted AUC; Biomarker combinations; Multicenter; Penalization; Prediction.

1 Introduction

Multicenter studies, where centers could be hospitals, clinics, or providers, have long been used in therapeutic settings as a way to increase power and improve generalizability, and are increasingly common in biomarker studies (e.g., Feldstein et al. (2009); Degos et al. (2010); Nickolas et al. (2012)). Additionally, it is now feasible to measure many biomarkers on each participant. As the performance of individual biomarkers is often modest, there is interest in developing combinations of biomarkers for prognosis, diagnosis, and screening. When studies of multiple biomarkers also involve multiple centers, the central question becomes how such biomarker combinations should be constructed.

One such study is the Translational Research Investigating Biomarker Endpoints in Acute Kidney Injury (TRIBE-AKI) study. The TRIBE-AKI study involves data from 1219 cardiac surgery patients at six centers in North America (Parikh et al., 2011). Study patients were followed for diagnosis of acute kidney injury (AKI) during hospitalization. For each patient, blood and urine were collected at multiple time points pre- and postoperatively, and about two dozen biomarkers were measured at each time point. AKI is typically diagnosed via changes in serum creatinine but these changes often do not happen until several days after the injury. The goal of the study is to identify combinations of biomarkers that can provide an earlier diagnosis of AKI.

Methods to construct biomarker combinations by maximizing the area under the receiver operating characteristic (ROC) curve (AUC) have been proposed. However, in a multicenter setting, there is interest in the conditional, or center-specific, performance. One such summary measure is the center-adjusted AUC (aAUC). We propose a method to construct linear biomarker combinations by targeting the aAUC. We then extend our method to allow for penalization of the variability in center-specific performance; this provides combinations with good overall performance and more similar performance across centers.

2 Background

Let D be a binary outcome, where "cases" have (or will experience) the outcome, denoted by $D = 1$ or the subscript D, and "controls" do not have (or will not experience) the outcome, denoted by $D = 0$ or the subscript $\bar{D}$.

2.1 Center-adjusted AUC

Without loss of generality, we assume that for a given predictor Z, higher values of Z are more indicative of D. Thus, for a particular threshold δ, the true and false positive rates are $P(Z > \delta | D = 1)$ and $P(Z > \delta | D = 0)$, respectively. The ROC curve for Z plots the true positive rate versus the false positive rate over the range of possible thresholds for Z; thus, it exists in the unit square (Pepe, 2003). The predictive capacity of Z is often summarized via the area under the ROC curve (AUC), a measure of the ability of Z to discriminate between cases and controls. The ROC curve for a useless predictor lies on the 45-degree line, and the corresponding AUC is 0.5 (Pepe, 2003). The ROC curve for a perfect predictor reaches the upper left-hand corner of the unit square, and its AUC is 1 (Pepe, 2003). The AUC can also be interpreted as the probability that Z for a randomly chosen case is larger than Z for a randomly chosen control (Pepe, 2003).

In the multicenter setting, Z can be evaluated marginally, by considering the AUC for Z pooled across centers, or conditionally, by summarizing center-specific AUCs. If we consider a measure of marginal performance (i.e., the AUC for Z pooled across centers), we allow center to potentially influence the assessment of the predictive capacity of Z, severely restricting interpretability and generalizability (Janes and Pepe, 2008). Instead, performance should be assessed conditionally and then summarized across centers; this is analogous to the center-adjusted odds ratio in the etiologic setting and the center-adjusted treatment effect in the therapeutic setting (Kahan, 2014; Janes and Pepe, 2008). One such summary measure is the center-adjusted AUC (aAUC).

The center-adjusted ROC ($aROC_Z$) and corresponding center-adjusted AUC ($aAUC_Z$) of Z, proposed by Janes and Pepe (2009), can be written as

$$\begin{aligned} aAUC_Z &= \int_0^1 aROC_Z(t)dt \\ &= \int_0^1 \sum_c ROC_{Z|C=c}(t)P(C=c|D=1)dt \\ &= \sum_c w_c AUC_{Z|C=c}, \end{aligned}$$

where C indicates center, t denotes the false positive rate, $ROC_{Z|C=c}$ and $AUC_{Z|C=c}$ denote the center-specific ROC and AUC, respectively, and $w_c = P(C=c|D=1)$ is the distribution of center among cases. When the center-specific AUCs are constant across centers, the adjusted AUC is simply that center-specific AUC (Janes, Longton, and Pepe, 2009). More generally, the aAUC is a weighted average of the center-specific AUCs (Janes et al., 2009). Weighting by the proportion of cases is appealing because centers with more cases tend to estimate the AUC with more precision than centers with fewer cases (Pepe, 2003). The aAUC is a summary of the accuracy of Z within each center (Janes and Pepe, 2008) and provides an estimate of the performance of Z in new centers, to the extent that the new centers are similar to those used to evaluate Z.

2.2 Biomarker Combinations

Many biomarker assays are now relatively affordable and/or can be used to measure multiple biomarkers at once. This has increased the ability of investigators to measure many biomarkers in each individual, leading to growing interest in developing biomarker combinations for diagnosis, prognosis, or screening.

For a collection of biomarkers $\mathbf{X}$, the combination $P(D = 1|\mathbf{X})$ (and monotone increasing functions of $P(D = 1|\mathbf{X})$) is optimal in terms of maximizing the true positive rate at each false positive rate (McIntosh and Pepe, 2002). Thus, to the extent that the linear logistic model holds, that is, $P(D = 1|\mathbf{X}) = \text{expit}(\boldsymbol{\theta}^\top \mathbf{X})$, the combination $\boldsymbol{\theta}^\top \mathbf{X}$ is optimal. As the linear logistic model may not hold, methods have been developed to construct biomarker combinations by maximizing the AUC without relying on this model (Pepe et al., 2006).

Methods have also been developed to identify combinations of biomarkers that maximize the AUC while accommodating covariates (Liu and Zhou, 2013; Schisterman, Faraggi, and Reiser, 2004). However, implementation of the method proposed by Liu and Zhou (2013) is computationally challenging or prohibitive for more than two biomarkers. The method proposed by Schisterman et al. (2004) assumes that the biomarkers have multivariate normal distributions and requires specification of the relationship between the covariates (i.e., center) and the biomarkers.

When the same data are used to construct a biomarker combination and evaluate its performance (with the aAUC, for example), the resulting estimate of performance is optimistically biased (Copas and Corbett, 2002). This optimistic bias, which we refer to as "resubstitution bias" (Kerr et al., 2015), can be addressed by using a bootstrapping procedure to estimate the optimistic bias and correct the apparent estimate of performance (Copas and Corbett, 2002; Harrell, 2001). Bootstrapping assumes the observations are exchangeable, but in the context of a multicenter study, observations from the same center may be correlated; thus, bootstrap resampling by center has been suggested (Bouwmeester et al., 2013; van Oirbeek and Lesaffre, 2010; Localio et al., 2001; Janes et al., 2009). However, similar results for the average cluster-specific AUC (where in our case, ‘cluster’ is center) have been found whether resampling is done on clusters or individual observations (Bouwmeester et al., 2013).

2.3 Smooth AUC Approximations

Logistic regression models are typically fit by maximizing the logistic likelihood. However, we are interested in using fitted combinations for diagnosis, prognosis, or screening, which motivates maximizing measures of predictive capacity, i.e., matching the objective function to the intended use of the combination (Pepe and Thompson, 2000; Pepe et al., 2006). Thus, as alluded to above, constructing biomarker combinations by directly maximizing the AUC is an appealing alternative to logistic regression. One benefit of directly maximizing the AUC is that the resulting combination is optimal regardless of whether the logistic model holds (Pepe et al., 2006). Furthermore, the AUC of a linear combination constructed by targeting the AUC will be at least as large as the AUC for the individual biomarkers (Pepe and Thompson, 2000). This simple, desirable property might not hold when a combination is constructed by maximizing the likelihood.

In practice, the true AUC for a given vector of coefficients $\boldsymbol{\theta}$ is unknown. Instead, we can consider the empirical AUC, which can be written

$$A\hat{U}C(\boldsymbol{\theta}) = \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} 1(\boldsymbol{\theta}^\top \mathbf{X}_{Di} > \boldsymbol{\theta}^\top \mathbf{X}_{\bar{D}j}),$$

where n_D and $n_{\bar{D}}$ are the number of cases and controls, respectively, $1(\cdot)$ is the indicator function, $\mathbf{X}_{Di}$ denotes the biomarker vector for the i^{th} case, and $\mathbf{X}_{\bar{D}j}$ denotes the biomarker vector for the j^{th} control. Since $A\hat{U}C$ involves indicator functions, direct maximization is challenging. However, smooth approximations to the empirical AUC have been proposed. Lin et al. (2011) used an approximation based on the the probit function to estimate $\boldsymbol{\theta}$:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, R_n(\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\left[\frac{1}{n_D n_{\bar{D}}}\sum_{i=1}^{n_D}\sum_{j=1}^{n_{\bar{D}}}\Phi\left\{\boldsymbol{\theta}^\top(\mathbf{X}_{Di}-\mathbf{X}_{\bar{D}j})/h\right\}\right],\end{aligned}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1\}$, Φ is the standard normal distribution function, and h is a tuning parameter. The function $\Phi(v/h)$ serves as an approximation to the indicator function $I(v > 0)$ and the tuning parameter h represents the trade-off between approximation accuracy and estimation feasibility and tends to zero as the sample size grows (Lin et al., 2011). Lin et al. (2011) noted that if h is too small, estimation will be unstable, and propose choosing the tuning parameter to be $h = \tilde{\sigma}n^{-1/3}$ where $\tilde{\sigma}$ is the sample standard error of $\tilde{\boldsymbol{\theta}}^\top\mathbf{X}$ for the starting value $\tilde{\boldsymbol{\theta}}$. In order to retain identifiability, constraints must be imposed on $\boldsymbol{\theta}$. Specifically, we constrain $||\boldsymbol{\theta}|| = 1$ as in (Fong, Yin, and Huang, 2016).

Due to the smoothness of R_n, gradient-based methods can be used to estimate $\boldsymbol{\theta}$. However, since R_n is not convex, convergence to a global maximum is not guaranteed. Other approximations have been proposed, including the logistic function (Ma and Huang, 2007) and the ramp function (Fong et al., 2016). The probit function approximation tends to be more accurate and stable than the logistic function approximation (Lin et al., 2011) and implementation is more straightforward than for the ramp function approximation.

3 Methods

The population consists of M centers ($M \in [m, \infty]$) where center c has N_c observations, $c = 1, ..., M$. We observe data from m centers with n_c observations from center c, giving n total observations. We consider a p-dimensional vector of biomarkers $\mathbf{X}$. Recall that cases are denoted by either $D = 1$ or the subscript D, and controls are denoted by either $D = 0$ or the subscript $\bar{D}$. Let $(\mathbf{X}, D)$ be the biomarkers and outcome for an arbitrary observation.

We use the subscript i on $\mathbf{X}$ and D to denote the biomarkers and outcome, respectively, for the i^{th} observation. We use the superscript c on $\mathbf{X}$ and D to denote the biomarkers and outcome, respectively, for an observation from cluster c. Throughout, we will assume that the center-specific disease prevalence is non-trivial; that is, $P(D = 1|C = c) \in (0, 1)$, $c = 1, ..., M$. We will consider linear biomarker combinations, as they are often a reasonable choice and have intuitive appeal for clinical collaborators.

3.1 Direct Maximization

We are interested in the aAUC for a combination of the biomarkers defined by $\boldsymbol{\theta}$:

$$
\begin{aligned}
aAUC(\boldsymbol{\theta}) &= \sum_{c=1}^{M} w_c AUC_c(\boldsymbol{\theta}) \\
AUC_c(\boldsymbol{\theta}) &= P(\boldsymbol{\theta}^\top \mathbf{X}_D^c > \boldsymbol{\theta}^\top \mathbf{X}_{\bar{D}}^c),
\end{aligned}
$$

where $w_c = P(C = c|D = 1)$. As with the unadjusted AUC, in practice the aAUC is unknown and we instead consider the empirical aAUC. The empirical aAUC, $aA\hat{U}C$, is based on empirical estimates of the center-specific AUCs, $A\hat{U}C_c$, and the weights, $\hat{w}_c$:

$$
\begin{aligned}
aA\hat{U}C(\boldsymbol{\theta}) &= \sum_{c=1}^{m} \hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) \\
A\hat{U}C_c(\boldsymbol{\theta}) &= \frac{1}{n_D^c n_{\bar{D}}^c} \sum_{i=1}^{n_D^c} \sum_{j=1}^{n_{\bar{D}}^c} 1(\boldsymbol{\theta}^\top \mathbf{X}_{Di}^c > \boldsymbol{\theta}^\top \mathbf{X}_{\bar{D}j}^c) \\
\hat{w}_c &= \frac{n_D^c}{n_D}.
\end{aligned}
$$

Again, $aA\hat{U}C$ is a function of $A\hat{U}C_c$, which involves indicator functions, making direct maximization challenging. However, we can use a smooth approximation to $A\hat{U}C_c$, which in turn provides a smooth approximation to $aA\hat{U}C$. In particular, we propose the SaAUC estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} aR_n(\boldsymbol{\theta}), , \tag{1}$$

where

$$\begin{aligned} aR_n(\boldsymbol{\theta}) &= \sum_{c=1}^{m} \hat{w}_c R^c_{n_c}(\boldsymbol{\theta}) \\ R^c_{n_c}(\boldsymbol{\theta}) &= \frac{1}{n^c_D n^c_{\bar{D}}} \sum_{i=1}^{n^c_D} \sum_{j=1}^{n^c_{\bar{D}}} \Phi\left\{\boldsymbol{\theta}^\top(\mathbf{X}^c_i - \mathbf{X}^c_j)/h_c\right\}, \end{aligned}$$

and h_c is a tuning parameter that tends to zero as n_c grows.

In the above formulation, each center has its own tuning parameter h_c. We choose these tuning parameters to be $h_c = \tilde{\sigma}_c n_c^{-1/3}$, where $\tilde{\sigma}_c$ is the sample standard error of $\tilde{\boldsymbol{\theta}}^\top \mathbf{X}^c$ for the starting value $\tilde{\boldsymbol{\theta}}$. The objective function (1) is a sum of smooth functions, and is therefore also smooth. We constrain $||\boldsymbol{\theta}|| = 1$ and use Lagrange multipliers to incorporate this constraint. Asymptotic results for this method are given in Section 3.3.

3.2 Penalization

In practice, it is unlikely that a given combination will have the same AUC in each center. This could be due to heterogeneity in the biomarker associations and/or heterogeneity in performance due to, for example, differences in the populations of patients at different centers. It may be desirable to construct a biomarker combination that has relatively similar performance across centers. In particular, it may be worth sacrificing a small amount of the overall performance (in terms of the aAUC) for less variability in the center-specific AUCs.

To accomplish this, we propose the following:

$$\hat{\boldsymbol{\theta}}_\lambda = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ aR_n(\boldsymbol{\theta}) - \lambda \sum_{c=1}^{m} \hat{w}_c \left(R^c_{n_c}(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta})\right)^2 \right\},$$

where λ is a fixed penalty parameter, $\lambda \geq 0$. Since $aR_n(\boldsymbol{\theta}) - \lambda \sum_{c=1}^{m} \hat{w}_c \left(R^c_{n_c}(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta})\right)^2$

is the difference of two smooth functions, it can be maximized using gradient-based methods. The goal of this penalized method is to construct a combination whose performance in a new center will be similar to what has been observed in previous centers. Of course, the notion of "similar" depends upon the degree of underlying variability across the population of centers, as well as the centers that have been sampled and can be used to estimate $\boldsymbol{\theta}_\lambda$.

3.3 Asymptotic Results

In the theorem below, we demonstrate good operating characteristics for the combination $\hat{\boldsymbol{\theta}}_\lambda$ in large samples. By setting $\lambda = 0$, we can obtain asymptotic results for the maximization of $aR_n(\boldsymbol{\theta})$ without penalization. Let

$$\tilde{Q}_n(\boldsymbol{\theta};\lambda) = aR_n(\boldsymbol{\theta}) - \lambda \sum_{c=1}^{m} \hat{w}_c \left(R^c_{n_c}(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta}) \right)^2$$
$$Q(\boldsymbol{\theta};\lambda) = aAUC(\boldsymbol{\theta}) - \lambda \sum_{c=1}^{M} w_c \left(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}) \right)^2.$$

We first present several conditions necessary for the theorem:

(A1) The m centers are randomly sampled from the population of M centers, and n_c observations are randomly sampled from center c, $c = 1, ..., m$.

(A2) $\sum_{c=1}^{m} |E(\hat{w}_c) - w_c| \to 0$ as $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n_c}/m \to \infty$.

(A3) The centers are independent and within each center, the observations $O^c_i = (D^c_i, \mathbf{X}^c_i)$, $i = 1, ..., n_c$, are independent and identically distributed $(p+1)$-dimensional random vectors such that there exists at least one component of $\mathbf{X}^c$, X^c_k for some $k \in \{1, ..., p\}$, with distribution that has everywhere positive Lebesgue density, conditional on the other $\mathbf{X}^c$ components.

(A4) The support of $\mathbf{X}^c$, $c = 1, ..., M$, is not contained in any proper linear subspace of $\mathbb{R}^p$.

(A5) For fixed $\lambda \geq 0$, both the maximum of $\tilde{Q}_n(\boldsymbol{\theta}; \lambda)$ and the maximum of $Q(\boldsymbol{\theta}; \lambda)$ over $B = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1, |\theta_k| > 0\}$ are attained.

Theorem 1. *Fix $\lambda \geq 0$ and suppose conditions (A1)–(A5) hold. Then $\sup_{\boldsymbol{\theta} \in B} Q(\boldsymbol{\theta}; \lambda) = Q(\hat{\boldsymbol{\theta}}_\lambda; \lambda) + o_p(1)$ as $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n_c}/m \to \infty$.*

The proof of the theorem is given in Web Appendix A. We have previously demonstrated that, under certain conditions, $A\hat{U}C_c(\boldsymbol{\theta})$ converges uniformly in probability to $AUC_c(\boldsymbol{\theta})$ and $aA\hat{U}C(\boldsymbol{\theta})$ converges uniformly in probability to $aAUC(\boldsymbol{\theta})$, and we use these results in the proof of the theorem (Meisner, Parikh, and Kerr, 2017). Briefly, the proof first demonstrates uniform convergence in probability of the difference between $Q(\boldsymbol{\theta}; \lambda)$ and the empirical analogue of $\tilde{Q}_n(\boldsymbol{\theta}; \lambda)$ (that is, with $A\hat{U}C_c$ in place of $R^c_{n_c}$) to zero. The proof then uses previous results for R_n to demonstrate uniform convergence in probability of the difference between $\tilde{Q}_n(\boldsymbol{\theta}; \lambda)$ and the empirical analogue of $\tilde{Q}_n(\boldsymbol{\theta}; \lambda)$ to zero. Combining these results gives the desired conclusion.

3.3.1 Choosing the Penalization Parameter λ

In other penalized estimation procedures, such as ridge regression or lasso, the penalty parameter λ is typically chosen via cross-validation, where the value of λ that gives the best cross-validated performance is selected. The motivation for cross-validation is that apparent measures of performance for a given model (that is, estimates of performance based on the data used to fit the model) will tend to be optimistic (Hastie, Tibshirani, and Friedman, 2016). Cross-validation is one way of addressing this problem.

For our penalized estimation method, we can extend the ideas behind cross-validation to the multicenter setting. As just described, the goal of cross-validation is typically to get an idea of the performance in new observations. In the case of data from multiple centers, we would like to get an idea of the performance in *new centers.* To that end, we propose the following procedure, which we call "leave one center out cross-validation" (LOCOCV):

1. Choose a sequence of λ values: $\{\lambda_1, \lambda_2, ..., \lambda_r\}$

2. For each value of λ:

   (a) For $i = 1, ..., m$, estimate the biomarker combination using the data from all but the i^{th} center.

   (b) Estimate the AUC of the fitted combination from (a) using the data from the i^{th} center.

3. Plot the m center-specific AUCs from (2b), the corresponding center-adjusted AUC, and the variability in the center-specific AUCs around the center-adjusted AUC (i) in the cross-validation "training" centers and (ii) in the cross-validation "test" centers as a function of λ.

4. Choose an appropriate value of λ, and use this value to estimate the biomarker combination using the data from all m centers.

It is difficult to define "appropriate" when choosing a value of λ. In some situations, it may be preferable to sacrifice a small amount of overall performance (aAUC) in return for substantial decrease in the variability of the center-specific AUCs. In other situations, any decline in overall performance may be very undesirable. Thus, we recommend using the cross-validation plot described above to choose λ, rather than an automated procedure, because the trade-offs for a larger or smaller value of λ may depend on the specific context.

An `R` package including code to implement these methods, `maxadjAUC`, will be publicly available.

4 Results

4.1 Direct Maximization

We used simulations to investigate the performance of the proposed direct maximization method in a variety of situations. These simulations were based in large part on the set-up used by Fong et al. (2016).

In each simulation, we generated a population of centers and individuals, and obtained training data by sampling from this population. In particular, we first sampled m centers from the population of M centers. Then, within each of the m sampled centers, we sampled n_c observations of the N_c observations available in each center (where N_c and n_c did not vary across centers). These observations formed the training data, in which the combinations were constructed. The fitted combinations were then evaluated in independent test data, which consisted of the N_c observations in each of the $M - m$ centers not used in the training data. We considered the following settings: (i) $M = 50, N_c = 5,000, m = 6, n_c = 200$, (ii) $M = 500, N_c = 500, m = 50, n_c = 50$, and (iii) $M = 5000, N_c = 200, m = 500, n_c = 20$.

Fong et al. (2016) noted that the presence of outliers may lead to diminished performance of logistic regression and similar methods, while methods based on maximizing the AUC may be less affected since the AUC is a rank-based measure. Thus, we considered simulations with and without outliers in the data-generating model. We focused on the setting of two biomarkers ($\mathbf{X} = (X_1, X_2)$) and considered

$$\begin{aligned} (\mathbf{X} \mid C) &= \{(1 - \Delta) \times \mathbf{Z}_0\} + \{\Delta \times \mathbf{Z}_1\} \\ (D \mid \mathbf{X}, C) &\sim \text{Bernoulli}\left[f\{\theta_0^C + 4X_1 - 3X_2 - (X_1 - X_2)^3\}\right] \end{aligned}$$

where $\mathbf{Z}_0$ and $\mathbf{Z}_1$ are independent bivariate normal random variables with mean zero and respective covariance matrices

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

$f(v) = \text{expit}(v)$, $\theta_0^C \sim \text{Uniform}(-1, 1)$, and Δ is an independent Bernoulli random variable with success probability π, where $\pi = 0.05$ when outliers were simulated and $\pi = 0$ otherwise. Other simulations with more complex center effects were considered, though the results were largely similar to those based on the scenario described above.

When $m = 6$, estimates from robust logistic regression were used as the starting values, and the proposed SaAUC method was compared to robust logistic regression and standard unconditional logistic regression, both with fixed center-specific intercepts. In particular, we used the robust logistic regression method proposed by Bianco and Yohai (1996). This method uses a deviance function that limits the influence individual observations have on the model fit, making it more robust to outliers than standard (likelihood-based) logistic regression. When $m = 50$ or $m = 500$, we also used conditional logistic regression both to provide starting values for and to compare with the SaAUC method. For all methods, a linear combination was fitted. The simulations were repeated 1000 times.

The results are presented in Table 1 where estimates from robust logistic regression were used as starting values for the SaAUC approach. Clearly, the proposed method outperforms both standard and robust logistic regression, both in terms of the center-adjusted AUC and the center-specific AUCs. There also appears to be less variability in performance across simulations when the SaAUC approach is used to construct combinations. In general, we found that this gave very similar results as when estimates from conditional logistic regression were used for $m = 50$ and $m = 500$ (Web Table 1). As was observed by Fong et al. (2016) for the AUC, when outliers were not present, the three methods produced combinations with comparable performance.

The proposed SaAUC method had excellent convergence rates (less than 0.03% of simulations failed). Robust logistic regression failed to converge in up to 3% of simulations for $m = 50$ and up to 15% for $m = 500$; when this happened, standard unconditional logistic regression was used to obtain starting values. In addition, when simulating data with outliers, in some instances the true biomarker combination was so large that it returned a non-value for the outcome D (in `R`, this occurs for expit(x) when $x > 800$). These observations had to be removed from the simulated dataset, though this happened for less than 0.01% of observations. Finally, for $m = 500$, some of the training data centers were concordant (that

Table 1: Mean (standard deviation) of the aAUC in test data and mean (standard deviation) of the minimum and maximum center-specific AUCs (AUC_c) across the centers in the test data based on combinations fitted by logistic regression (GLM), robust logistic regression (rGLM), and the SaAUC method (SaAUC). Robust logistic regression estimates were used as the starting values for the SaAUC method.

Outliers	$aAUC(\hat{\boldsymbol{\theta}}_{GLM})$	$AUC_c(\hat{\boldsymbol{\theta}}_{GLM})$		$aAUC(\hat{\boldsymbol{\theta}}_{rGLM})$	$AUC_c(\hat{\boldsymbol{\theta}}_{rGLM})$		$aAUC(\hat{\boldsymbol{\theta}}_{SaAUC})$	$AUC_c(\hat{\boldsymbol{\theta}}_{SaAUC})$	
		Min	Max		Min	Max		Min	Max
				$m = 6$					
Yes	0.6244	0.6065	0.6424	0.6492	0.6315	0.6666	0.6856	0.6684	0.7025
	(0.012)	(0.013)	(0.013)	(0.030)	(0.031)	(0.030)	(0.007)	(0.008)	(0.007)
No	0.7032	0.6866	0.7197	0.7032	0.6866	0.7196	0.7030	0.6864	0.7195
	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)
				$m = 50$					
Yes	0.6233	0.5444	0.6992	0.6473	0.5692	0.7215	0.6843	0.6082	0.7564
	(0.008)	(0.014)	(0.012)	(0.027)	(0.030)	(0.026)	(0.004)	(0.011)	(0.009)
No	0.7036	0.6301	0.7731	0.7036	0.6301	0.7731	0.7035	0.6299	0.7730
	(0.001)	(0.009)	(0.008)	(0.001)	(0.009)	(0.008)	(0.001)	(0.010)	(0.008)
				$m = 500$					
Yes	0.6221	0.4683	0.7659	0.6333	0.4798	0.7756	0.6796	0.5287	0.8154
	(0.004)	(0.015)	(0.013)	(0.013)	(0.020)	(0.017)	(0.004)	(0.015)	(0.012)
No	0.7038	0.5574	0.8330	0.7038	0.5574	0.8330	0.7037	0.5573	0.8329
	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)

is, all cases or all controls) and were removed from the analysis. Up to 11% of simulations had one or two concordant training centers.

4.2 Penalized Estimation

We explored our proposed penalized estimation procedure via simulated datasets. In particular, we used individual datasets generated under a variety of models to explore how the method may perform in practice.

As was done in the earlier simulations, we first generated a population of centers and individuals, and obtained training data by sampling from this population. Specifically, we considered a population of $M = 50$ centers with $N_c = 5,000$ observations in each and sampled $m = 6$ centers and $n_c = 200$ observations in each to form the training data, with the observations in the remaining $M - m = 44$ centers serving as test data.

We considered nearly 400 individual datasets; different data-generating mechanisms were used and included variations on the link function, the distribution of the biomarkers across centers, and the degree of heterogeneity in the true biomarker combination across centers. We simulated four independent normally distributed biomarkers with equal variance and throughout, the true, optimal biomarker combination in each center was linear. Estimates from robust logistic regression were used as starting values for the penalized estimation procedure. For each simulation, we applied the LOCOCV procedure described above. We considered 50 values of λ equally-spaced (on the log scale) between 0.1 and 200. This range of values is somewhat arbitrary. In other penalized estimation procedures, it is common to choose the maximum value of λ to be the value that returns coefficient estimates of 0. The analogous requirement in the current setting would be the value of λ that gives center-specific AUCs of 0.5 in all centers. This is only expected to occur when all of the biomarker coefficients are 0, which cannot happen due to the constraint $||\boldsymbol{\theta}|| = 1$ in the penalized estimation procedure. The key point is that the range of λ values used here is meant to be illustrative, not prescriptive.

We present a handful of examples here, and include several more in the Web Figures 1–16. All of the plots we present have the same layout: the left plot gives the training data results, the middle plot gives the results of the LOCOCV procedure, and the right plot gives the test data results. In each plot, the horizontal axis shows $\log_{10}\lambda$. The left vertical axis displays the AUC, and corresponds to the gray lines (center-specific AUCs for the penalized estimation procedure) and the black lines (center-adjusted AUCs for the penalized estimation procedure, robust logistic regression ("rGLM"), and standard logistic regression ("GLM")). The right vertical axis shows the variability in the center-specific performance on the standard deviation scale and corresponds to the red lines (variability relative to the adjusted AUC in the training centers) and blue lines (variability relative to the adjusted AUC in the test centers). In the test data, the centers are so large that the AUCs calculated in these centers are presumed to be equal to the population values. In the training data and cross-validation procedure, on the other hand, the AUCs are empirical estimates. For example, in the test data, for a combination $\hat{\boldsymbol{\theta}}$ estimated in the training data, the variability relative to the training centers is $\sum_{c=1}^{M-m} w_c(AUC_c(\hat{\boldsymbol{\theta}}) - a\widehat{AUC}(\hat{\boldsymbol{\theta}}))^2$ and the variability relative to the test centers is $\sum_{c=1}^{M-m} w_c(AUC_c(\hat{\boldsymbol{\theta}}) - aAUC(\hat{\boldsymbol{\theta}}))^2$, where the w_c's are the weights for the centers in the test data, $a\widehat{AUC}$ is the adjusted AUC among the centers in the training data, and $aAUC$ and AUC_c are the adjusted AUC and the center-specific AUC, respectively, among the centers in the test data. Finally, throughout, the dashed lines represent the standard logistic regression results and the dot-dashed lines represent the robust logistic regression results.

Figures 1 and 2 present examples where the LOCOCV procedure does a particularly nice job of mimicking the patterns in the test data. We encountered some datasets where the penalized estimation procedure did not work as well. For instance, in a small number of datasets, the variability increases with increasing λ in the test data, despite the patterns seen in the training data and the LOCOCV results; Figure 3 presents one such example. In this situation, a value of λ may be chosen that results in a fitted combination with worse overall performance and more variability in center-specific performance than would be obtained without penalization. However, in this example, the drop in overall performance is not large and the variability is fairly small even when λ is large. Our simulations indicate that when the centers in the training data are not representative of the population of centers, the results from the cross-validation procedure may not reflect the patterns in the test data; such discrepancies would be expected in general when a resampling procedure is applied to a non-representative sample.

Problems with convergence were uncommon in our simulations. Out of nearly 400 examples considered, convergence issues were encountered in fewer than 6%. Such issues generally only arose with the more extreme scenarios we considered and primarily occurred during cross-validation. In practice, this may require modification of the range of λ values considered. None of the results presented here had any convergence failures.

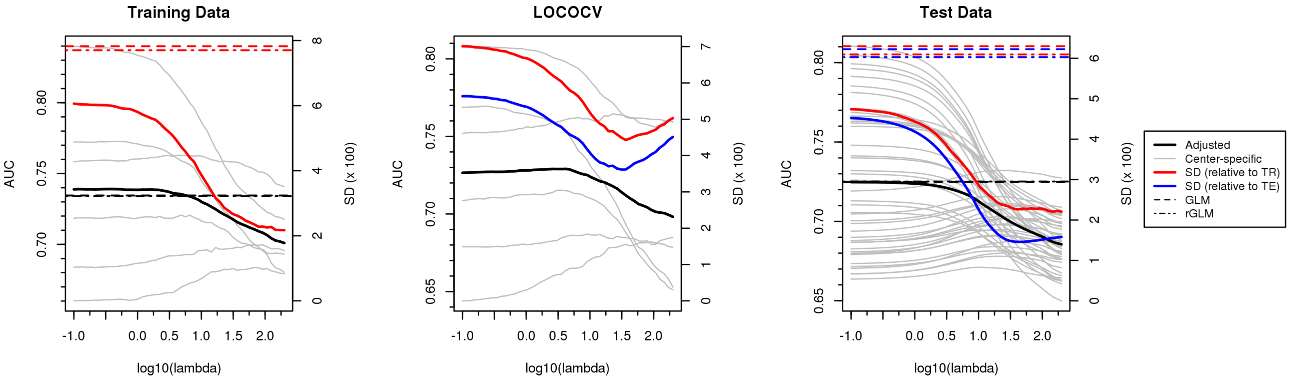

Figure 1: Penalized estimation example 1. In this example, the penalized estimation procedure produces combinations with reduced variability across centers with minimal loss in overall performance (for modest λ values). Importantly, the LOCOCV results mimic what is seen in the test data. This figure appears in color in the electronic version of this article.

4.3 TRIBE-AKI Data

To illustrate the methods we have developed, we used data from the TRIBE-AKI study and constructed combinations of three biomarkers measured immediately after surgery: urine NGAL, plasma h-FABP, and plasma TNI. These data are used as illustration and not to report new findings of the TRIBE-AKI study. We removed observations with missing values for any of the three biomarkers (leaving 962 observations), log-transformed the biomarker values, and scaled the biomarkers to have equal variance. The prevalence of AKI in each center was between 7.8% and 22.9%, and the sizes of the centers ranged from 53 to 483 patients.

We applied standard logistic regression ("GLM"), robust logistic regression ("rGLM"), and the proposed SaAUC method to the TRIBE-AKI study data. The fitted combinations (with normalized coefficients) for GLM, rGLM, and SaAUC were

$$0.0720 * \text{NGAL} + 0.9917 * \text{h-FABP} - 0.1068 * \text{TNI}$$

$$0.0720 * \text{NGAL} + 0.9917 * \text{h-FABP} - 0.1068 * \text{TNI}$$

$$0.0107 * \text{NGAL} + 0.9585 * \text{h-FABP} - 0.2849 * \text{TNI},$$

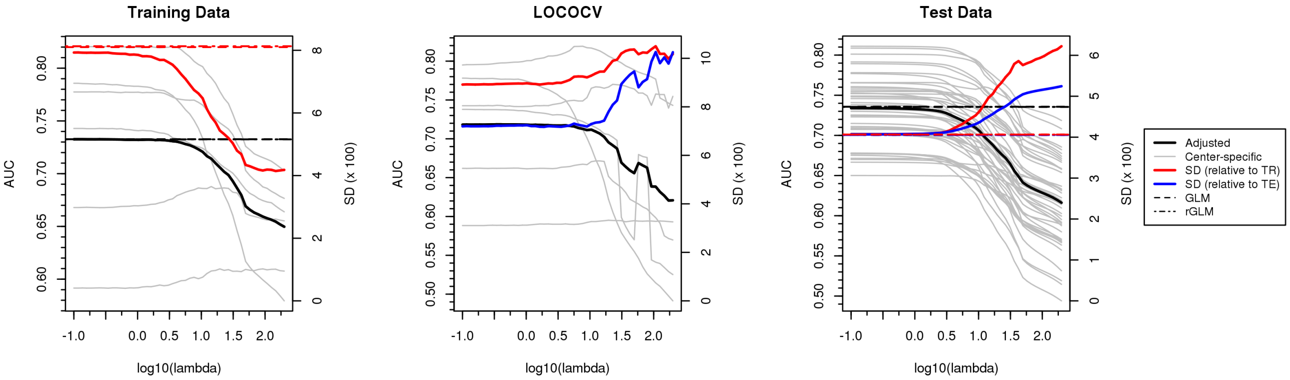

Figure 2: Penalized estimation example 2. This is an example where the LOCOCV results closely mimic the patterns seen in the test data, indicating the importance of performing cross-validation. This figure appears in color in the electronic version of this article.

respectively. The apparent estimated center-adjusted AUCs for these combinations are 0.6878, 0.6878 and 0.6918, respectively. After optimism correction, the center-adjusted AUC estimates are 0.6819, 0.6820 and 0.6825. Thus, it seems that in these data, there is little difference in the performance of the combinations (though there are clear differences in the fitted combinations themselves). Furthermore, there appears to be more optimism in the apparent adjusted AUC estimate for the combination fitted by the SaAUC method, which might be expected in general since the SaAUC method optimizes a smooth approximation to this estimate.

Finally, we applied the proposed penalized estimation method to the TRIBE-AKI study data (Figure 4). The results from the LOCOCV procedure support choosing $\lambda \approx 10^{1.5}$, which is expected to give a reduction in variability in center-specific performance of about 25–30%, with essentially no loss in overall (center-adjusted) performance. In particular, the LOCOCV results indicate that when $\lambda = 0.1$ (the smallest value considered by LOCOCV), the center-specific AUC estimates ranged from 0.6042 to 0.7250, but when $\lambda = 10^{1.5}$, the center-specific AUC estimates were between 0.6270 and 0.6986. Using $\lambda = 10^{1.5}$ in the full

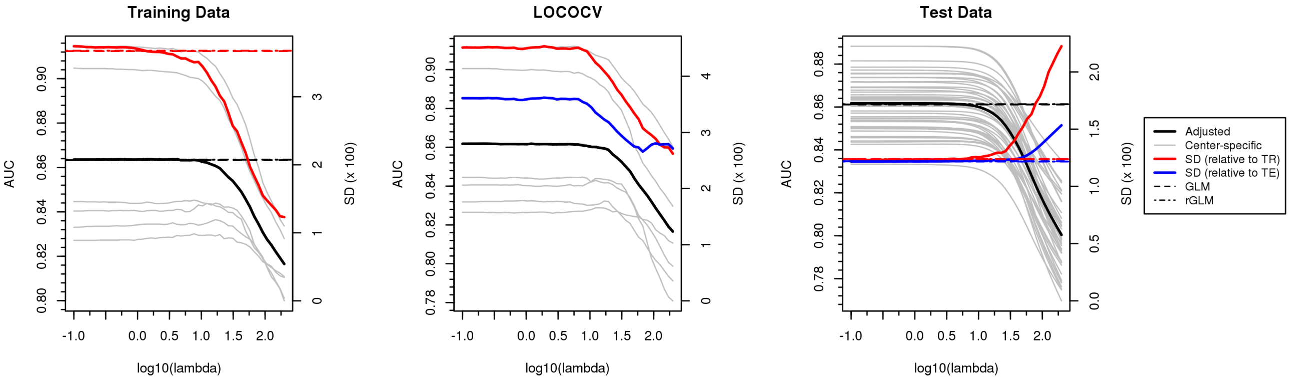

Figure 3: Penalized estimation example 3. This is an example where the penalization procedure does not work as well, since the variability in performance across centers increases with increasing λ, despite the patterns seen in the training data and the LOCOCV. This figure appears in color in the electronic version of this article.

TRIBE-AKI study dataset yielded the combination

$$-0.1067 * \text{NGAL} + 0.9911 * \text{h-FABP} + 0.0798 * \text{TNI}.$$

5 Discussion

We have developed a method to construct biomarker combinations by maximizing a smooth approximation to the center-adjusted AUC. This method is directly applicable to the covariate-adjusted AUC for any discrete covariate, and so could be applied beyond the multicenter setting. In addition, we have incorporated a penalty term that can be used to encourage similarity in performance across centers. This penalized estimation approach could be useful in other settings with discrete nuisance covariates, such as batch. We used data on biomarkers measured after cardiac surgery to construct combinations for the diagnosis acute kidney injury, demonstrating the feasibility of our methods.

A limitation of the methods we have proposed is that they cannot be used to generate predicted probabilities, as they do not relate the biomarker combination to $P(D = 1)$. Thus,

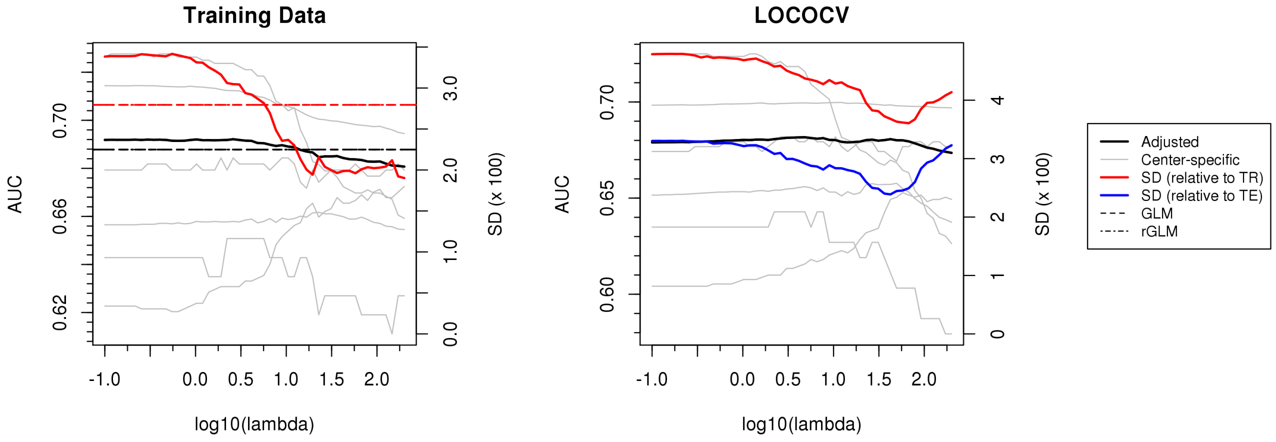

Figure 4: Penalized estimation method applied to the TRIBE-AKI study data. The results from the LOCOCV procedure support choosing $\lambda \approx 10^{1.5}$, which is expected to give a reduction in variability in center-specific performance of about 25–30%. This figure appears in color in the electronic version of this article.

our methods provide tools for risk stratification within each center rather than risk prediction models. In addition, in multicenter studies, different sampling schemes could be used (e.g., case-control or stratified case-control sampling). The estimated weights $\hat{w}_c$ would potentially be affected by different sampling procedures, and may not reflect $P(C = c|D = 1)$. This would in turn affect the interpretation of the center-adjusted AUC, though it would still be a summary of the conditional performance. Our methods would then be optimizing this summary of conditional performance. The sampling scheme could also affect the validity of the asymptotic results we have provided. Furthermore, if a study involves matching, our methods would need to be modified to adjust the AUC for the matching in addition to center (Janes and Pepe, 2008).

The adjusted AUC is a reasonable estimand even when the center-specific AUCs of a given combination are not the same across centers, though it is helpful to consider the center-specific AUCs, as these provide some insight into how the combination might be expected to perform in a new center. In addition, differences in performance across centers may be scientifically meaningful and merit further investigation. When assessing center-specific AUCs, it is important to also consider the sizes of the centers, as estimates of center-specific AUCs from small centers may be highly variable. One feature of our penalization approach is the use of the weights $\hat{w}_c$ in the penalty function, which reflect the proportion of cases in each center and so will tend to give less weight to small centers. Furthermore, the optimal combination (in terms of the center-specific AUC) may be different for each center. Importantly, however, our aim is not to identify the optimal combination in every center; instead, we are interested in constructing a single combination that performs well across centers.

Since our smooth approximation function is not convex, further research is needed on the choice of starting values. It may also be possible to extend the method proposed by Fong et al. (2016), which optimizes the convex ramp function approximation to the AUC, to the center-adjusted AUC. This may lead to further improvements in performance over logistic regression, as was seen in Fong et al. (2016) for the unadjusted AUC. When the centers are very small, as when "centers" are clinicians, the empirical center-specific AUC will be unreliable. Research is needed into the use of other (possibly parametric) methods to estimate the center-specific AUC by borrowing information across centers, which may be useful when the centers are small. Extensions of the methods we have proposed to other center-adjusted measures of performance, such as the partial AUC or the true positive rate for a fixed false positive rate, is another avenue for future research.

6 Supplementary Materials

Appendices, Tables, and Figures referenced in Sections 3.3, 4.1, and 4.2 are available with this paper.

Acknowledgements

The authors wish to acknowledge the TRIBE-AKI study investigators: Steven G. Coca (Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, New

York), Amit X. Garg (Institute for Clinical Evaluative Sciences Western, London, Ontario, Canada; Division of Nephrology, Department of Medicine, and Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada), Jay Koyner (Section of Nephrology, Department of Medicine, University of Chicago Pritzker School of Medicine, Chicago, Illinois), and Michael Shlipak (Kidney Health Research Collaborative, San Francisco Veterans Affairs Medical Center, University of California, San Francisco, San Francisco, California).

Urine NGAL, plasma h-FABP, and plasma cardiac troponin I assays used in the TRIBE-AKI study were donated by Abbott Diagnostics, Randox Laboratories, and Beckman Coulter, respectively.

This work was supported by the National Institutes of Health (F31 DK108356, R01 HL085757, K24 DK090203). The opinions, results, and conclusions reported in this article are those of the authors and are independent of the funding sources.

References

Bianco, A. M. and Yohai, V. J. (1996). Robust Estimation in the Logistic Regression Model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods,* pages 17–34. New York: Springer-Verlag.

Bouwmeester, W., Moons, K. G. M., Kappen, T. H., van Klei, W. A., Twisk, J. W. R., Eijkemans, M. J. C., et al. (2013). Internal validation of risk models in clustered data: a comparison of bootstrap schemes. *American Journal of Epidemiology* **177,** 1209–17.

Copas, J. B. and Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* **89,** 315–331.

Degos, F., Perez, P., Roche, B., Mahmoudi, A., Asselineau, J., Voitot, H., et al. (2010). Diagnostic accuracy of FibroScan and comparison to liver fibrosis biomarkers in chronic viral hepatatis: a multicenter prospective study (the FIBROSTIC study). *Hepatology* **53,** 1013–1021.

Feldstein, A. E., Wieckowska, A., Lopez, A. R., Liu, Y.-C., Zein, N. N., and McCullough, A. J. (2009). Cytokeratin-18 fragment levels as noninvasive biomarkers for nonalcoholic steatohepatitis: a multicenter validation study. *Hepatology* **50,** 1072–1078.

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

Harrell, F. E. (2001). *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning.* Springer Science & Business Media, New York.

Janes, H., Longton, G., and Pepe, M. (2009). Accommodating covariates in ROC analysis. *Stata Journal* **9,** 17–39.

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

Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* **96,** 1–12.

Kahan, B. C. (2014). Accounting for centre-effects in multicentre trials with a binary outcome - when, why and how? *BMC Medical Research Methodology* **14**.

Kerr, K. F., Meisner, A., Thiessen-Philbrook, H., Coca, S. G., and Parikh, C. R. (2015). RiGoR: reporting guidelines to address common sources of bias in risk model development. *Biomarker Research* **3**.

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

Liu, D. and Zhou, X.-H. (2013). ROC analysis in biomarker combination with covariate adjustment. *Academic Radiology* **20,** 874–882.

Localio, A. R., Berlin, J. A., Ten Have, T. R., and Kimmel, S. E. (2001). Adjustments for center in multicenter studies: an overview. *Annals of Internal Medicine* **135,** 112–123.

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

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

Meisner, A., Parikh, C. R. and Kerr, K. F. (2017). Biomarker combinations for diagnosis and prognosis in multicenter studies: principles and methods. *UW Biostatistics Working Paper Series,* Working Paper 419.

Nickolas, T. L., Schmidt-Ott, K. M., Canetta, P., Forster, C., Singer, E., Sise, M., et al. (2012). Diagnostic and prognostic stratification in the emergency department using urinary biomarkers of nephron damage. *Journal of the American College of Cardiology* **59,** 246–255.

Parikh, C. R., Coca, S. G., Thiessen-Philbrook, H., Shlipak, M. G., Koyner, J. L., Wang, Z., et al. (2011). Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. *Journal of the American Society of Nephrology* **22,** 1748–1757.

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

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

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

Schisterman, E. F., Faraggi, D., and Reiser, B. (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine* **23,** 3319–3331.

van Oirbeek, R. and Lesaffre, E. (2010). An application of Harrell's C-index to PH frailty models. *Statistics in Medicine* **29,** 3160–3171.

Web-based Supplementary Materials for "Developing Biomarker Combinations in Multicenter Studies via Direct Maximization and Penalization"

Allison Meisner[1], Chirag R. Parikh[2,3], and Kathleen F. Kerr[1]

[1]Department of Biostatistics, University of Washington, Seattle, Washington, U.S.A.

[2]Program of Applied Translational Research, Department of Medicine, Yale School of Medicine, New Haven, Connecticut, U.S.A.

[3]Department of Internal Medicine, Veterans Affairs Medical Center, West Haven, Connecticut, U.S.A.

Web Appendix A

Proof of Theorem 1. First we will show $\sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda)-Q(\boldsymbol{\theta};\lambda)\right| = o_p(1)$. Let

$$Q_n(\boldsymbol{\theta};\lambda) = aA\hat{U}C(\boldsymbol{\theta}) - \lambda\sum_{c=1}^{m}\hat{w}_c\left(A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta})\right)^2.$$

We can write

$$\sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda)-Q(\boldsymbol{\theta};\lambda)\right| \le \sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda)-Q_n(\boldsymbol{\theta};\lambda)\right| + \sup_{\boldsymbol{\theta}\in B}\left|Q_n(\boldsymbol{\theta};\lambda)-Q(\boldsymbol{\theta};\lambda)\right|.$$

Under conditions (A1)–(A4), we have shown (Lemma 1 and Theorem 1 in Meisner, Parikh, and Kerr (2017)) that $\sup_{\boldsymbol{\theta}\in B}\left|aA\hat{U}C(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})\right| = o_p(1)$ and $\sup_{\boldsymbol{\theta}\in B}\left|A\hat{U}C_c(\boldsymbol{\theta}) - AUC_c(\boldsymbol{\theta})\right| = o_p(1), \quad c = 1, ..., M$. We can write

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta}\in B}|Q_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)| \\
&\quad \leq \sup_{\boldsymbol{\theta}\in B}\left|aA\hat{U}C(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})\right| \\
&\qquad + \lambda \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{M} w_c(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}))^2 - \sum_{c=1}^{m} \hat{w}_c(A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}))^2\right| \\
&\quad \leq \sup_{\boldsymbol{\theta}\in B}\left|aA\hat{U}C(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})\right| \\
&\qquad + \lambda \sum_{c=m+1}^{M} \sup_{\boldsymbol{\theta}\in B}\left|w_c(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}))^2\right| \\
&\qquad + \lambda \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\left\{w_c(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}))^2 - \hat{w}_c(A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}))^2\right\}\right|,
\end{aligned}
$$

where $\sum_{c=m+1}^{M}\sup_{\boldsymbol{\theta}\in B}|w_c(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}))^2| = o(1)$ as $m \to M$. Then by Theorem 1 in Meisner et al. (2017),

$$
\sup_{\boldsymbol{\theta}\in B}|Q_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)| \leq o_p(1) + o(1) + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B}\left|w_c Y_1^c(\boldsymbol{\theta})^2 - \hat{w}_c(Y_2^c(\boldsymbol{\theta}) + Y_1^c(\boldsymbol{\theta}) + Y_3(\boldsymbol{\theta}))^2\right|,
$$

where $Y_1^c(\boldsymbol{\theta}) = AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})$, $Y_2^c(\boldsymbol{\theta}) = A\hat{U}C_c(\boldsymbol{\theta}) - AUC_c(\boldsymbol{\theta})$, and $Y_3(\boldsymbol{\theta}) = aAUC(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta})$; note that $|Y_1^c(\boldsymbol{\theta})| \leq 1$, $|Y_2^c(\boldsymbol{\theta})| \leq 1$ and $|Y_3(\boldsymbol{\theta})| \leq 1$. Then

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in B} |Q_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)| \\
&\quad \leq o_p(1) + o(1) + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} \big| (w_c - \hat{w}_c) Y_1^c(\boldsymbol{\theta})^2 \\
&\qquad - \hat{w}_c \left\{ Y_2^c(\boldsymbol{\theta})^2 + Y_3(\boldsymbol{\theta})^2 + 2Y_1^c(\boldsymbol{\theta})Y_2^c(\boldsymbol{\theta}) + 2Y_1^c(\boldsymbol{\theta})Y_3(\boldsymbol{\theta}) + 2Y_2^c(\boldsymbol{\theta})Y_3(\boldsymbol{\theta}) \right\} \big| \\
&\quad \leq o_p(1) + o(1) + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} \big| (w_c - \hat{w}_c) Y_1^c(\boldsymbol{\theta})^2 \big| + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} \big| -\hat{w}_c Y_2^c(\boldsymbol{\theta})^2 \big| \\
&\qquad + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} \big| -\hat{w}_c Y_3(\boldsymbol{\theta})^2 \big| + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_1^c(\boldsymbol{\theta}) Y_2^c(\boldsymbol{\theta})| \\
&\qquad + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_1^c(\boldsymbol{\theta}) Y_3(\boldsymbol{\theta})| + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_2^c(\boldsymbol{\theta}) Y_3(\boldsymbol{\theta})| .
\end{aligned}$$

We have (by Lemma 1 and Theorem 1 in Meisner et al. (2017)) $\sup_{\boldsymbol{\theta}\in B} |Y_2^c(\boldsymbol{\theta})| = o_p(1), c = 1, ..., M$ and $\sup_{\boldsymbol{\theta}\in B} |Y_3(\boldsymbol{\theta})| = o_p(1)$. This gives

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in B} \big| (w_c - \hat{w}_c) \left[Y_1^c(\boldsymbol{\theta})\right]^2 \big| &= |w_c - \hat{w}_c| \sup_{\boldsymbol{\theta}\in B} \left[Y_1^c(\boldsymbol{\theta})\right]^2 \leq |w_c - \hat{w}_c| \\
\sup_{\boldsymbol{\theta}\in B} \big| -\hat{w}_c \left[Y_2^c(\boldsymbol{\theta})\right]^2 \big| &= \hat{w}_c \sup_{\boldsymbol{\theta}\in B} \left[Y_2^c(\boldsymbol{\theta})\right]^2 \leq \hat{w}_c \sup_{\boldsymbol{\theta}\in B} |Y_2^c(\boldsymbol{\theta})| = \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B} \big| -\hat{w}_c \left[Y_3(\boldsymbol{\theta})\right]^2 \big| &= \hat{w}_c \sup_{\boldsymbol{\theta}\in B} \left[Y_3(\boldsymbol{\theta})\right]^2 \leq \hat{w}_c \sup_{\boldsymbol{\theta}\in B} |Y_3(\boldsymbol{\theta})| = \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_1^c(\boldsymbol{\theta}) Y_2^c(\boldsymbol{\theta})| &= \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_1^c(\boldsymbol{\theta}) Y_3(\boldsymbol{\theta})| &= \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B} |-2\hat{w}_c Y_2^c(\boldsymbol{\theta}) Y_3(\boldsymbol{\theta})| &= \hat{w}_c o_p(1).
\end{aligned}$$

Since $\sum_{c=1}^{m} \hat{w}_c = 1$ for every m,

$$\sup_{\boldsymbol{\theta}\in B} |Q_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)| \leq o_p(1) + o(1) + \lambda \sum_{c=1}^{m} |w_c - \hat{w}_c| .$$

Furthermore, we have previously shown (Theorem 1 in Meisner et al. (2017)) that $\sum_{c=1}^{m} |w_c - \hat{w}_c| = o_p(1)$ as $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n}_c/m \to \infty$. Thus, $\sup_{\boldsymbol{\theta} \in B} |Q_n(\boldsymbol{\theta}; \lambda) - Q(\boldsymbol{\theta}; \lambda)| = o_p(1)$.

Now consider $\sup_{\boldsymbol{\theta} \in B} \left| \tilde{Q}_n(\boldsymbol{\theta}; \lambda) - Q_n(\boldsymbol{\theta}; \lambda) \right|$. We will first show $\sup_{\boldsymbol{\theta} \in B} \left| aR_n(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}) \right| = o_p(1)$. Ma and Huang (2007) demonstrated that $\sup_{\boldsymbol{\theta} \in B} \left| R^c_{n_c}(\boldsymbol{\theta}) - A\hat{U}C_c(\boldsymbol{\theta}) \right| = o_p(1)$ as $n_c \to \infty$. We can write

$$\sup_{\boldsymbol{\theta} \in B} \left| aR_n(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}) \right| \leq \sum_{c=1}^{m} \hat{w}_c \sup_{\boldsymbol{\theta} \in B} \left| R^c_{n_c}(\boldsymbol{\theta}) - A\hat{U}C_c(\boldsymbol{\theta}) \right| \leq \sum_{c=1}^{m} \hat{w}_c o_p(1) = o_p(1)$$

since $\sum_{c=1}^{m} \hat{w}_c = 1$ for every m.

Now consider $\sup_{\boldsymbol{\theta} \in B} \left| \tilde{Q}_n(\boldsymbol{\theta}; \lambda) - Q_n(\boldsymbol{\theta}; \lambda) \right|$. We can write

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in B} \left| \tilde{Q}_n(\boldsymbol{\theta}; \lambda) - Q_n(\boldsymbol{\theta}; \lambda) \right| \\
&\quad \leq \sup_{\boldsymbol{\theta} \in B} \left| aR_n(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}) \right| \\
&\qquad + \lambda \sup_{\boldsymbol{\theta} \in B} \left| \sum_{c=1}^{m} \left\{ \hat{w}_c (A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta}))^2 - \hat{w}_c (R^c_{n_c}(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta}))^2 \right\} \right| \\
&\quad \leq o_p(1) + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| \hat{w}_c Z_1^c(\boldsymbol{\theta})^2 - \hat{w}_c (Z_2^c(\boldsymbol{\theta}) + Z_1^c(\boldsymbol{\theta}) + Z_3(\boldsymbol{\theta}))^2 \right|,
\end{aligned}$$

where $Z_1^c(\boldsymbol{\theta}) = A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta})$, $Z_2^c(\boldsymbol{\theta}) = R^c_{n_c}(\boldsymbol{\theta}) - A\hat{U}C_c(\boldsymbol{\theta})$, and $Z_3(\boldsymbol{\theta}) = aA\hat{U}C(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta})$; note that $|Z_1^c(\boldsymbol{\theta})| \leq 1$, $|Z_2^c(\boldsymbol{\theta})| \leq 1$ and $|Z_3(\boldsymbol{\theta})| \leq 1$. This gives

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in B} \left| \tilde{Q}_n(\boldsymbol{\theta}; \lambda) - Q_n(\boldsymbol{\theta}; \lambda) \right| \leq o_p(1) &+ \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| -\hat{w}_c Z_2^c(\boldsymbol{\theta})^2 \right| \\
&+ \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| -\hat{w}_c Z_3(\boldsymbol{\theta})^2 \right| + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| -2\hat{w}_c Z_1^c(\boldsymbol{\theta}) Z_2^c(\boldsymbol{\theta}) \right| \\
&+ \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| -2\hat{w}_c Z_1^c(\boldsymbol{\theta}) Z_3(\boldsymbol{\theta}) \right| + \lambda \sum_{c=1}^{m} \sup_{\boldsymbol{\theta} \in B} \left| -2\hat{w}_c Z_2^c(\boldsymbol{\theta}) Z_3(\boldsymbol{\theta}) \right|.
\end{aligned}$$

We have that $\sup_{\boldsymbol{\theta}\in B}|Z_2^c(\boldsymbol{\theta})| = o_p(1), c = 1, ..., M$ and $\sup_{\boldsymbol{\theta}\in B}|Z_3(\boldsymbol{\theta})| = o_p(1)$. This gives

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in B}\left|-\hat{w}_c\left[Z_2^c(\boldsymbol{\theta})\right]^2\right| &= \hat{w}_c \sup_{\boldsymbol{\theta}\in B}\left[Z_2^c(\boldsymbol{\theta})\right]^2 \leq \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B}\left|-\hat{w}_c\left[Z_3(\boldsymbol{\theta})\right]^2\right| &= \hat{w}_c \sup_{\boldsymbol{\theta}\in B}\left[Z_3(\boldsymbol{\theta})\right]^2 \leq \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B}\left|-2\hat{w}_c Z_1^c(\boldsymbol{\theta}) Z_2^c(\boldsymbol{\theta})\right| &= \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B}\left|-2\hat{w}_c Z_1^c(\boldsymbol{\theta}) Z_3(\boldsymbol{\theta})\right| &= \hat{w}_c o_p(1) \\
\sup_{\boldsymbol{\theta}\in B}\left|-2\hat{w}_c Z_2^c(\boldsymbol{\theta}) Z_3(\boldsymbol{\theta})\right| &= \hat{w}_c o_p(1).
\end{aligned}$$

Since $\sum_{c=1}^m \hat{w}_c = 1$ for every m, we have $\sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda) - Q_n(\boldsymbol{\theta};\lambda)\right| = o_p(1)$.

Combining these results, we have $\sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)\right| = o_p(1)$. Then

$$\begin{aligned}
\left|Q(\hat{\boldsymbol{\theta}}_\lambda;\lambda) - \sup_{\boldsymbol{\theta}\in B} Q(\boldsymbol{\theta};\lambda)\right| &\leq \left|\sup_{\boldsymbol{\theta}\in B} Q(\boldsymbol{\theta};\lambda) - \sup_{\boldsymbol{\theta}\in B}\tilde{Q}_n(\boldsymbol{\theta};\lambda)\right| + \left|\sup_{\boldsymbol{\theta}\in B}\tilde{Q}_n(\boldsymbol{\theta};\lambda) - Q(\hat{\boldsymbol{\theta}}_\lambda;\lambda)\right| \\
&\leq \sup_{\boldsymbol{\theta}\in B}\left|Q(\boldsymbol{\theta};\lambda) - \tilde{Q}_n(\boldsymbol{\theta};\lambda)\right| + \left|\tilde{Q}_n(\hat{\boldsymbol{\theta}}_\lambda;\lambda) - Q(\hat{\boldsymbol{\theta}}_\lambda;\lambda)\right| \\
&\leq o_p(1) + \sup_{\boldsymbol{\theta}\in B}\left|\tilde{Q}_n(\boldsymbol{\theta};\lambda) - Q(\boldsymbol{\theta};\lambda)\right| = o_p(1),
\end{aligned}$$

giving $\sup_{\boldsymbol{\theta}\in B} Q(\boldsymbol{\theta};\lambda) = Q(\hat{\boldsymbol{\theta}}_\lambda;\lambda) + o_p(1)$ as $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n_c}/m \to \infty$. □

Web Table 1

Table 1: Mean (standard deviation) of the aAUC in test data and mean (standard deviation) of the minimum and maximum center-specific AUCs (AUC_c) across the centers in the test data based on combinations fitted by conditional logistic regression ($\hat{\boldsymbol{\theta}}_{GLM}$) and the SaAUC method ($\hat{\boldsymbol{\theta}}_{SaAUC}$). Conditional logistic regression estimates were used as the starting values for the SaAUC method.

Outliers	$aAUC(\hat{\boldsymbol{\theta}}_{GLM})$	$AUC_c(\hat{\boldsymbol{\theta}}_{GLM})$		$aAUC(\hat{\boldsymbol{\theta}}_{SaAUC})$	$AUC_c(\hat{\boldsymbol{\theta}}_{SaAUC})$	
		Min	Max		Min	Max
			$m = 50$			
Yes	0.6233	0.5444	0.6992	0.6824	0.6062	0.7547
	(0.008)	(0.014)	(0.012)	(0.004)	(0.011)	(0.009)
No	0.7036	0.6301	0.7731	0.7035	0.6299	0.7730
	(0.001)	(0.009)	(0.008)	(0.001)	(0.010)	(0.008)
			$m = 500$			
Yes	0.6221	0.4684	0.7659	0.6764	0.5253	0.8128
	(0.004)	(0.015)	(0.013)	(0.003)	(0.015)	(0.012)
No	0.7038	0.5574	0.8330	0.7037	0.5573	0.8329
	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)

Web Figure 1

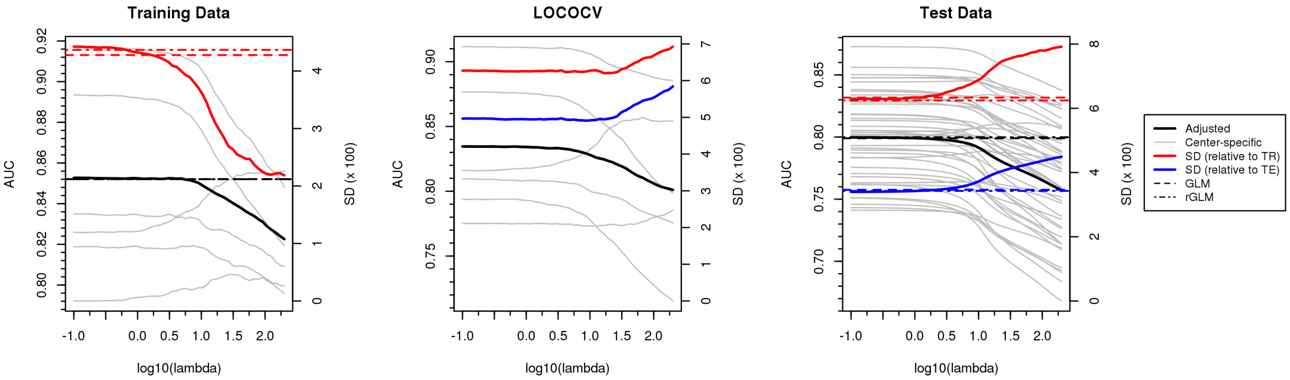

Figure 1: Penalized estimation example 3. This is an example where the LOCOCV procedure does very well in mimicking the patterns seen in the test data.

Web Figure 2

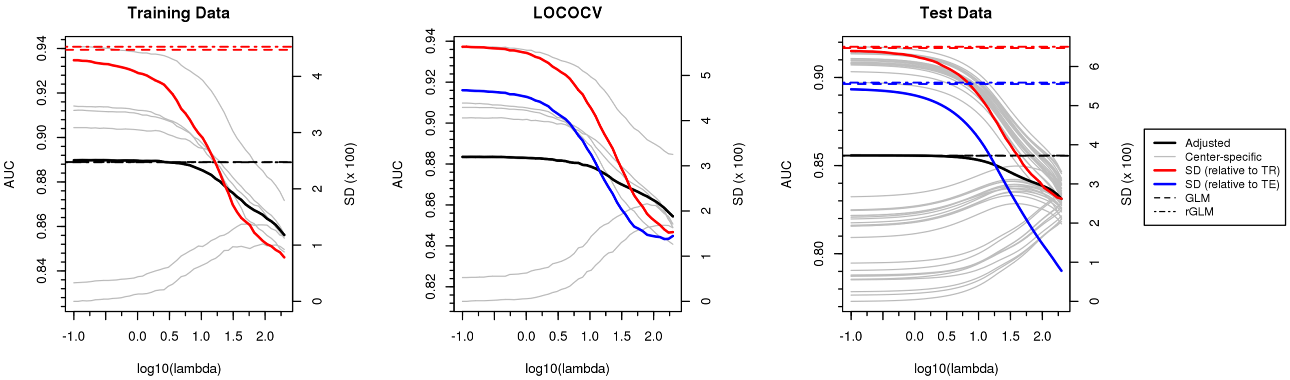

Figure 2: Penalized estimation example 4. This illustrates a setting where there is a clear benefit to penalizing as there is a reduction in variability with little loss in overall performance.

Web Figure 3

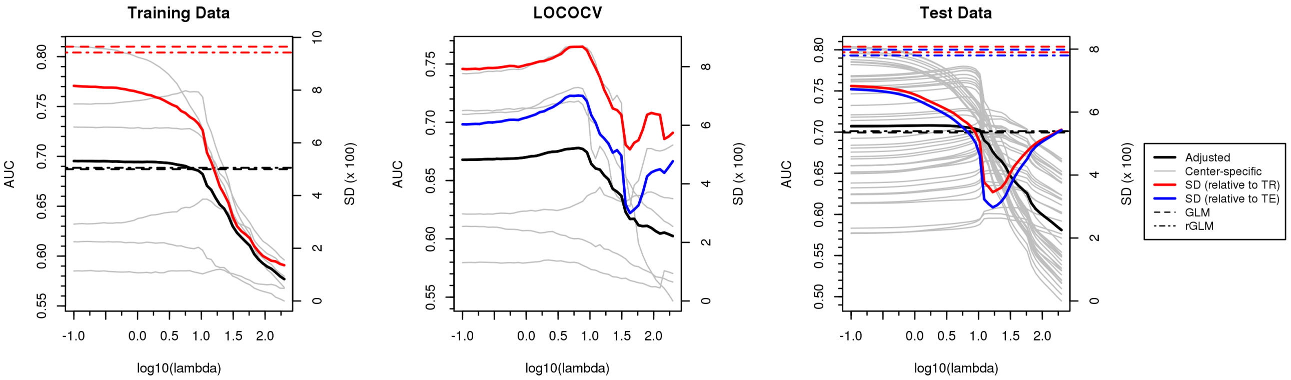

Figure 3: Penalized estimation example 5. This is an example where the LOCOCV results are inconclusive in terms of which value of λ should be chosen.

Web Figure 4

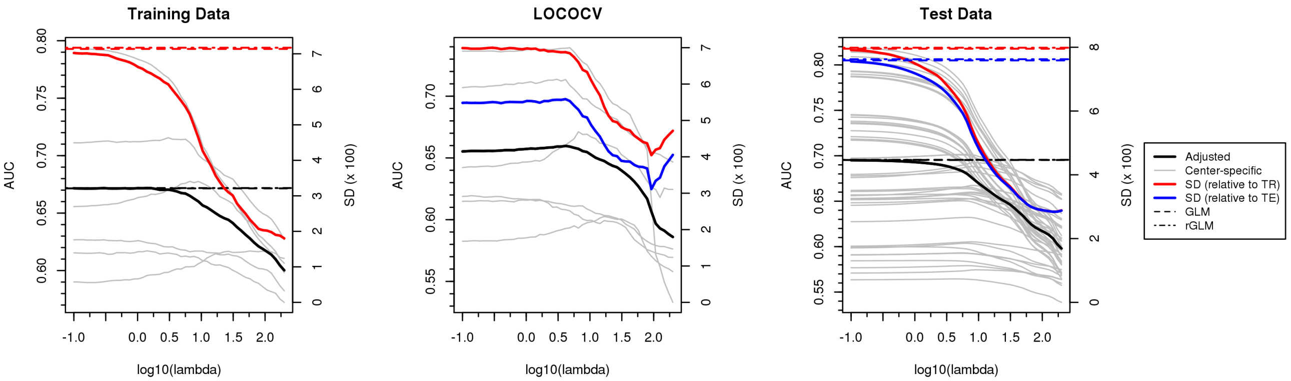

Figure 4: Penalized estimation example 6. This is an example where the penalization procedure does not work as well, since in the test data, the aAUC decreased more quickly with increasing λ than was indicated by the LOCOCV procedure and the training data.

Web Figure 5

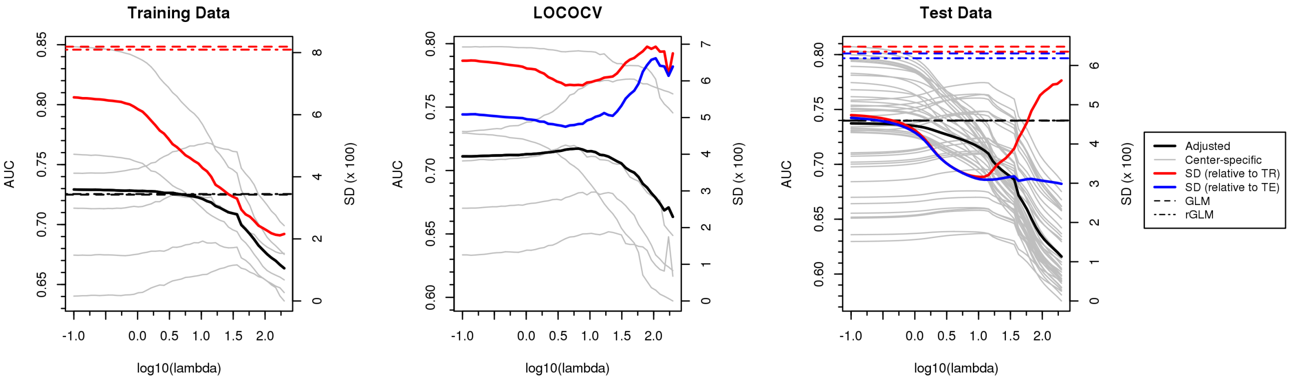

Figure 5: Penalized estimation example 7. In this example, the LOCOCV procedure does a nice job capturing the trends seen in the test data.

Web Figure 6

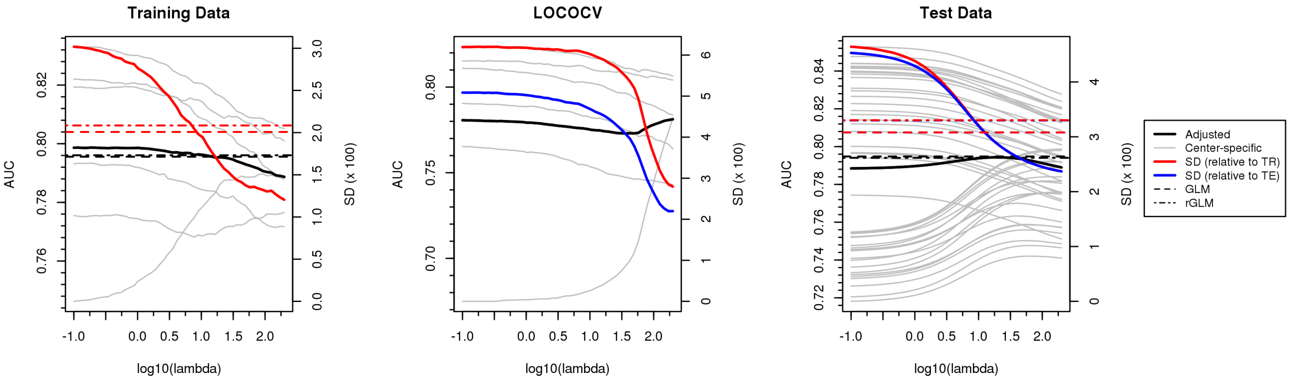

Figure 6: Penalized estimation example 8. In this example, there is a clear benefit to penalization.

Web Figure 7

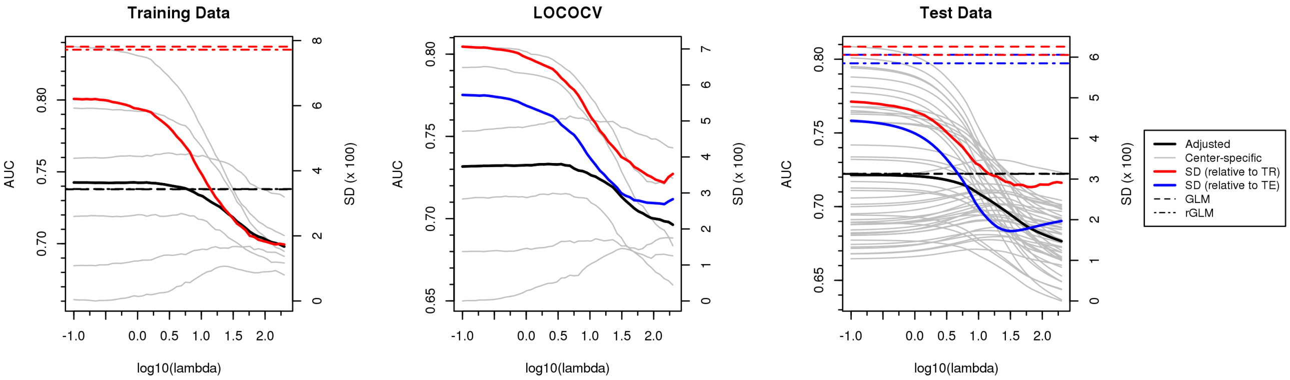

Figure 7: Penalized estimation example 9. In this example, there is a clear benefit to penalization for a range of λ values.

Web Figure 8

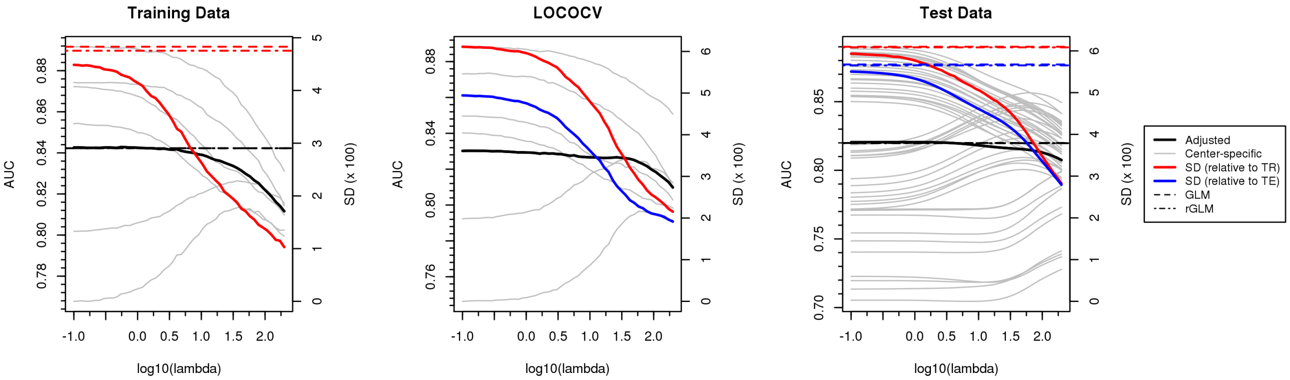

Figure 8: Penalized estimation example 10. In this example, there is a clear benefit to penalization.

Web Figure 9

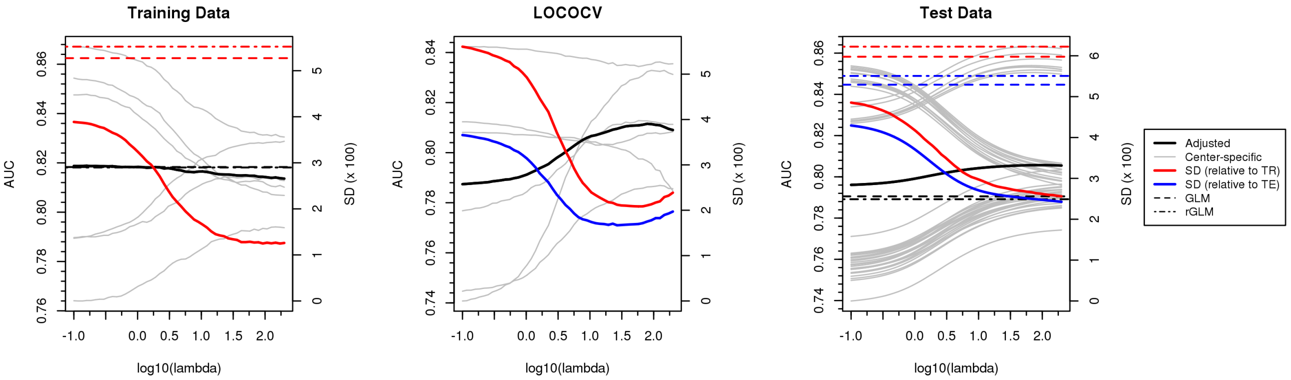

Figure 9: Penalized estimation example 11. Here, there is a clear benefit to penalization, and the overall performance even increases slightly as λ increases.

Web Figure 10

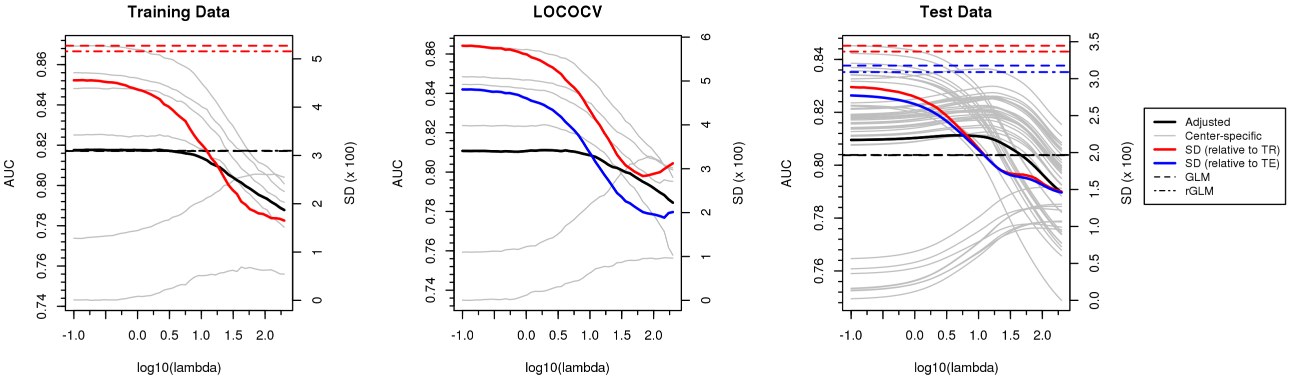

Figure 10: Penalized estimation example 12. In this example, there is a definite benefit to penalization for a range of λ values.

Web Figure 11

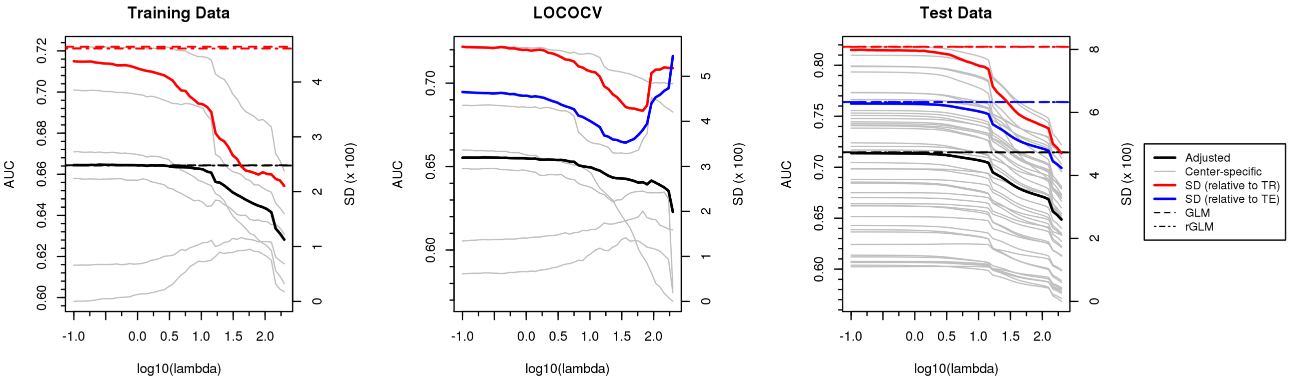

Figure 11: Penalized estimation example 13. In this example, the LOCOCV procedure returns somewhat inconclusive results, making the choice of λ less clear.

Web Figure 12

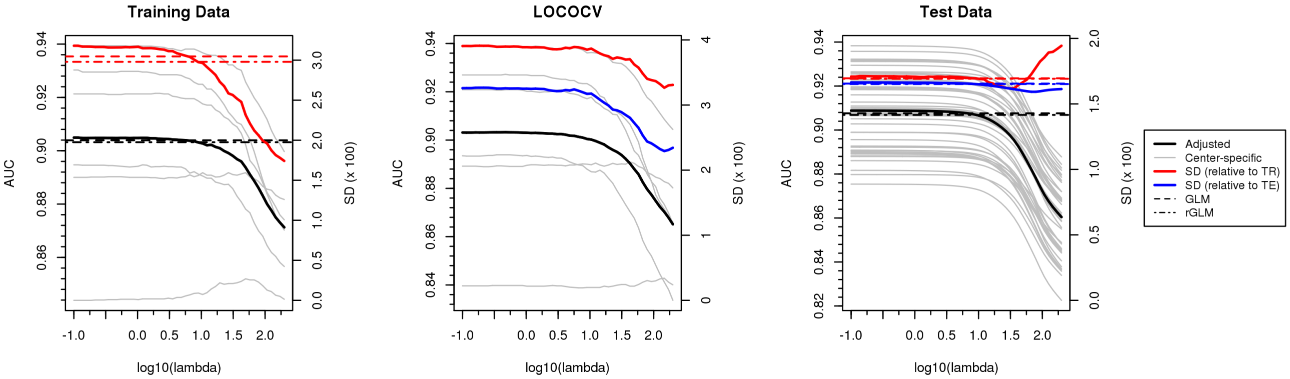

Figure 12: Penalized estimation example 14. In this example, the LOCOCV procedure is a bit misleading, when compared to the results in the test data.

Web Figure 13

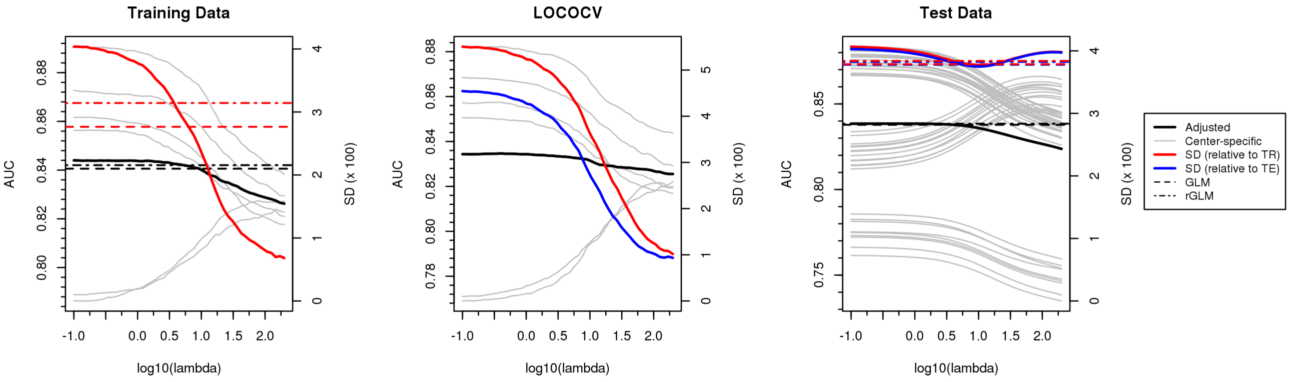

Figure 13: Penalized estimation example 15. In this example, the LOCOCV procedure is a bit misleading, when compared to the results in test data.

Web Figure 14

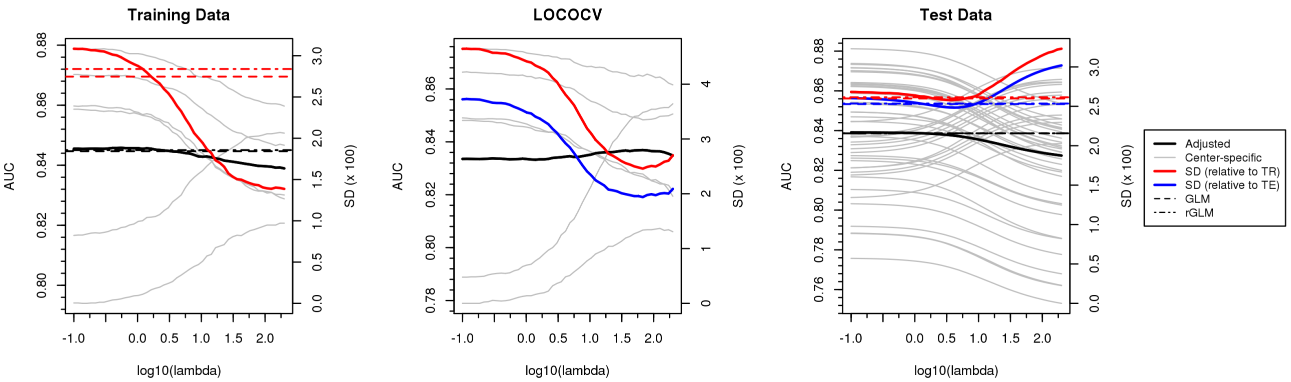

Figure 14: Penalized estimation example 16. In this example, the variability in performance in test data increases slightly with increasing λ, though this is not reflected in the LOCOCV results.

Web Figure 15

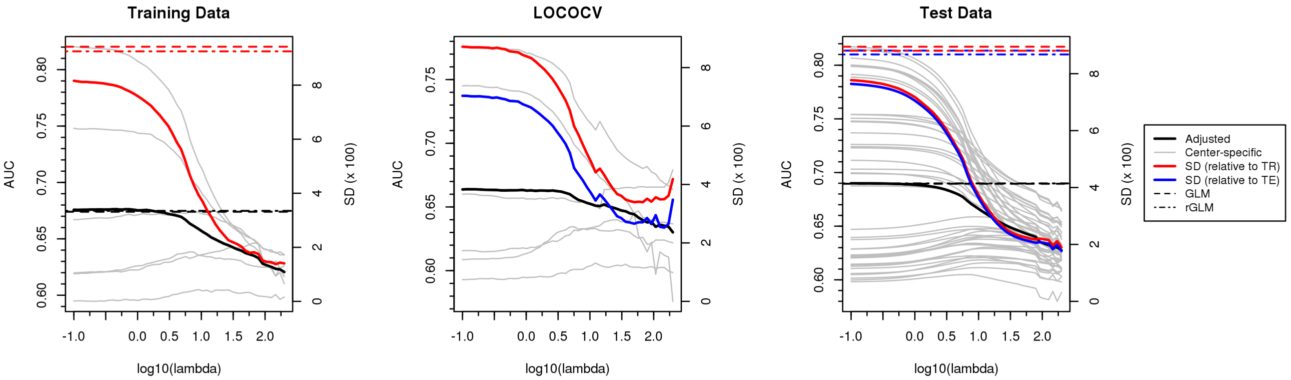

Figure 15: Penalized estimation example 17. In this example, the overall performance in test data decreases more rapidly with increasing λ than is suggested by the LOCOCV results.

Web Figure 16

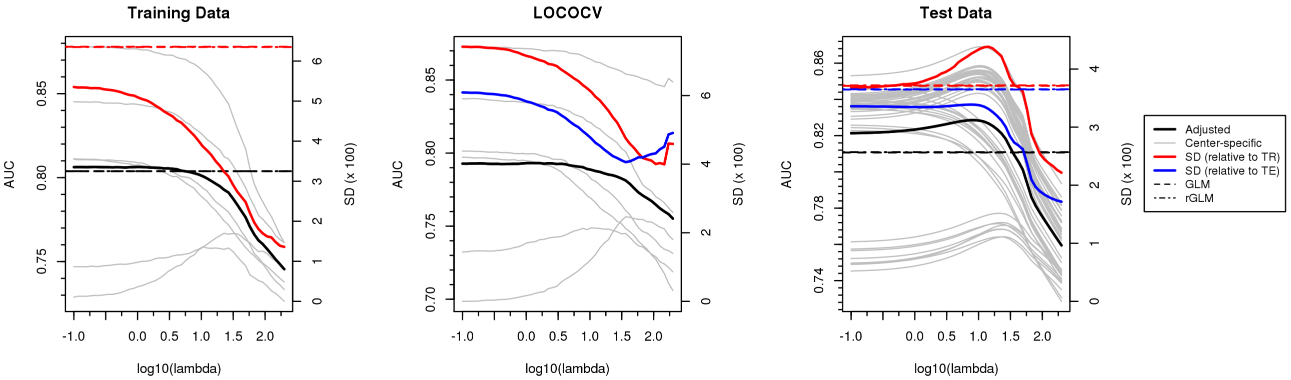

Figure 16: Penalized estimation example 18. In this example, the relationship between λ and variability in performance in the test data is somewhat unusual and is not seen in the LOCOCV results.

References

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

Meisner, A., Parikh, C. R. and Kerr, K. F. (2017). Biomarker combinations for diagnosis and prognosis in multicenter studies: principles and methods. *UW Biostatistics Working Paper Series,* Working Paper 419.