Combining Biomarkers for Diagnosis, Prognosis, and Screening: Methods for Direct Maximization, Multilevel Outcomes, and Multicenter Studies

Allison Meisner

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Kathleen F. Kerr, Chair

Marco Carone

Holly E. Janes

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Combining Biomarkers for Diagnosis, Prognosis, and Screening: Methods for Direct Maximization, Multilevel Outcomes, and Multicenter Studies

Allison Meisner

Chair of the Supervisory Committee:
Associate Professor Kathleen F. Kerr
Biostatistics

Interest in using biomarkers for prognosis, diagnosis, and screening continues to grow in many clinical areas. However, most biomarkers have only modest predictive capacity and therefore are not clinically useful. As the cost of measuring individual biomarkers declines, investigators are increasingly interested in combining biomarkers to create tools that are clinically useful. Creating such tools involves constructing a combination of a set of biomarkers and evaluating its predictive capacity; in some settings where a large number of biomarkers are available, combination selection may be necessary. In this dissertation, we consider particular challenges that may arise in the construction, evaluation, and selection of biomarker combinations and propose methods to address these challenges. We first propose a distribution-free method to construct biomarker combinations by maximizing the true positive rate while constraining the false positive rate at some clinically acceptable level. We also consider the potential role of multilevel outcomes in combination construction and selection when there is interest in predicting a particular level of the outcome due to its clinical importance. Finally, we address issues related to the use of biomarker data from multiple centers. We describe the potential role of center in these studies, demonstrate problems with currently used methods for constructing biomarker combinations, present appropriate likelihood-based methods for constructing combinations, and consider how to correctly eva-

luate the performance of combinations in this setting. We then move beyond the maximum likelihood framework and propose a method that directly maximizes a center-adjusted measure of performance while allowing for penalization of variability in performance across centers. This research provides investigators with novel insights and methods that will facilitate the development of biomarker combinations for diagnosis, prognosis, and screening.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

It is difficult to adequately convey my gratitude for my committee chair, Katie Kerr. During my time at UW, Katie has been not just a dissertation advisor, but also an RA supervisor, grant co-sponsor, teaching mentor, sounding board, sanity check, etc., etc. Thank you for nearly five years of advice, encouragement, and enthusiasm. I can't imagine a better graduate student experience. I am truly grateful for my committee: Marco Carone, Jim Hughes, Holly Janes, Yvonne Lin, and Margaret Pepe. Thank you for your support, engagement, and insights.

I am deeply indebted to Chirag Parikh, and the rest of the TRIBE-AKI study team, for providing a unique opportunity to become involved with such interesting and important clinical research. You have set an impossible standard for all future collaborations. I have been supported by an F31 grant from the National Institutes of Diabetes and Digestive and Kidney Diseases, for which I am grateful. I must note that I would not have succeeded in receiving this grant had it not been for the boundless support of my co-sponsors, Katie Kerr and Chirag Parikh.

Any success I may have is due to the support of my family. To say that my parents have been supportive is a vast understatement. They have encouraged every academic endeavor, from building dioramas in grade school to moving to Seattle for graduate school. Thank you for always believing in me. I am very grateful to my sister, Julianne, for her support, and feel so fortunate to have a sister who is also a best friend, whose own academic success motivates me, and whose sense of humor has provided a great deal of laughter and perspective during this time. I am also indebted to my close friend, Lauren Kunz, who has been in the trenches with me for nearly a decade, and whose understanding (and occasionally sympathy) has been

invaluable.

Of course, none of this would have been possible, let alone have happened, without the unwavering love and support of my husband, Jonathan. I will forever be indebted to him, but hope to at least express my appreciation. Last, but not least, I would like to convey my gratitude for my four-legged support team, our dog, Honey. She has brought endless joy and a measure of balance to my life, which have helped me more than she will ever know.

DEDICATION

to Jonathan

"That's what good waffles do, they stick together." – The Simpsons

Chapter 1

OVERVIEW

There is great interest in using biomarkers for diagnosis, prognosis, and screening in many clinical areas. Examples include using biomarkers to diagnose various types of cancer (Bünger et al., 2011), to predict risk of a future cardiac event (Blankenberg et al., 2010), or to allow early identification of certain chromosomal abnormalities, such as those related to Down syndrome, through prenatal screening (Pennings et al., 2009). In many applications, the capacity of a single biomarker to identify individuals who have or will experience the clinical outcome, i.e., the "predictive capacity" of a biomarker, is inadequate for clinical use. As a result, there is a great deal of interest in using biomarker combinations to achieve adequate predictive capacity, yielding clinical tools that can be used to inform patients about their likelihood of having a disease or their risk of experiencing an outcome (e.g., Gruenewald et al. (2006); Zethelius et al. (2008)). In this dissertation, we discuss current approaches and propose novel statistical methods related to the development of biomarker combinations for diagnosis, prognosis, and screening.

A related area of research is the use of biomarkers to predict therapeutic response, often called "predictive biomarkers." We do not consider such biomarkers here, and the use of the term "predictive" herein relates only to the performance of biomarkers in the context of diagnosis, prognosis, and screening.

There are several steps that must be taken before a biomarker combination can be adopted clinically. The combination must be constructed. In other words, a rule or procedure for combining multiple biomarker measurements into one result must be developed. This is often done by maximizing a logistic likelihood. Having been constructed, the biomarker

combination must be evaluated; that is, its predictive capacity must be assessed. We focus on the ability of a biomarker combination to discriminate between individuals who have or will experience some clinical outcome and those who do not have or will not experience the outcome. If, in addition, a large number of biomarkers is available, combination selection may be necessary. We consider methods related to these three components of biomarker combination development.

In Chapter 2, we propose a method to construct biomarker combinations by maximizing the true positive rate while constraining the false positive rate at some clinically acceptable level. The true and false positive rates are clinically useful measures of biomarker performance, and are related to the idea of discrimination described above. The acceptable false positive rate is generally low, but will vary with the setting (i.e., diagnosis, prognosis, or screening) and clinical context. Our proposal utilizes smooth approximations to the empirical true and false positive rates to provide computational feasibility. By targeting measures of predictive capacity in the construction of biomarker combinations, the resulting combination may demonstrate improved performance.

Most often, diagnosis, prognosis, and screening are considered in the context of a binary outcome. However, clinical outcomes can be multilevel, but in many settings, there is interest in predicting a particular level of the outcome. In this situation, it is common practice to dichotomize the outcome and proceed as though it were truly binary. In Chapter 3, we consider leveraging the additional information present in multilevel outcomes to yield better combinations. In particular, we evaluate whether regression methods for multilevel outcomes should be preferred to the commonly used binary logistic regression in the construction of biomarker combinations. Additionally, we propose an algorithm for selecting a biomarker combination based on the ability of the candidate combinations to discriminate between multiple levels of the outcome, as opposed to relying solely on their ability to narrowly predict the targeted level of the outcome.

In Chapters 4 and 5, we consider some issues related to using multicenter data to construct and evaluate biomarker combinations. In Chapter 4, we discuss the role that center can play

in multicenter biomarker studies and we consider likelihood-based approaches to constructing biomarker combinations in this setting. Existing approaches include ignoring center and using a logistic regression model to construct the combination or accounting for center via random intercept logistic regression. We establish the shortcomings of these approaches and propose instead using fixed intercept logistic regression. Furthermore, we illustrate the problems that can arise when center is ignored in the evaluation of biomarker combinations and recommend instead using center-adjusted measures of predictive capacity to evaluate biomarker combinations in the presence of multicenter data.

In Chapter 5, we move beyond the maximum likelihood framework and propose methods to construct biomarker combinations by maximizing a center-adjusted measure of performance. Similar to Chapter 2, we accomplish this by using a smooth approximation to this center-adjusted measure of performance, maintaining computational feasibility even when the number of biomarkers is large. While this method was developed for the multicenter setting, it can be applied to any situation where there is a discrete covariate and there is interest in optimizing covariate-adjusted performance. This method also allows for penalization of the variability in center-specific performance, which provides some assurance that when we apply the combination we have constructed to a new center, its performance will be closer to the overall performance previously observed. Again, this method is generally applicable to situations where there is a discrete nuisance covariate.

This dissertation research was motivated in large part by challenges encountered in the analysis of data from the Translational Research Investigating Biomarker Endpoints in Acute Kidney Injury (TRIBE-AKI) study, a study of acute kidney injury (AKI) following cardiac surgery (Parikh et al., 2011). The study involves more than 1200 adults at six medical centers in North America. The aims of the study include using biomarkers measured in blood and urine to provide an earlier diagnosis of AKI, which is thought to occur during the surgery but is currently not diagnosed until several days after the surgery (Parikh et al., 2011). While the outcome of severe AKI is generally of greatest interest due to the associated morbidity and mortality (Coca et al., 2012), patients can also experience mild AKI (Parikh et al., 2011),

giving an outcome with three levels. We use data from the TRIBE-AKI study to illustrate methods in Chapters 3-5.

We close in Chapter 6 by summarizing the work and discussing future research directions, including extensions of the methods we have proposed and ways in which these methods could be used to address related problems.

Chapter 2

COMBINING BIOMARKERS BY MAXIMIZING THE TRUE POSITIVE RATE FOR A FIXED FALSE POSITIVE RATE

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.

2.1 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 (Hsu and Hsueh, 2013; Wang and Chang, 2011).

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 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.

2.2 Background

2.2.1 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 $P(Z > \delta \mid D = 1)$ and the false positive rate as $P(Z > \delta \mid D = 0)$ (Pepe, 2003). For a given δ, the true positive rate is also referred to as the sensitivity, and 1-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 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 and Hsueh, 2013). This consideration led to the development of the partial AUC, the area under the ROC curve over some range of false positive rate values, (t_0, t_1) (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, which may have more clinical relevance than the AUC or the partial AUC. Additionally, in contrast 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.

2.2.2 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 and Liu (1993) demonstrated that when the vector of p biomarkers $\mathbf{X}$ 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 and 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 $\text{logit}\{P(D = 1|\mathbf{X})\} = \boldsymbol{\theta}^\top \mathbf{X}$ holds for some p-dimensional $\boldsymbol{\theta}$ (McIntosh and Pepe, 2002). If the covariance matrices are proportional but not equal, the likelihood ratio is a nonlinear function of the biomarkers, as shown in Appendix A.1.1 for the setting where $p = 2$, and the optimal biomarker combination with respect to the ROC curve will be nonlinear (McIntosh and 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 (Anderson and Bahadur, 1962; Su and Liu, 1993). Thus, methods to optimize the AUC have been proposed. When the biomarkers are conditionally multivariate normal with nonproportional covariance matrices, Su and 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 (Fong et al., 2016; Lin et al., 2011; Ma and Huang, 2007).

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 (Hsu and Hsueh, 2013; Yu and Park, 2015) while others make no such assumptions (Komori and Eguchi, 2010; Wang and Chang, 2011). 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 (Komori and Eguchi, 2010; Wang and Chang, 2011). One challenge faced by 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 and Hsueh, 2013).

In recent years, the AUC has been heavily criticized because it does not 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 and Janes, 2013). Moreover, the AUC includes, and may in fact be dominated by, regions of the ROC curve that are not clinically relevant (Pepe and 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 and 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; in other words, constructing biomarker combinations by optimizing the relevant measure of performance (Hwang et al., 2013; Liu et al., 2005; Ricamato and Tortorella, 2011; Wang and Chang, 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.

2.3 Methodology

2.3.1 Description

We will assume a non-trivial disease prevalence, $P(D = 1) \in (0, 1)$, throughout. Cases will be denoted by either $D = 1$ or the subscript D, and controls will be denoted by either $D = 0$ or the subscript $\bar{D}$.

We propose constructing a linear biomarker combination of the form $\boldsymbol{\theta}^\top \mathbf{X}$ for a p-dimensional $\mathbf{X}$ 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 $\mathbf{X}$ as a function of $\boldsymbol{\theta}$ and δ:

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

Since the true positive rate and false positive rate for a given combination $\boldsymbol{\theta}$ and threshold δ are invariant to scaling of the parameters $(\boldsymbol{\theta}, \delta)$, we must constrain $(\boldsymbol{\theta}, \delta)$ 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

$$(\boldsymbol{\theta}_{t,0}, \delta_{t,0}) = \underset{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}}{\arg\max}\, TPR(\boldsymbol{\theta}, \delta),$$

where $\Omega_{t,0} = \{\boldsymbol{\theta} \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\boldsymbol{\theta}|| = 1, FPR(\boldsymbol{\theta}, \delta) \leq t\}$. This provides the optimal combination, $\boldsymbol{\theta}_{t,0}$, and the optimal threshold, $\delta_{t,0}$.

Of course, in practice, the true and false positive rates are unknown and so $\hat{\boldsymbol{\theta}}_{t,0}$ and $\hat{\delta}_{t,0}$ cannot be computed. We can replace these unknowns by their empirical estimates,

$$\hat{TPR}_{n_D}(\boldsymbol{\theta}, \delta) = \frac{1}{n_D}\sum_{i=1}^{n_D} 1(\boldsymbol{\theta}^\top \mathbf{X}_{Di} > \delta), \quad \hat{FPR}_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) = \frac{1}{n_{\bar{D}}}\sum_{j=1}^{n_{\bar{D}}} 1(\boldsymbol{\theta}^\top \mathbf{X}_{\bar{D}j} > \delta),$$

where n_D is the number of cases and $n_{\bar{D}}$ is the number of controls, giving the total sample size $n = n_D + n_{\bar{D}}$. We can then define

$$(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) = \underset{(\boldsymbol{\theta},\delta)\in\Omega'_{t,n_{\bar{D}}}}{\arg\max}\ \hat{TPR}_{n_D}(\boldsymbol{\theta}, \delta)$$

where $\Omega'_{t,n_{\bar{D}}} = \{\boldsymbol{\theta} \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\boldsymbol{\theta}|| = 1, \hat{FPR}_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) \leq t\}$. It is possible to conduct a grid search over $(\boldsymbol{\theta}, \delta)$ 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 $(\boldsymbol{\theta}, \delta)$. Derivative-based methods therefore cannot be readily used. However, smooth approximations to indicator functions exist and have been used for AUC maximization (Fong et al., 2016; Lin et al., 2011; Ma and Huang, 2007). 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 $h \to 0$ as $n \to \infty$ (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

$$T\tilde{P}R_{n_D}(\boldsymbol{\theta},\delta) = \frac{1}{n_D}\sum_{i=1}^{n_D}\Phi\left(\frac{\boldsymbol{\theta}^\top \mathbf{X}_{Di} - \delta}{h}\right), \quad F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta) = \frac{1}{n_{\bar{D}}}\sum_{j=1}^{n_{\bar{D}}}\Phi\left(\frac{\boldsymbol{\theta}^\top \mathbf{X}_{\bar{D}j} - \delta}{h}\right).$$

Thus, we propose to compute

$$(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) = \underset{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}}{\arg\max}\ T\tilde{P}R_{n_D}(\boldsymbol{\theta},\delta), \tag{2.1}$$

where $\Omega_{t,n_{\bar{D}}} = \{\boldsymbol{\theta} \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\boldsymbol{\theta}|| = 1, F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta) \leq t\}$. We can obtain $(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t)$ by applying gradient-based methods that incorporate the constraints imposed by $\Omega_{t,n_{\bar{D}}}$ using, for example, Lagrange multipliers. Estimation can be accomplished with existing software, such as the `Rsolnp` package in `R`. The choice of tuning parameter h is discussed below.

2.3.2 Asymptotic Properties

We present a theorem which concludes that under certain conditions, the combination obtained by maximizing 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 in probability in large samples by the acceptable level t. In addition, its true positive rate tends in probability to the true positive rate of the combination obtained by maximizing the true positive rate while constraining the false positive rate.

Before stating the theorem, we give the following conditions. Some of these conditions may be affected by the use of certain sampling schemes (e.g., cohort sampling) or the presence of discrete or collinear biomarkers. Let $\mathbf{X}_{Di}$ denote the vector of biomarkers for the i^{th} case, and let $\mathbf{X}_{\bar{D}j}$ denote the vector of biomarkers for the j^{th} control.

(A1) The observations are randomly sampled conditional on D, $n_D + n_{\bar{D}} \to \infty$ and

$n_D/n_{\bar{D}} \to \rho \in (0,1)$.

(A2) The observations $\mathbf{X}_{Di}$, $i = 1, ..., n_D$, are independent and identically distributed p-dimensional random vectors with distribution function F_D, and the observations $X_{\bar{D}j}$, $j = 1, ..., n_{\bar{D}}$, are independent and identically distributed p-vector random variables with distribution function $F_{\bar{D}}$.

(A3) For each $d \in \{0,1\}$, there exists no proper linear subspace S of $\mathbb{R}^p$ such that $P(\mathbf{X} \in S \mid D = d) = 1$.

(A4) $\max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta)$, $\max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} TPR(\boldsymbol{\theta},\delta)$, and $\max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta)$ exist.

(A5) For each $d \in \{0,1\}$, the distribution function of $(\boldsymbol{\theta}^\top \mathbf{X} | D = d)$ and its inverse are Lipschitz for all $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1\}$.

(A6) For every $(\boldsymbol{\theta}, \delta) \in \Omega = \{\boldsymbol{\theta} \in \mathbb{R}^p, \delta \in \mathbb{R} : ||\boldsymbol{\theta}|| = 1\}$, $TPR(\boldsymbol{\theta},\delta)$ is Lipschitz with respect to $(\boldsymbol{\theta},\delta)$.

Theorem 2.1. *Under conditions (A1)-(A6), we have that for every fixed $t \in (0,1)$,*

$$FPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \leq t + o_p(1)$$

and

$$\left| \max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - TPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \right| \to 0 \ w.p.\ 1.$$

The proof of Theorem 2.1 is given in Appendix A.1.4. The proof relies on two lemmas, which are stated and proved in Appendices A.1.2 and A.1.3. Lemma 2.1 demonstrates almost sure convergence to zero of the difference between the maximum of a function over a fixed set and the maximum of the function over a stochastic set that converges to the fixed set in an appropriate sense. Lemma 2.2 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 2.1 then demonstrates that Lemma A1 holds for the relevant function and sets, relying in part on the conclusions of Lemma 2.2. The conclusions of Lemmas 2.1 and 2.2 are then used to demonstrate the claims of Theorem 2.1.

2.3.3 Implementation Details

In order to implement these methods, certain considerations must first be addressed, including the choice of tuning parameter h and starting values $(\tilde{\boldsymbol{\theta}}, \tilde{\delta})$ 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 $\tilde{\boldsymbol{\theta}}^\top \mathbf{X}$. In simulations, we considered both $h = \tilde{\sigma} n^{-1/3}$ and $h = \tilde{\sigma} n^{-1/2}$ and found the latter to yield a slightly better approximation with no impact on the convergence of the optimization routine. Thus, we use $h = \tilde{\sigma} n^{-1/2}$. We must also identify initial values $(\tilde{\boldsymbol{\theta}}, \tilde{\delta})$ for our procedure. As done in Fong et al. (2016), for $\tilde{\boldsymbol{\theta}}$, we use normalized estimates from robust logistic regression, which is described in greater detail below. Based on this initial value $\tilde{\boldsymbol{\theta}}$, we choose $\tilde{\delta}$ such that $\tilde{FPR}_{n_{\bar{D}}}(\tilde{\boldsymbol{\theta}}, \tilde{\delta}) = t$, where `R`'s root-finding function `uniroot` can be used to find $\tilde{\delta}$.

Finally, we have also found that when $\tilde{FPR}_{n_{\bar{D}}}$ is bounded by t, the performance of the optimization routine is 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, such that

$$\tilde{FPR}_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) \leq t + \alpha.$$

Since the effective sample size for the smooth approximation to the empirical false positive rate is $n_{\bar{D}}$, we chose to scale α with $n_{\bar{D}}$, and have found $\alpha = 1/(2n_{\bar{D}})$ 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; this may require pre-selecting biomarkers or incorporating a penalty term to encourage selection. 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.

2.4 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 model, 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 $(D \mid X_1, X_2) \sim \text{Bernoulli}\left[f\left\{\beta_0 + 4X_1 - 3X_2 - 0.8(X_1 - X_2)^3\right\}\right]$, where $\Delta \sim \text{Bernoulli}(\pi)$, $\pi = 0.05$ when outliers are simulated and $\pi = 0$ otherwise,

$$Z_0 \sim N\left(\mathbf{0}, 0.2 \times \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right), \quad Z_1 \sim N\left(\mathbf{0}, 2 \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

and Δ, Z_0, and Z_1 are independent. 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 $f(v) = f_1(v) \equiv \text{expit}(v)$, $\beta_0 = 0$, with and without outliers is given in Figure 2.1.

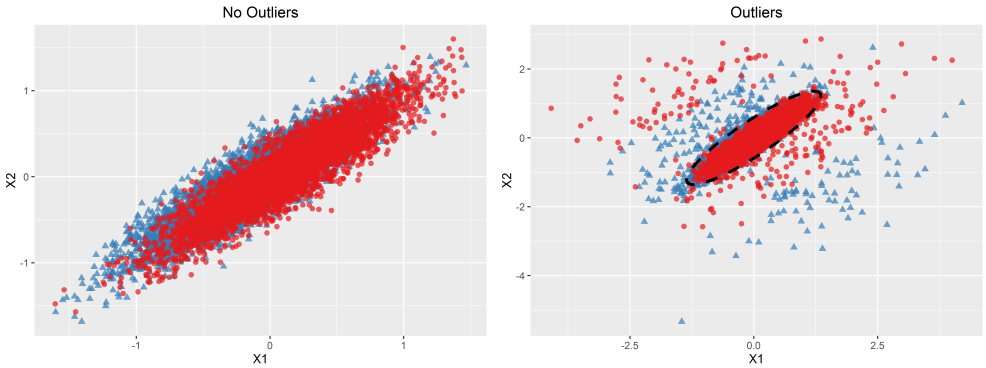

Figure 2.1: Datasets with $f(v) = f_1(v) \equiv \text{expit}(v)$, $\beta_0 = 0$, without (left plot) and with outliers (right plot). Cases are represented by red circles, and controls are represented by blue triangles. The plot with outliers also includes an ellipse (dashed black line) indicating the 99% confidence region for the distribution of (X_1, X_2) without outliers.

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 and 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 (2.1), 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 2.1 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 recalculated in the training data based on the fitted combination, as opposed to estimated directly.

Table 2.2 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

Table 2.1: 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 $\beta_0 = 0$ 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 recalculated 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 2.2: 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 $\beta_0 = 0$ 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 recalculated 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)

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 recalculated.

The results for low and high prevalence are presented in Appendix B.1. The results are generally similar to those presented in Tables 2.1 and 2.2, though there are some differences. When the prevalence is low and outliers are present, the differences between the methods

are smaller than in Tables 2.1 and 2.2. When the prevalence is low and there are no outliers, the differences in terms of the false positive rate are smaller than in Tables 2.1 and 2.2. When the prevalence is high, the differences in terms of the false positive rate are slightly larger than in Tables 2.1 and 2.2. 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 2.1 when outliers are present and are slightly larger when outliers are not present.

For some data-generating models, 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 model, 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 recalculated based on quantiles of the fitted combination.

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

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \middle| D = 1 \sim N\left(\begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}, 0.25 \times \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right),$$
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \middle| D = 0 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.25 & 0 \\ 0 & 1.25 \end{pmatrix} \right),$$

with $\mu_{X_1} = 2^{1/2}\Phi^{-1}\{AUC_{X_1}\}$ and $\mu_{X_2} = 2^{1/2}\Phi^{-1}\{AUC_{X_2}\}$, where $AUC_{X_1} = 0.6$ is the marginal AUC for X_1 and $AUC_{X_2} = 0.8$ is the marginal AUC for X_2. This data-generating model corresponds to a situation in which the biomarkers are highly correlated in cases, but essentially constitute noise in controls. Under this data-generating model, 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 recalculated 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.

2.5 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 distributions of the predictors are depicted in Figure 2.2.

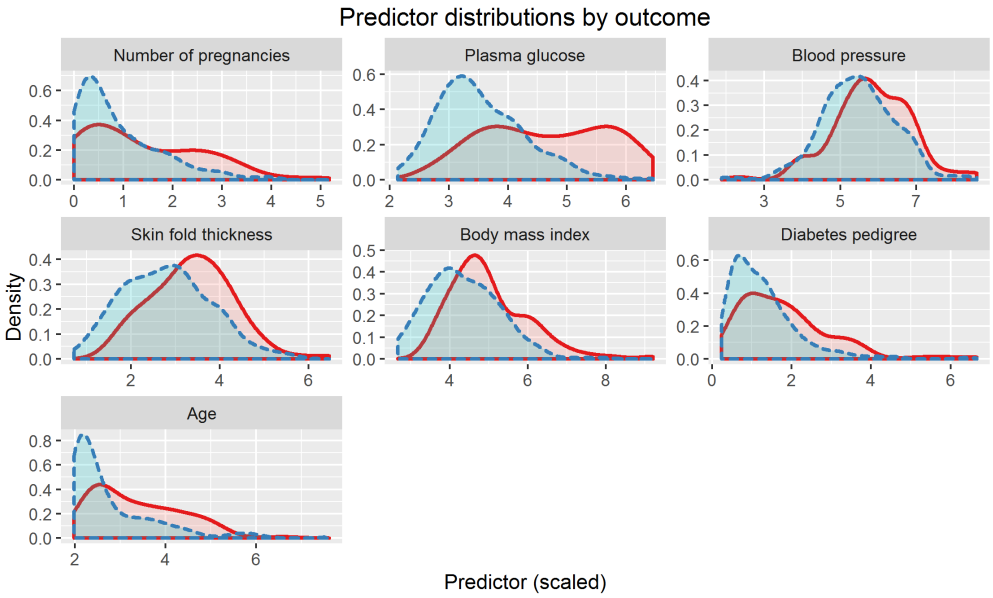

Figure 2.2: Stratified distributions of the scaled predictors measured in the diabetes study for the observations in the training data. The predictors are number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, 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.

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 2.3, where the fitted combinations from standard and robust logistic

Table 2.3: 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

regression have been normalized to aid in comparison.

Using thresholds based on $FPR = 0.10$ 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 recalculated 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 methods considered, indicating potentially important differences in the controls between the training and test data.

2.6 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 and 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. 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.

Chapter 3

USING MULTILEVEL OUTCOMES TO DEVELOP AND SELECT BIOMARKER COMBINATIONS FOR SINGLE-LEVEL PREDICTION

Abstract

Biomarker studies may involve multilevel outcomes, such as no, mild, or severe disease, yet there is often interest in diagnosing or predicting one particular level of the outcome due to its clinical significance. The standard approach to constructing biomarker combinations in this context involves dichotomizing the outcome and using standard binary regression methods. We assessed whether information can be usefully gained from instead using multilevel regression methods to construct biomarker combinations for the purpose of predicting a single level of the outcome. Furthermore, when more than a few biomarkers are available, it is often necessary to select among several candidate biomarker combinations. Generally this selection is done on the basis of the ability of each candidate combination to predict the outcome level of interest. We propose an algorithm that more fully uses the multilevel outcome to inform combination selection. We apply this algorithm to data from a study of acute kidney injury after cardiac surgery, where the kidney injury may be absent, mild or severe.

3.1 Introduction

In some clinical settings, a patient can experience one of several outcomes. For example, a patient can have no, mild, or severe disease. In the setting of cancer diagnosis, a patient can be disease-free, have a benign mass, or have a malignancy. However, there is often greatest

clinical interest in predicting one level of the outcome in particular, typically the level that poses the greatest threat in terms of morbidity and mortality. In the examples just given, this may be severe disease or the presence of a malignant tumor. Thus, investigators are interested in "single-level prediction," but a multilevel outcome is available. The question becomes whether and how the information from the multilevel outcome can be leveraged to improve prediction of the outcome level of interest.

In addition, it is becoming increasingly common for several biomarkers to be measured in each participant in a study. Often the goal of such studies is to identify a combination of biomarkers (or a subset of the available biomarkers) that can be used diagnose disease or predict some clinical outcome. In the presence of a multilevel outcome, particular challenges emerge; we consider two such challenges.

The first challenge relates to the construction of biomarker combinations, specifically, how the biomarkers should be combined when a multilevel outcome is available but there is interest in single-level prediction. The most common approach is to dichotomize the outcome and fit a single logistic regression model. Of course, this discards some information present in the multilevel outcome. We will evaluate the potential benefits of using other regression methods for constructing biomarker combinations in the presence of multilevel outcomes.

The second challenge is how biomarker combinations should be selected. In many studies, the number of biomarkers measured is quite large and a combination of only a few biomarkers is sought. Typically, investigators consider, for example, all possible pairs of biomarkers, and choose the pair with the best performance in terms of single-level prediction. As with combination construction, it may be possible to leverage the additional information in the multilevel outcome to aid in combination selection. We propose an algorithm for doing so, and provide examples of scenarios where this method is beneficial.

In addition, we illustrate the application of this combination selection method to data from the Translational Research Investigating Biomarker Endpoints in Acute Kidney Injury (TRIBE-AKI), a study of acute kidney injury after cardiac surgery (Parikh et al., 2011). This study aims to use biomarkers measured immediately after surgery to provide an earlier

diagnosis of AKI. Clinical definitions of AKI include both mild and severe types, though severe AKI is often of greatest clinical interest due to its impact on long-term morbidity and mortality (Coca et al., 2012). As a result, there is interest in using biomarkers to diagnose severe AKI.

3.2 Background

3.2.1 Constructing Combinations

We focus on "clinically ordinal" multilevel outcomes, that is, outcomes whose levels can be ordered by, for example, their clinical significance. We anticipate that such ordering may allow information about one level of the outcome to be gleaned from the others.

Binary Logistic Models

When an outcome variable D has K levels, it is common for investigators to dichotomize the outcome and fitting either one binary logistic regression model or a series of such models (Armstrong and Sloan, 1989; Bartfay et al., 1999; Maas et al., 2010; Risselada et al., 2010; Steyerberg, 2008). These include:

1. **Single regression model.** A single binary logistic model based on dichotomizing D at some clinically relevant outcome level, k': $\text{logit}\{P(D \leq k'|\mathbf{x})\} = \alpha + \boldsymbol{\beta}^T\mathbf{x}$ (Manor et al., 2000; McHugh et al., 2010; Norris et al., 2006; Roozenbeek et al., 2011; Scott et al., 1997).
2. **Each level vs. others.** K binary logistic models comparing each outcome to the combination of the other outcomes: $\text{logit}\{P(D = k|\mathbf{x})\} = \alpha_k + \boldsymbol{\beta}_k^T\mathbf{x}, \;\; k = 1, ..., K$ (Biesheuvel et al., 2008; Roukema et al., 2008).
3. **Each level vs. reference.** $(K-1)$ binary logistic models comparing each outcome to a reference level k': $\log\frac{P(D=k|\mathbf{x})}{P(D=k'|\mathbf{x})} = \alpha_k + \boldsymbol{\beta}_k^T\mathbf{x}, \;\; k \neq k'$ (Begg and Gray, 1984; Bull and Donner, 1993).

4. **Sequential models.** $(K-1)$ sequential binary logistic models where each outcome level k is compared to the combination of the levels above it: $\text{logit}\left\{P(D=1|\mathbf{x})\right\} = \alpha_1 + \boldsymbol{\beta}_1^T\mathbf{x}, \text{logit}\left\{P(D=2|D>1,\mathbf{x})\right\} = \alpha_2 + \boldsymbol{\beta}_2^T\mathbf{x}$, etc. (Roukema et al., 2008).

If $k'=1$ in (3), the resulting $K-1$ models together are equivalent to the baseline-category logit model (defined below), though the resulting estimates are expected to differ in terms of efficiency (Lunt, 2005). Approaches that compare only two levels at a time, such as (3), may be preferable as other models "lump" several potentially heterogeneous outcome levels together (van Calster et al., 2010).

Models for Ordinal Outcomes

Several models are available that fully model D (i.e., do not collapse different levels of the outcome together) while utilizing the ordered nature of D. Using one of these models to incorporate the ordering in the outcome could lead to greater parsimony and efficiency (Agresti, 2013). Ordinal methods do not assume equal spacing between the levels of D; they simply use the ordering of D (Harrell, 2013). We will consider the cumulative logit model, the continuation-ratio logit model, the adjacent-category logit model, and the stereotype model (Agresti, 2013). These models account for both the categorical nature of the outcome and the ordering (Risselada et al., 2010).

Cumulative Logit Model The cumulative logit model considers (Agresti, 2013)

$$\text{logit}\left\{P(D\leq k|\mathbf{x})\right\}, k=1,...,K-1.$$

Thus, for each value of k, this is equivalent to a binary logistic regression model (Agresti, 2013; Harrell, 2013). The cumulative logit model is one way to simultaneously use all $(K-1)$ cumulative logits in a single model (Agresti, 2013):

$$\text{logit}\left\{P(D\leq k|\mathbf{x})\right\} = \alpha_k + \boldsymbol{\beta}^T\mathbf{x}, k=1,...,K-1. \tag{3.1}$$

The model includes separate intercepts α_k for each level of the outcome (these intercepts are increasing in k), and uses a single parameter vector $\boldsymbol{\beta}$ to capture the relationship between the predictors and the levels of the outcome (Agresti, 2013). The resulting odds ratios,

$$\exp\left[\text{logit}\left\{P(D \leq k|\mathbf{x}_1)\right\} - \text{logit}\left\{P(D \leq k|\mathbf{x}_2)\right\}\right],$$

are often called the cumulative odds ratios (Agresti, 2013). Under model (3.1), the log cumulative odds ratio is proportional to the distance between the predictor values being compared, and the proportionality constant does not depend on k: (Agresti, 2013)

$$\text{logit}\left\{P(D \leq k|\mathbf{x}_1)\right\} - \text{logit}\left\{P(D \leq k|\mathbf{x}_2)\right\} = \boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2).$$

As a result of this proportionality (sometimes referred to as the parallel slopes assumption), the model given in (3.1) is also called the proportional odds model (Agresti, 2013). The proportional odds property can also be described as follows: the model considers each possible dichotomization of the outcome, assuming that the odds ratio for a higher versus lower value of D is the same wherever the outcome is dichotomized (Maas et al., 2010). When there is interest in estimating the risk of the highest level of D relative to the other levels, the cumulative logit may be a reasonable choice (Ananth and Kleinbaum, 1997).

The cumulative logit model can be motivated by considering a latent variable: suppose there exists a latent variable D^*, and the α_k are the cutpoints defining the value of D based on D^* (Agresti, 2013). When the distribution of the errors ϵ in the model $D^* = \boldsymbol{\beta}^T\mathbf{x} + \epsilon$ is standard logistic, the cumulative logit model follows; the only difference from (3.1) is that the linear predictor is $\alpha_k - \boldsymbol{\beta}^T\mathbf{x}$ (Agresti, 2013). The negative sign in front of $\boldsymbol{\beta}$ resulting from the latent variable model formulation gives the coefficients their usual interpretation in terms of direction: when $\beta_j > 0$ and the model is parameterized as $\alpha_k - \boldsymbol{\beta}^T\mathbf{x}$, D is larger for larger values of x_j (Agresti, 2013; Liu and Agresti, 2005). This latent variable interpretation is not necessary for the model to be useful (Armstrong and Sloan, 1989).

It is straightforward to generalize the model in (3.1) to include separate effects (i.e., a separate $\boldsymbol{\beta}$ vector, $\boldsymbol{\beta}_k$, for each value of k), which may be useful for assessing the assumption of proportional odds (Agresti, 2013; Ananth and Kleinbaum, 1997; Armstrong and Sloan, 1989; Liu and Agresti, 2005). In general, allowing separate effects should be done with care as it could lead to crossing of cumulative probability curves for some values of the predictors, violating the ordering of the cumulative probabilities (Agresti, 2013; Liu and Agresti, 2005). In addition, allowing separate effects comes at the expense of parsimony (Agresti, 2013). More flexible models, such as the partial proportional odds model, could be used, though this requires specification of the predictors for which proportionality should not be assumed (Agresti, 2013). It is possible to compare the estimates from binary models for each possible dichotomization of the outcome to those from the cumulative logit model to assess whether the assumption of proportional odds is reasonable (Steyerberg, 2008). Even when the proportional odds assumption does not hold, the model can be useful and powerful (Harrell, 2013). In particular, the odds ratio estimate obtained from the cumulative logit model can be conceptualized as a summary odds ratio over the different binary splits of the outcome, though the estimates arise through different likelihoods (Armstrong and Sloan, 1989; Maas et al., 2010; Strömberg, 1996). Under case-control sampling, the estimates provided by the cumulative logit model may be biased since they are affected by the sampling fractions of the outcome levels; such bias may affect the resulting predicted probabilities (Scott et al., 1997; Strömberg, 1996).

Adjacent-Category Logit Model The adjacent-category logit model can be written as (Agresti, 2013):

$$\text{logit}\left\{P(D = k | D = k \text{ or } D = k+1, \mathbf{x})\right\} = \alpha_k + \boldsymbol{\beta}^T\mathbf{x}.$$

The set of logits produced by the adjacent-category logit model are equivalent to those produced by the baseline-category logit model (defined below), except that the adjacent-

category logit model assumes a common $\boldsymbol{\beta}$ (Agresti, 2013). Thus, the adjacent-category logit model takes advantage of the ordinal outcome to achieve parsimony, without involving cumulative probabilities (Agresti, 2013; Liu and Agresti, 2005). The adjacent-category logit is more natural when there is interest in describing the effect of the predictor in terms of odds relating to particular response categories (Liu and Agresti, 2005).

Importantly, the issue discussed above for the cumulative logit model with separate effects $\boldsymbol{\beta}_k$, namely, that this model could lead to crossing of the cumulative probability curves, violating the ordering of the cumulative probabilities, is not a problem for the adjacent-category logit model with separate effects (since this model is equivalent to the baseline-category logit model and does not involve cumulative probabilities) (Agresti, 2013). Furthermore, the adjacent-category logit model can be used with data from case-control studies (Agresti, 2013).

Continuation-Ratio Logit Model The continuation-ratio logit model may be useful when a sequential mechanism determines the outcome, i.e., when individuals have to "pass through" one level of the outcome to get to the next (Agresti, 2013; Harrell, 2013). The model can be written as follows(Agresti, 2013):

$$\text{logit}\left\{P(D = k|D \geq k, \mathbf{x})\right\} = \alpha_k + \boldsymbol{\beta}^T\mathbf{x}, \quad k = 1, ..., K-1.$$

The continuation-ratio logit model considers conditional probabilities as opposed to cumulative probabilities (Harrell, 2013). As with the cumulative logit model, the continuation-ratio logit model restricts the regression coefficients to be the same for all k and the α_k are ordered in k (Ananth and Kleinbaum, 1997; Feldmann and Steudel, 2000; Harrell, 2013). The requirement that $\boldsymbol{\beta}$ be the same for each k can be relaxed; such relaxation gives the sequential binary approach described earlier (Armstrong and Sloan, 1989; Harrell, 2013). As with the cumulative logit model, the estimates provided by the continuation-ratio logit model may be biased under case-control sampling (Scott et al., 1997).

Each of the three models described above involve comparisons between different subsets of levels of D: the cumulative logit model compares $D \leq k$ to $D > k$, the adjacent-category logit model compares $D = k$ to $D = k + 1$, and the continuation-ratio logit model compares $D = k$ to $D > k$. In the context of estimating associations, the choice of model will depend upon the scientific question (Ananth and Kleinbaum, 1997). For example, if individuals are classified as having no, mild, or severe disease ($K = 3$), the continuation ratio logit would compare those with severe disease to those with mild disease, and those with mild disease to those with no disease. The cumulative logit model, on the other hand, would compare those with severe disease to those with no or mild disease, and those with mild or severe disease to those with no disease. In the prediction setting, the predicted probabilities based on these models have the same interpretation regardless of which model is used, and the choice of model is essentially a question of model fit.

Stereotype Model The stereotype model was proposed by Anderson (1984) as a sort of compromise between models that incorporate the ordinality of the outcome and more flexible models (i.e., the baseline-category logit model defined below). The stereotype model actually includes a "hierarchy" of models that vary in flexibility, as defined by the dimension of the model (Anderson, 1984). The dimension of the model can range from one (the model typically thought of as the stereotype model), to the maximum dimension d, which is related to the number of predictors p and outcome levels K (Anderson, 1984). A stereotype model of maximum dimension is a reparameterization of the baseline-category logit model (defined below) (Lunt, 2005). Anderson (1984) recommended choosing the dimension empirically, although the term "stereotype model" is generally reserved for the one-dimensional model and we focus on that model here. The one-dimensional model can be written (Anderson, 1984)

$$\log\{P(D = k|\mathbf{x})/P(D = K|\mathbf{x})\} = \alpha_k + \phi_k \boldsymbol{\beta}^T \mathbf{x}, k = 1, ..., K - 1. \tag{3.2}$$

Thus, in this model, the $\boldsymbol{\beta}_k$ are restricted in the sense that $\boldsymbol{\beta}_k = \phi_k \boldsymbol{\beta}$, which may be reasonable in a variety of settings (Anderson, 1984).

Identifiability constraints must be imposed on the ϕ_k; typically, these are $\phi_1 = 0, \phi_K = 1$ (Anderson, 1984). The definition of the stereotype model also typically includes the assumption that $\phi_1 < \phi_2 < ... < \phi_K$; when this holds, the one-dimensional model given in (3.2) is an ordered model (Anderson, 1984). In his examples, Anderson (1984) did not assume ordering among the ϕ; rather, he fit the one-dimensional model and evaluated whether the estimates $\hat{\phi}$ were ordered. Indeed, most statistical packages that fit stereotype models do not make this ordering assumption (e.g., R and STATA) and others have noted that this ordering does not need to be specified a priori (Armstrong and Sloan, 1989; Lunt, 2005). Thus, Anderson (1984) recommended fitting a fairly flexible model and assessing whether the data suggest ordering. In other words, the model allows users to judge whether the outcome levels are totally ordered or not, giving a data-driven analysis of ordering (Feldmann and Steudel, 2000). The stereotype model may be particularly useful for ordered outcomes that are not grouped continuous variables since this model does not involve comparisons between subsets of the levels of D, but rather compares each level of D to a reference level, and so may provide a better fit in this setting (Guisan and Harrell, 2000). In addition, the stereotype model can be used with case-control data (Scott et al., 1997).

Models for Nominal Outcomes

The baseline-category logit model is a very flexible approach that does not incorporate the ordered nature of D (Agresti, 2013). This model simultaneously describes the log odds for all $\begin{pmatrix} K \\ 2 \end{pmatrix}$ pairs of categories (Agresti, 2013; Steyerberg, 2008). The baseline-category logit model treats the outcome D as a multinomial variable with probabilities $(\pi_1(\mathbf{x}), ..., \pi_K(\mathbf{x}))$ and can be written as follows (Agresti, 2013):

$$\log\{P(D = k|\mathbf{x})/P(D = K|\mathbf{x})\} = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x}, k = 1, ..., K - 1.$$

The left-hand side can be interpreted as the logit of the conditional probability $P(D = k|D = k \text{ or } D = K|\mathbf{x})$ (Agresti, 2013). The baseline-category logit model allows the effect of the predictors to vary with the level of the outcome (Agresti, 2013). In addition, these logits can be used to derive the logit for the comparison of any two response categories (Agresti, 2013). When the baseline-category logit model is applied to an ordinal outcome, the model does not make use of the ordering of the outcome (Bender and Grouven, 1998). Essentially, the baseline-category logit model considers the categorical nature of the outcome, but ignores any ordering (Risselada et al., 2010).

Models for Continuous Outcomes

When the number of outcome levels K is large, linear regression models may be appealing (Guisan and Harrell, 2000; Risselada et al., 2010; Scott et al., 1997). These models consider the ordering of the outcome but ignore the categorized nature of the outcome (Risselada et al., 2010). In other words, these models are only strictly valid if the intervals between consecutive outcome levels are considered equivalent (Armstrong and Sloan, 1989). Other problems may arise, including predictions beyond the reasonable range (Guisan and Harrell, 2000). Furthermore, the categorical nature of the outcome could give misleading results, depending upon how the ordinal outcome is quantified (Scott et al., 1997). As a result of these issues, we do not consider this approach further.

Comparing Modeling Approaches

A good deal of research has been done to compare the models described above, often in terms of efficiency. Indeed, in many settings, the parameter estimates and/or predicted probabilities provided by several models are similar (Manor et al., 2000; Norris et al., 2006; Scott et al., 1997) and so the focus is often on the efficiency of these estimates. The efficiency of parameter estimates and predicted probabilities may influence the precision of the corresponding estimates of performance, and so may be informative when deciding which modeling approach to use.

Begg and Gray compared the efficiency of parameter estimates and predicted probabilities from the baseline-category logit model to those from binary logistic models comparing each outcome level to a reference level (Begg and Gray, 1984). These two models are parametrically equivalent if the reference level is K, and return parameter estimates that are asymptotically equivalent except for their asymptotic variance matrix (Begg and Gray, 1984). In the scenarios considered by Begg and Gray (1984), the efficiency of the separate logistic models was quite good for the parameter estimates and predicted probabilities (efficiencies above 90% in most cases), though there were some instances where the efficiency was $60-70\%$ for the predicted probabilities. The efficiency depended upon the prevalence of the baseline level, the number of outcome levels and the number of predictors (Begg and Gray, 1984). Notably, Begg and Gray's work was motivated in part by the limited computing resources of the era; modern computers and software can accommodate the baseline-category logit model with ease.

Bull and Donner (1993) followed up on the work of Begg and Gray (1984) by considering variations in the true parameter vectors for each level of the outcome. Specifically, they considered "collinear" coefficient vectors (e.g., the same predictor is useful for all levels, while another predictor is not useful for any level) and "orthogonal" coefficient vectors (e.g., one predictor is useful for some levels, while another predictor is useful for the other levels) (Bull and Donner, 1993). They derived expressions for the relative efficiency when the predictors are multivariate normal and found the relative efficiency to be much lower when the coefficient vectors were collinear (as low as 20% for coefficient estimates) compared to when they were orthogonal (above 70%) (Bull and Donner, 1993). All coefficient vectors can be considered to fall between these two extremes; thus, the results presented by Bull and Donner (1993) represent bounds on the efficiency. These results indicate that the baseline-category logit model is more efficient than separate binary logistic models due to the baseline-category logit model's ability to use information from other levels when estimating the parameters for one level (Bull and Donner, 1993). Thus, given current computing capabilities, there seems to be little reason to pursue the each level vs. reference binary logistic strategy instead of fitting

a baseline-category logit model.

Armstrong and Sloan (1989) demonstrated that in the situation where there is one binary predictor, a single binary logistic regression model can have more than 75% of the efficiency of the cumulative logit model, but only if the dichotomization used in the binary model is close to the optimal point. In general, they found that efficiency depended upon the number of outcome levels and how the outcome was dichotomized (Armstrong and Sloan, 1989). If the outcome was dichotomized such that the number of observations above and below the cutpoint was approximately equal, the efficiency of the binary logistic regression model relative to the cumulative logit model was 75-80% but if the cutpoint was not optimized, the efficiency could be as low as 30% (Armstrong and Sloan, 1989). Armstrong and Sloan (1989) note that "conventionally used" cutpoints may be far from optimal. Likewise, McHugh et al. (2010) compared the cumulative logit model to a single binary logistic regression model and found substantial efficiency gains for the cumulative logit model. These findings are supported by the work of Strömberg (1996), who advised caution when dichotomizing, and in particular warned against extreme cutpoints. These results indicate that collapsing a multilevel outcome into a single binary outcome generally lowers efficiency due to the inherent loss of information, and in some situations the effect could be severe (Maas et al., 2010; McHugh et al., 2010). Some studies have found that a single binary logistic model and a model that does not dichotomize the outcome provide similar results in individual datasets (Manor et al., 2000), but in general such dichotomization is expected to reduce efficiency, sometimes substantially.

Campbell and Donner (1989) focused on the classification efficiency of the baseline-category logit model and the stereotype model. They noted that if the outcome is ordinal, incorporating the ordinality of the outcome should be expected to improve classification performance; however, if the outcome is indeed ordinal, this would be evident by the parameter estimates provided by an unordered (i.e., nominal) model (Campbell and Donner, 1989). Thus, they evaluated whether incorporating ordinality affects classification performance (Campbell and Donner, 1989). Efficiency was measured in terms of the asymptotic

rate of excess classification errors due to estimation of the parameters, where classification was done on the basis of optimal classification boundaries between the outcome levels (Campbell and Donner, 1989). Relative efficiency was defined as the classification efficiency of the baseline-category logit model relative to the stereotype model (ratio of errors under the stereotype model to the errors under the baseline-category logit model) (Campbell and Donner, 1989). The stereotype model was found to be more efficient (relative efficiency of less than 75% for most clinically interesting scenarios), and the relative efficiency decreased as the number of levels K increased, the distance separating the response populations decreased, or the number of predictors increased (Campbell and Donner, 1989).

Armstrong and Sloan (1989) conclude that in general, if the order of categories can be specified with confidence, models making this ordering a strong assumption are preferable to more flexible models. In other words, it is reasonable to expect that when the outcome is ordinal, information is gained when this ordinality is used by the model (Scott et al., 1997). Harrell et al. (1998) note that models can exhibit lack of fit for some predictors and yet still provide quite accurate predicted probabilities. On the other hand, it has been noted that ordinal models become "increasingly unrealistic" as the number of outcome levels and/or predictors increases (Campbell and Donner, 1989; McHugh et al., 2010).

Applications in Risk Prediction

Previous research has considered the use of multilevel outcomes in the context of diagnostic and prognostic modelling. Multilevel outcomes are frequently encountered in this area, and it is common practice to dichotomize the outcome at some accepted cutpoint and fit a binary logistic regression model (Biesheuvel et al., 2008; Risselada et al., 2010; Steyerberg, 2008; van Calster et al., 2010). We will focus on the existing work in the context of fitting models for diagnosis and prognosis with a multilevel outcome, and then using these models for single-level prediction. It is important to keep in mind that several of the models described above compare different levels of the outcome. For example, the baseline-category logit model, the adjacent-category logits model, the stereotype model, and the binary logistic model

(3) compare two levels of the outcome; the cumulative logit model and the binary logistic models (1) and (2) consider all of the outcome levels; and the continuation-ratio logit and the binary logistic model (4) compare between two and K outcome levels. Thus, the parameter estimates from these models will often not be comparable (Armstrong and Sloan, 1989). However, for prediction research, the predicted probabilities are the chief consideration. In light of this, we emphasize the results of previous research in terms of differences in these predicted probabilities and related performance measures.

We focus on the area under the receiver operating characteristic (ROC) curve (AUC) as a measure of predictive capacity. The AUC assesses the ability of a model to discriminate between individuals who have or will experience the outcome of interest and those who do not have or will not experience the outcome of interest; the AUC for a model that is able to perfectly separate these groups is 1, while the AUC for a useless model is 0.5 (Pepe, 2003). In the context of a multilevel outcome where there is interest in single-level prediction, it is important to carefully define the group "without the outcome of interest"; depending upon the scientific question and clinical setting, this could be individuals with some reference level of the outcome, or it could be all the individuals without the targeted outcome level.

Although these investigations have considered the use of multilevel outcomes in prediction research, they have primarily involved individual datasets and/or have not focused on the setting where the goal is single-level prediction.

There have been several empirical comparisons of modeling strategies for multilevel outcomes in the setting of diagnostic and prognostic research using individual datasets, but there has not yet been a systematic evaluation of these approaches in the context of single-level prediction. For example, Biesheuvel et al. (2008) compared a baseline-category logit model to the sequential binary logistic strategy. They estimated the performance of the resulting models in terms of the AUC comparing each outcome level with the other levels combined (Biesheuvel et al., 2008). They found fairly similar estimates of the odds ratios and AUCs for both modeling approaches (Biesheuvel et al., 2008). In addition, they found good agreement among the predicted probabilities provided by the two modeling methods,

though the agreement was somewhat lower for the least common outcome (the AUCs for this outcome were also slightly different for the two methods) (Biesheuvel et al., 2008). Roukema et al. (2008) fit diagnostic models using the baseline-category logit model, the sequential binary logistic strategy, and the each level vs. others binary logistic strategy. Each model was evaluated by considering the AUC for each outcome level versus the combination of the other outcome levels (Roukema et al., 2008). They found similar discriminatory power for all three modeling strategies (Roukema et al., 2008). However, Roukema et al. (2008) employed variable selection procedures for all of the models, making comparisons difficult. The baseline-category logit model can accommodate the use of different variables to predict different levels of the outcome by constraining certain coefficients; this has been termed the "relaxed" baseline-category logit model (Bull and Donner, 1993). Thus, although separate binary logistic regressions may be preferred on the basis of the ability to perform variable selection for each level of the outcome, this is not necessarily a limitation of the baseline-category logit model (Begg and Gray, 1984).

Schuit et al. (2012) fit a baseline-category logit model and calculated an AUC for each outcome level relative to the reference level used in the baseline-category logit model, while Steyerberg et al. (2008) used a cumulative logit model to fit a prognostic model, and then calculated the AUCs for two clinically relevant cutpoints of the outcome. van Calster et al. (2010) used the baseline-category logit model and a series of binary logistic regression models for each level of the outcome versus each of the other levels individually. The AUC for each pair of levels was evaluated; these were not presented for the baseline-category logit model, though other measures of performance that were reported were slightly better for the separate binary models (van Calster et al., 2010). Importantly, van Calster et al. (2010) allowed for different variable selection in each of the binary logistic regression models, making the comparisons of the fitted models challenging, and the authors indicate that this added flexibility may have contributed to the improved performance of the separate binary models. Risselada et al. (2010) considered using cumulative logits instead of a single binary logistic model, and argued that inaccuracies due to mild violations of the proportional odds

assumption are expected to be less severe than would be due to arbitrary dichotomization of an ordinal outcome, since dichotomization involves a greater loss of information than restricting $\boldsymbol{\beta}$ to be the same for all levels of the outcome.

Begg and Gray (1984) noted that in the context of discrimination it is important to compare the predicted probabilities provided by different methods. Thus, their results on the relative efficiency of predicted probabilities are an important consideration in the prediction setting, and provide some insight into the expected behavior of measures of discrimination (i.e., AUC) for predicted probabilities based on models for multilevel outcomes.

3.2.2 Combination Selection

When there is interest in constructing a biomarker combination using a subset of the available biomarkers, some form of selection is required. This is motivated not only by logistical and financial concerns, but often also biologically: it is common for a few biomarkers to have strong predictive effects, while a large number have weak or modest effects. In the context of regression models, investigators often use an automated variable selection procedure, such as forward, backward, or stepwise selection (Harrell, 2013). These procedures generally use p-values to decide which biomarkers should be included in the combination, and thus are a form of model selection. When the goal is to use biomarker combinations for risk prediction, it seems appropriate to use a different criterion for model selection. For a binary outcome, one possibility is to use the AUC for model selection (e.g., (Gevaert et al., 2006)). That is, for each candidate combination, the AUC for the fitted combination is estimated, and the fitted combination with the highest AUC is chosen. For example, one strategy could be to consider each possible pair of biomarkers, fit each combination using logistic regression, and select the combination with the highest AUC.

Two challenges arise in utilizing such an approach. The first is that when the same data are used to construct a biomarker combination and estimate the AUC for that combination, the resulting AUC estimate will be optimistically biased because the same data were used to estimate and evaluate the combination; we refer to this as "resubstitution bias" (Kerr

et al., 2015). Methods such as bootstrapping can be used to estimate the optimism due to resubstitution bias and correct the apparent AUC estimate (Harrell, 2013).

An additional challenge applies to model selection in general. When a model is selected on the basis of its estimated performance, that estimated performance will be optimistically biased due to the model selection since the same data were used to evaluate and select the combination. This has been studied in the bioinformatics/machine learning literature, where cross-validated (CV) estimates of the classification error rate are often used as the basis for model selection. Although cross-validation can be used to address resubstitution bias, as a result of model selection bias, the estimated CV error rate for the model selected on the basis of its favorable CV error rate will be optimistic relative to the model's true error rate in independent data (Bernau et al., 2013; Boulesteix and Strobl, 2009; Cawley and Talbot, 2010; Chatfield, 1995; Ding et al., 2014; Jelizarow et al., 2010; Varma and Simon, 2006). Cawley and Talbot (2010) call this issue "overfitting the model selection criterion." Previous papers have found that this bias in the CV error estimates can be quite large, (Bernau et al., 2013; Cawley and Talbot, 2010; Dupuy and Simon, 2007; Jelizarow et al., 2010; Varma and Simon, 2006) and is related to sample size and the number of competing models (Cawley and Talbot, 2010; Chatfield, 1995; Ding et al., 2014; Jelizarow et al., 2010). This bias is due in part to variability in the estimate of performance, as low variability in the estimate of performance ensures that the optimum of the estimated selection criterion is an honest estimate of the optimum of the true performance of the models considered (Cawley and Talbot, 2010). As Varma and Simon (2006) note, the fact that this bias is due to variability in the estimate means that bias should be expected to occur with other, non-CV resampling procedures, though its magnitude may differ. Although this issue has not yet been fully characterized in the clinical risk prediction setting, where AUC is generally preferred over classification error, the problem is expected to persist. In general, when some form of model selection is done and the performance of the chosen model is evaluated without accounting for model selection, that is, treating the model as though it were pre-specified, optimistic bias is expected (Chatfield, 1995; Harrell, 2013; Lukacs et al., 2010; Steyerberg et al., 2003;

Ye, 1998).

3.3 Methods

Without loss of generality, we will suppose that for an outcome D with K levels, "single-level prediction" relates to predicting $D = K$.

3.3.1 Constructing Combinations

We have described several options for constructing biomarker combinations under the regression framework (i.e., estimating the linear predictor). In particular, we can dichotomize the outcome and use one of the four binary logistic model strategies, we can treat the outcome as ordered and use one of the four ordinal logistic models, or we can use the more flexible baseline-category logit model. Using a binary logistic modeling strategy requires either combining several levels of the outcome into one "control" group, or fitting several models to subsets of the data. Likewise, the ordinal models require restricting the nature of the relationship between the biomarkers and the outcome so as to achieve parsimony. The baseline-category logit model, on the other hand, imposes no such restrictions and includes all of the data in a single model; of course, this comes at a cost of having to estimate additional parameters. We will consider the impact of these modeling choices on the performance of the resulting combinations in a variety of scenarios via simulation. The key question is whether more sophisticated modeling approaches can offer improvements in performance in terms of single-level prediction over the naïve approach, that is, fitting a single binary logistic regression.

Since we are interested in predicting $D = K$, we considered $k' = K - 1$ in the context of the single logistic regression model strategy. Furthermore, for the purposes of predicting $D = K$, the single logistic regression model strategy and the each level vs. others binary logistic strategy are identical, and so the latter was not considered further. Finally, as the baseline-category logit model is a reparameterization of the each level vs. reference binary logistic models, and the former is generally more efficient than the latter, we did not

include the each level vs. reference binary logistic strategy in our investigation. Thus, we considered seven different modeling strategies: the single logistic regression model approach ("SingleLR"), the sequential binary logistic strategy ("LRSeq"), the cumulative logit model ("CumLogit"), the adjacent-category logit model ("AdjCatLogit"), the continuation-ratio logit model ("ContRatLogit"), the stereotype model ("Stereo"), and the baseline-category logit model ("BaselineCat").

We considered two broad simulation scenarios. In the first scenario, the biomarkers were simulated such that the assumption of proportional odds did not hold; in the second scenario, the data were simulated under the cumulative logit model where the assumption of proportional odds held. In both scenarios, we considered two biomarkers, X_1 and X_2. We considered outcomes with either 3 or 5 levels, that is, $K = 3$ or $K = 5$. The combinations were constructed using training data with 200, 400, 800 or 1600 observations and evaluated in test data with 10^4 observations. We simulated data such that $P(D = 1) = 0.1$ or 0.5 and $P(D = K) = 0.05$ or 0.3; when $K = 5$, $P(D = 2) = P(D = 3) = P(D = 4)$.

We used each of the modeling strategies to fit a linear combination of the predictors $\mathbf{X}$ in the training data, yielding the estimates $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. We then applied these estimates to the test data to determine $\hat{P}(D = K|\mathbf{X}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. Finally, we assessed the ability of $\hat{P}(D = K|\mathbf{X}, \hat{\boldsymbol{\beta}})$ to discriminate between $(D = K)$ and $(D < K)$ in the test data via the AUC. The AUC comparing other levels of the outcome (i.e., $D = K$ vs. $D = 1$) may be useful in certain clinical settings, but the AUC for $D = K$ vs. $D < K$ is often the most relevant measure in the context of single-level prediction.

In the simulations where the assumption of proportional odds did not hold (the first scenario mentioned above), the biomarkers had bivariate normal distributions, conditional on D. In particular, for $K = 3$, we considered

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 1 \sim N\left(\mathbf{0}, \Sigma_{i,1}\right), \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 2 \sim N\left(\boldsymbol{\mu}, \Sigma_{i,2}\right), \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 3 \sim N\left(\mathbf{2}, \Sigma_{i,3}\right),$$

where $\boldsymbol{\mu} \in \{-\mathbf{1}, \mathbf{0}, ..., \mathbf{2}, \mathbf{3}\}$ and $i = 1, 2, 3, 4$, corresponding to four different possibilities for the set of covariance matrices Σ_X. In particular, $\Sigma_X = \Sigma_i$, $i = 1, 2, 3, 4$ where

$$\Sigma_1 : \Sigma_{1,j} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ j = 1, 2, 3$$

$$\Sigma_2 : \Sigma_{2,j} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \ j = 1, 2, 3$$

$$\Sigma_3 : \Sigma_{3,1} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_{3,2} = \Sigma_{3,3} = 2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

$$\Sigma_4 : \Sigma_{4,1} = \Sigma_{4,2} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \Sigma_{4,3} = 2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

For $K = 5$, we considered

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 1 \sim N(\mathbf{0}, \Sigma_{i,1}), \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 2 \sim N(\mathbf{0.5}, \Sigma_{i,2}), \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 3 \sim N(\mathbf{1}, \Sigma_{i,3}),$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 4 \sim N(\boldsymbol{\mu}, \Sigma_{i,4}), \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \Big| D = 5 \sim N(\mathbf{2}, \Sigma_{i,5}),$$

where $\boldsymbol{\mu} \in \{-\mathbf{1}, \mathbf{0}, ..., \mathbf{2}, \mathbf{3}\}$, $i = 1, 2, 3, 4$ and

$$\Sigma_1 : \Sigma_{1,j} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ j = 1, 2, 3, 4, 5$$

$$\Sigma_2 : \Sigma_{2,j} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \ j = 1, 2, 3, 4, 5$$

$$\Sigma_3 : \Sigma_{3,1} = \Sigma_{3,2} = \Sigma_{3,3} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \Sigma_{3,4} = \Sigma_{3,5} = 2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix},$$

$$\Sigma_4 : \Sigma_{4,1} = \Sigma_{4,2} = \Sigma_{4,3} = \Sigma_{4,4} = 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \Sigma_{4,5} = 2 \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

If the biomarkers are conditionally bivariate normal with equal covariance matrices, the baseline-category logit model holds and can be used to characterize $P(D = k|\mathbf{X}), \ k = 1, ..., K$.

To evaluate data under the cumulative logit model with proportional odds (second scenario), we simulated bivariate normal biomarkers

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left(\mathbf{1}, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix} \right).$$

Then the outcome was simulated as a multinomial random variable, where the success probabilities of the K levels were determined by $\beta_{0k} + \boldsymbol{\beta}^\top \mathbf{X}$ such that the cumulative logit model held. Three sets of coefficients $\boldsymbol{\beta}$ were considered ($\boldsymbol{\beta} = (1, 2), (1, 1.5), (1, -1)$) and the values of β_{0k} were chosen such that the desired prevalences (given above) were achieved in a large dataset.

The simulations were repeated 1000 times.

3.3.2 Combination Selection

As with combination construction, it may be possible to leverage multilevel outcome data in selecting biomarker combinations. In particular, when a modest number of biomarkers (e.g., 20-30) are available, and there is interest in using a subset of these biomarkers for single-level prediction, different approaches to combination selection can be considered.

We consider a setting where several candidate combinations exist. A candidate combination is taken to be a set of biomarkers. One strategy is to dichotomize the outcome, fit each candidate biomarker combination using binary logistic regression, estimate the AUC for $D = K$ vs. $D < K$ (including correcting for optimism due to resubstitution bias), and select the combination with the highest estimated AUC. Of course, this optimal estimated AUC will be optimistic relative to the AUC for the same fitted combination in independent data due to model selection bias. We propose an alternative strategy where combination selection is done on the basis of not only the $D = K$ vs. $D < K$ AUC, but also the $D = K - 1$ vs. $D < K - 1$ AUC, the $D = K - 1$ vs. $D < K - 2$ AUC, and so on. We anticipate that in some settings, the estimated $D = K$ vs. $D < K$ AUC for the combination selected in this way will have less optimism due to model selection bias and so may be preferred. In particular, if some of the same biomarkers are associated with multiple levels of the outcome, our proposed method could offer improvements over the standard approach. Conversely, our approach may not be suitable in settings where different biomarkers are predictive of different levels of the outcome. In addition, we expect our approach to be useful when, as described above, many biomarkers have modest associations with the outcome, and the candidate combinations include subsets of these biomarkers.

More precisely, for $K = 3$, we define our algorithm as follows.

1. In the training data, fit all candidate biomarker combinations via binary logistic regression comparing $D = 3$ to $D < 3$.
2. Based on the combinations fit in (1), estimate (i) the apparent AUC for $D = 3$ vs. $D < 3$ and (ii) the apparent AUC for $D = 2$ vs. $D = 1$ in the training data.

3. Estimate the optimistic bias due to resubstitution: generate B bootstrap samples from the training data.
   (a) In each bootstrap sample, fit all candidate biomarker combinations via binary logistic regression comparing $D = 3$ to $D < 3$.
   (b) For each of the fitted combinations from (a), estimate (i) the AUC for $D = 3$ vs. $D < 3$ and (ii) the AUC for $D = 2$ vs. $D = 1$ in both the bootstrap data and the training data.
   (c) Estimate the optimism as the average difference in the AUCs in the bootstrap data and the training data.
4. Correct the apparent AUCs from (2) using the optimism estimated in (3).
5. Determine the ranks for the two optimism-corrected AUCs from (4) across all candidate biomarker combinations. The "standard" approach involves choosing the biomarker combination with the best AUC for $D = 3$ vs. $D < 3$. The "new" approach involves choosing the biomarker combination with the best sum of ranks for the two AUCs.
6. Apply the chosen combinations to test data and estimate the AUC for $D = 3$ vs. $D < 3$. Estimate the optimism as the difference between the AUC in the test data and the optimism-corrected estimate for the $D = 3$ vs. $D < 3$ AUC from the training data.

In practice, test data may not be available, so it may not be possible to complete step (6). An `R` package including code to implement this method, `multiselect`, will be publicly available. This approach could also be applied to the setting where the candidate combinations are a set of fitted combinations, removing the need for model fitting and correcting for resubstitution bias in our algorithm.

We used simulations to investigate the potential benefit of the proposed method. We considered five examples as a proof of concept; these are not intended to be exhaustive. In the first two examples, the cumulative logit model with proportional odds held, and in the other three, it did not. Throughout the simulations, there were $p = 30$ biomarkers and we considered the set of candidate combinations to be all possible pairs of these biomarkers. We

used $B = 50$ bootstrap replicates, a training set of 400 observations, and a test set of 10^4 observations. We repeated the simulations 500 times.

In Example 1, we had

$$\mathbf{X} \sim N(\mathbf{1}, 2\Sigma),$$

where $\mathbf{X}$ was a vector of dimension 30, and Σ was a 30×30 matrix where the diagonal elements were 1 and the off-diagonal elements were 0.3. The linear predictor was then $\boldsymbol{\beta}^\top \mathbf{X}$, where $\beta_1 = 1, \beta_2 = 2, \beta_3 = ... = \beta_{16} = 0.5, \beta_{17} = ... = \beta_{30} = 0.1$. The outcome was simulated under the cumulative logit model such that $P(D = 1) = 0.6$, $P(D = 2) = 0.3$, and $P(D = 3) = 0.1$ in a large dataset. Example 2 was identical to Example 1, except that $P(D = 2) = 0.335$ and $P(D = 3) = 0.065$.

In Example 3, we had $P(D = 1) = 0.6$, $P(D = 2) = 0.335$, and $P(D = 3) = 0.065$. Additionally,

$$(\mathbf{X}|D = 1) \sim N(\mathbf{0}, 2\Sigma)$$
$$(\mathbf{X}|D = 2) \sim N(\boldsymbol{\beta}_1, 2\Sigma)$$
$$(\mathbf{X}|D = 3) \sim N(\boldsymbol{\beta}_2, 2\Sigma),$$

where $\mathbf{X}$ was a vector of dimension 30, Σ was as defined above for Example 1, and

$$\beta_{1,1} = 1.5, \beta_{1,2} = 1, \beta_{1,3} = ... = \beta_{1,16} = 0.5, \beta_{1,17} = ... = \beta_{1,30} = 0.1$$

$$\beta_{2,1} = 2, \beta_{2,2} = 2, \beta_{2,3} = ... = \beta_{2,16} = 0.8, \beta_{2,17} = ... = \beta_{2,30} = 0.1.$$

Example 4 was identical to Example 3, except that

$$\beta_{1,1} = 1, \beta_{1,2} = 1, \beta_{1,3} = ... = \beta_{1,16} = 0.5, \beta_{1,17} = ... = \beta_{1,30} = 0.1$$

$$\beta_{2,1} = 2, \beta_{2,2} = 2, \beta_{2,3} = ... = \beta_{2,16} = 0.8, \beta_{2,17} = ... = \beta_{2,30} = 0.2.$$

Finally, Example 5 was identical to Example 3, except that

$$\beta_{1,1} = 1, \beta_{1,2} = 1, \beta_{1,3} = ... = \beta_{1,16} = 0.5, \beta_{1,17} = ... = \beta_{1,30} = 0$$

$$\beta_{2,1} = 2, \beta_{2,2} = 2, \beta_{2,3} = ... = \beta_{2,16} = 0.8, \beta_{2,17} = ... = \beta_{2,30} = 0.2.$$

3.4 Results

3.4.1 Constructing Combinations

First we consider the scenario where the cumulative logit model does not hold. We present the results for a training set size of 400; the results for the other sample sizes were generally similar. We focus on the results Σ_1 and Σ_3; the results for Σ_2 and Σ_4 show largely similar patterns and are presented in Appendix B.2.1.

Figures 3.1–3.4 present the results for $K = 3$. Figures 3.1 and 3.3 give the results in the scenario where the outcome $D = 3$ is rare ($P(D = 3) = 0.05$). Here we see that when $\boldsymbol{\mu} = \mathbf{2}$, the standard approach (the single binary logistic regression model) may do slightly worse than some of the ordinal approaches. However, when $\boldsymbol{\mu} = -\mathbf{1}$, $\boldsymbol{\mu} = \mathbf{0}$, or $\boldsymbol{\mu} = \mathbf{1}$, the standard approach is comparable to or better than the other approaches. When $\boldsymbol{\mu} = \mathbf{3}$, the performance of the standard method relative to the other methods depends heavily on the prevalence of $D = 1$, i.e., $P(D = 1) = 0.1$ vs. $P(D = 1) = 0.5$.

Figures 3.2 and 3.4 present the results for the scenario where the outcome $D = 3$ is common ($P(D = 3) = 0.3$). In this setting, the performance of the methods is quite similar for $\boldsymbol{\mu} = -\mathbf{1}$, $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\mu} = \mathbf{1}$. When $\boldsymbol{\mu} = \mathbf{2}$ and $P(D = 1) = 0.1$, the performance of the standard approach is very slightly worse than some of the ordinal approaches. As before, when $\boldsymbol{\mu} = 3$, the performance of the standard method relative to the other methods depends heavily on $P(D = 1)$.

Figures 3.5–3.8 present the results for $K = 5$. Figures 3.5 and 3.7 give the results for the scenario where the outcome $D = 5$ is rare ($P(D = 5) = 0.05$). Here we see that when $\boldsymbol{\mu} = \mathbf{2}$, the standard approach (the single binary logistic regression model) may do slightly worse

than some of the ordinal approaches. However, when $\boldsymbol{\mu} = -\mathbf{1}$, $\boldsymbol{\mu} = \mathbf{0}$ or $\boldsymbol{\mu} = \mathbf{1}$, the standard approach is comparable to or better than the other logistic or ordinal approaches. When $\boldsymbol{\mu} = \mathbf{3}$, the standard method does not do as well as some of the alternative approaches.

Figures 3.6 and 3.8 present the results for the scenario where the outcome $D = 5$ is common ($P(D = 5) = 0.3$). In this setting, the performance of the methods is comparable for $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\mu} = \mathbf{1}$, and $\boldsymbol{\mu} = \mathbf{2}$, though there are some small gains for the sequential binary logistic approach in some scenarios. For $\boldsymbol{\mu} = -\mathbf{1}$, the standard method is comparable to or better than the sequential binary logistic approach and some of the ordinal approaches, though some very small gains are seen for the sequential binary logistic approach when $P(D = 1) = 0.5$ and $\Sigma_X = \Sigma_3$. When $\boldsymbol{\mu} = \mathbf{3}$, the standard method is consistently outperformed by the sequential binary logistic approach, the stereotype model, and the baseline-category logit model.

Taken together, the results in Figures 3.1–3.8 indicate that when the cumulative logit model does not hold, if there is some ordering in the outcome by the biomarkers (that is, $\boldsymbol{\mu}$ is not extreme), the standard approach does reasonably well in general, but when $\boldsymbol{\mu}$ is extreme, another approach can perform better in some situations. In particular, when $\boldsymbol{\mu} = \mathbf{2}$, some of the ordinal approaches offered improvements over the standard approach, though these were typically small. When $\boldsymbol{\mu} = \mathbf{3}$, the performance of the standard approach was quite poor relative to some of the other approaches, including the sequential binary logistic approach, the stereotype model, and the baseline-category logit model. It seems likely that these extreme values of $\boldsymbol{\mu}$ (i.e., $\boldsymbol{\mu} = \mathbf{2}$ and $\boldsymbol{\mu} = \mathbf{3}$) had more of an effect on performance than values of $\boldsymbol{\mu}$ that were extreme in the other direction (that is, $\boldsymbol{\mu} = -\mathbf{1}$ or $\boldsymbol{\mu} = \mathbf{0}$) because we considered only the prediction of $D = K$, and higher values of $\boldsymbol{\mu}$ have the effect of making $D = 2$ (when $K = 3$) or $D = 3$ (when $K = 5$) more similar to $D = K$, to the detriment of the ability of the standard approach to predict $D = K$. Thus, it seems that in some settings, more sophisticated modeling can offer improvements over the standard approach, but very often, the standard approach does well.

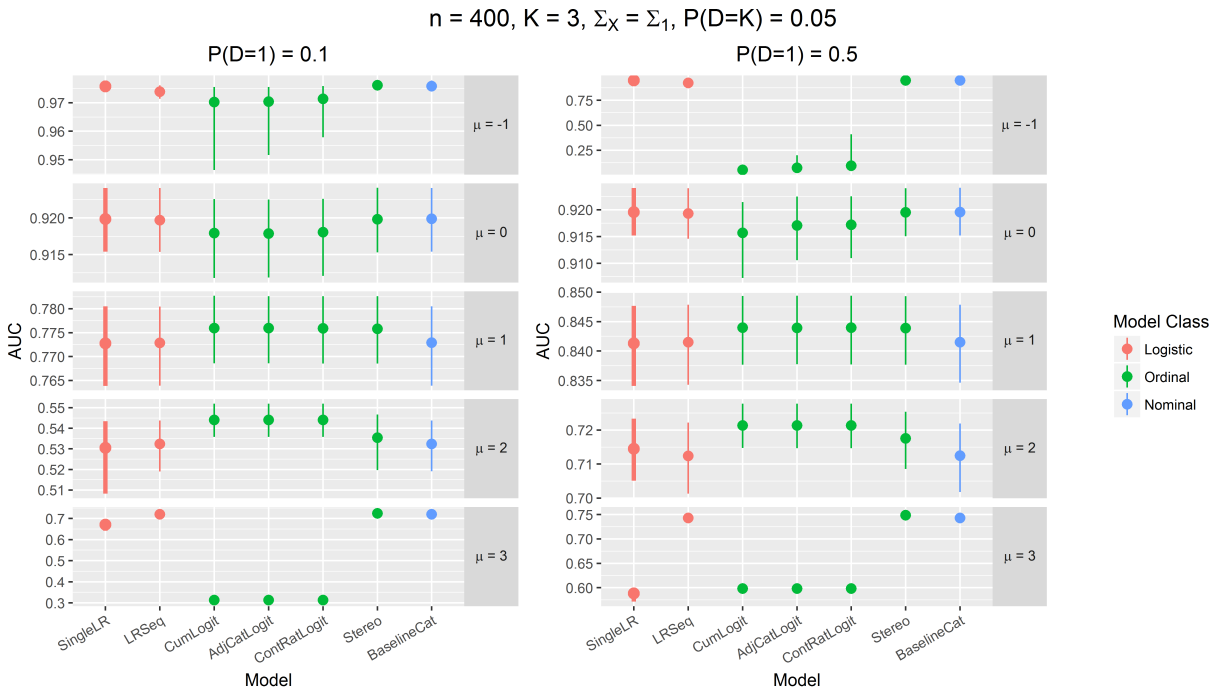

Figure 3.1: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.05$ and $\Sigma_X = \Sigma_1$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

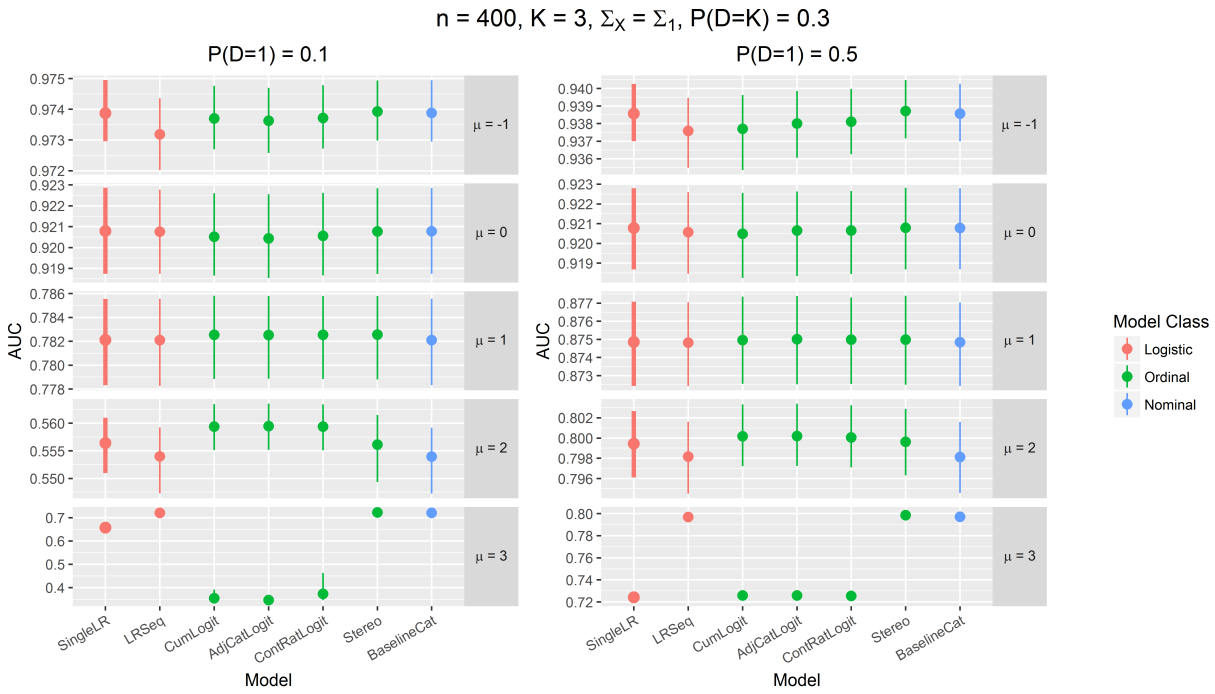

Figure 3.2: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.3$ and $\Sigma_X = \Sigma_1$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

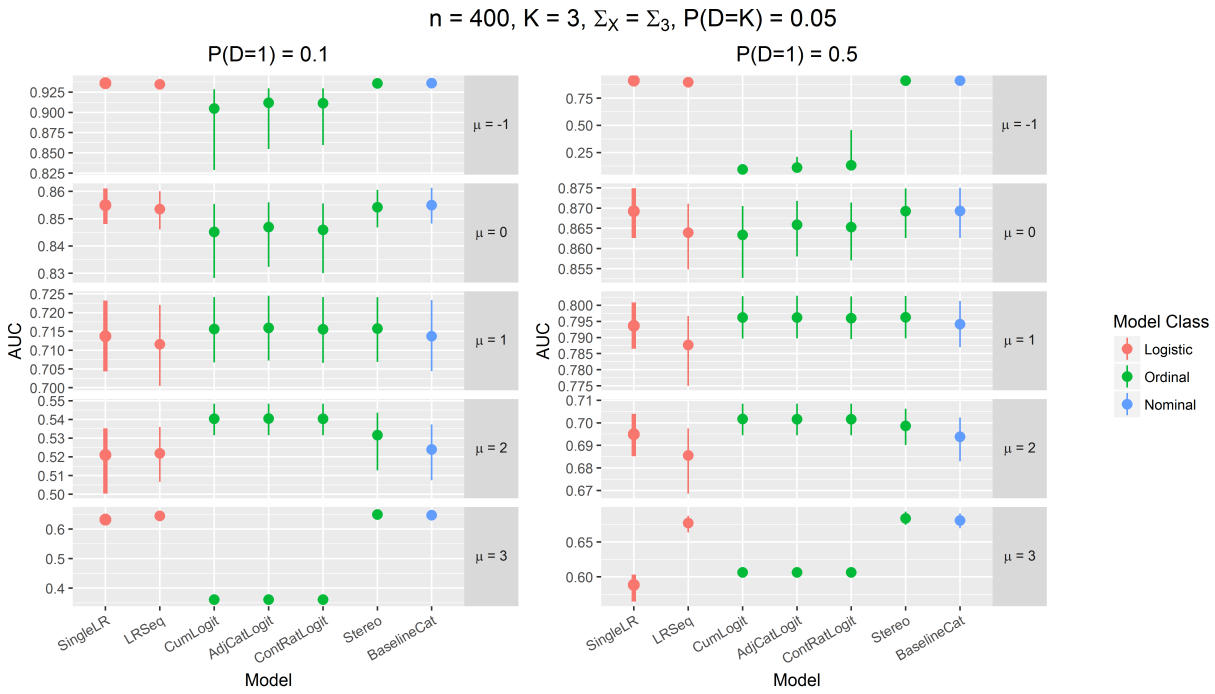

Figure 3.3: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.05$ and $\Sigma_X = \Sigma_3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

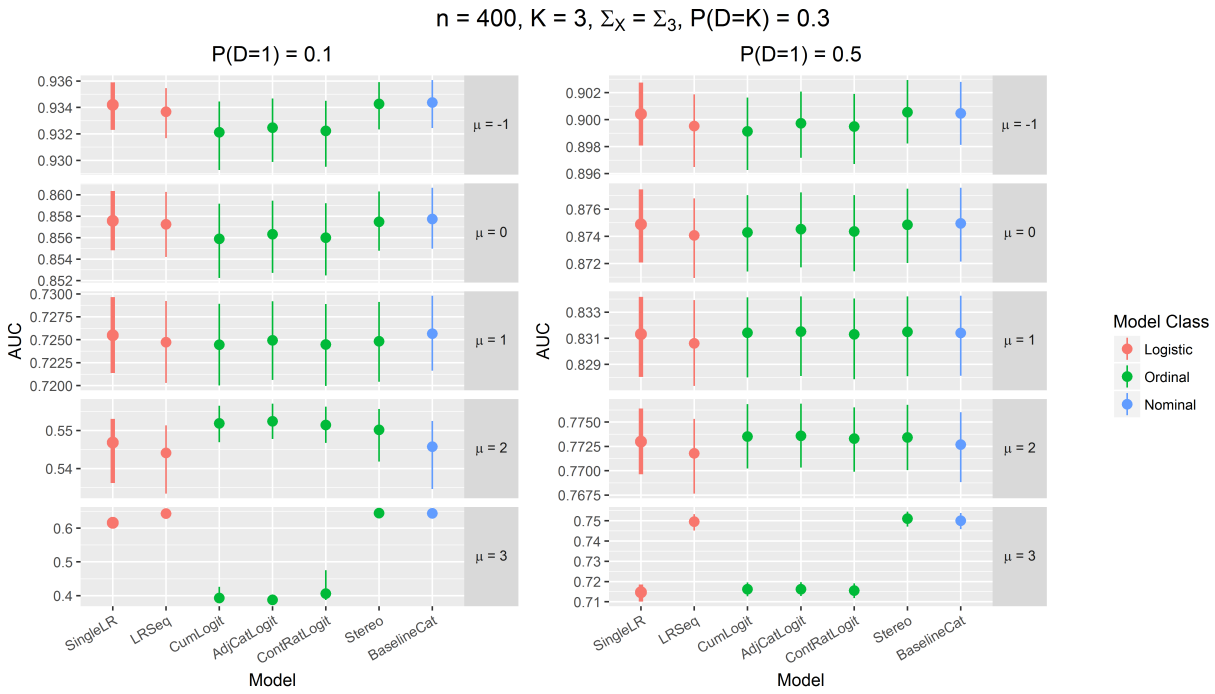

Figure 3.4: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.3$ and $\Sigma_X = \Sigma_3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

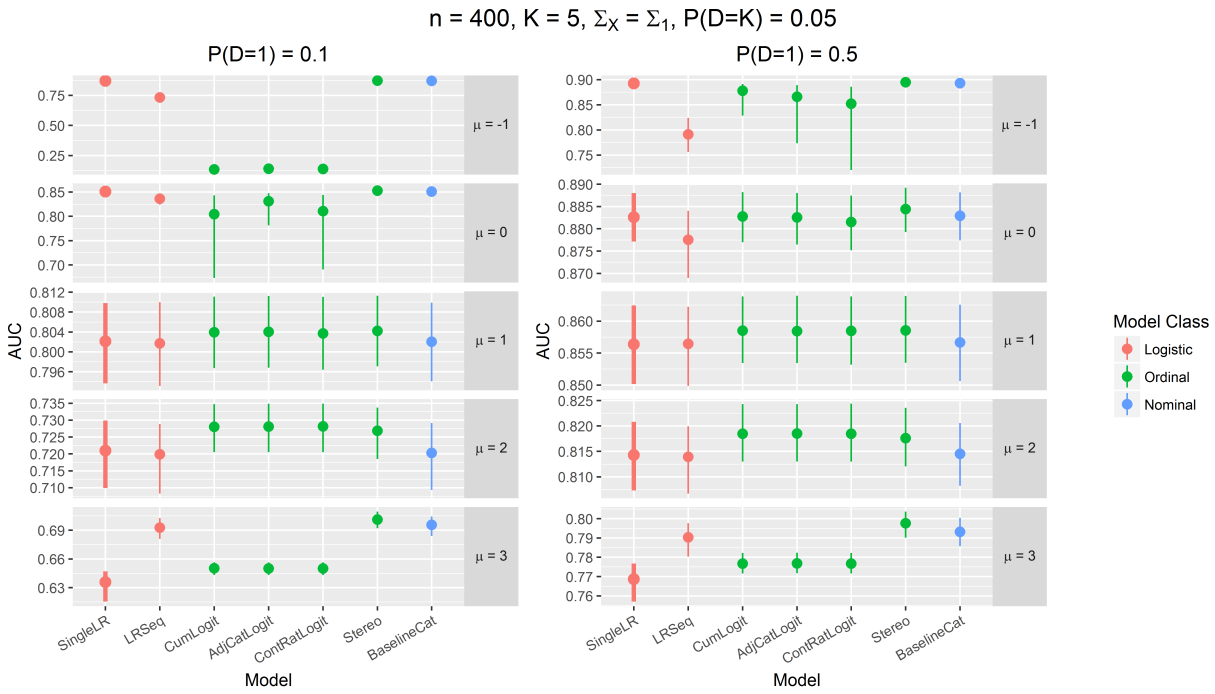

Figure 3.5: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.05$ and $\Sigma_X = \Sigma_1$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

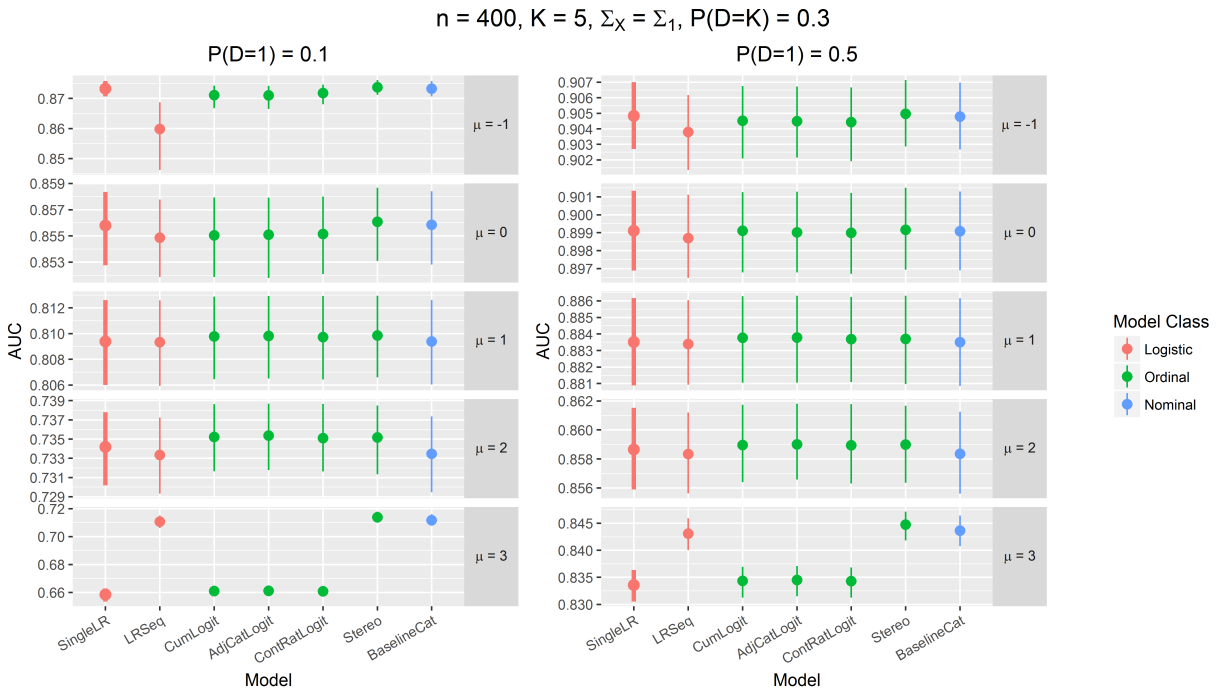

Figure 3.6: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.3$ and $\Sigma_X = \Sigma_1$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

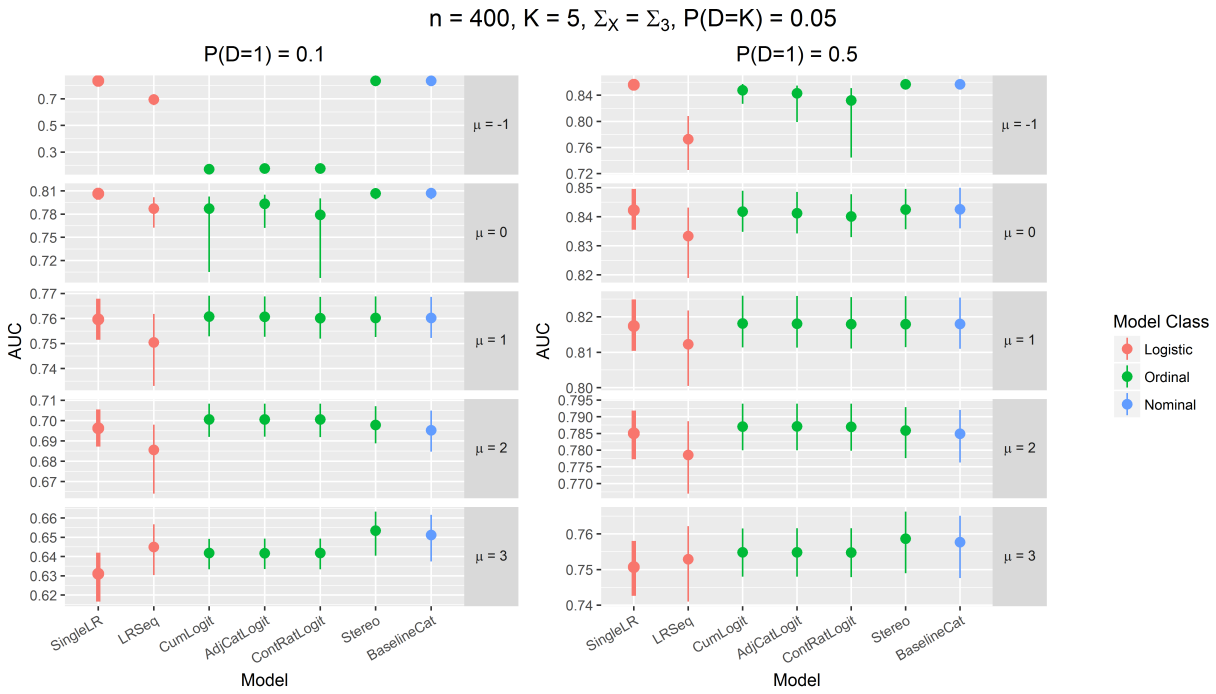

Figure 3.7: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.05$ and $\Sigma_X = \Sigma_3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

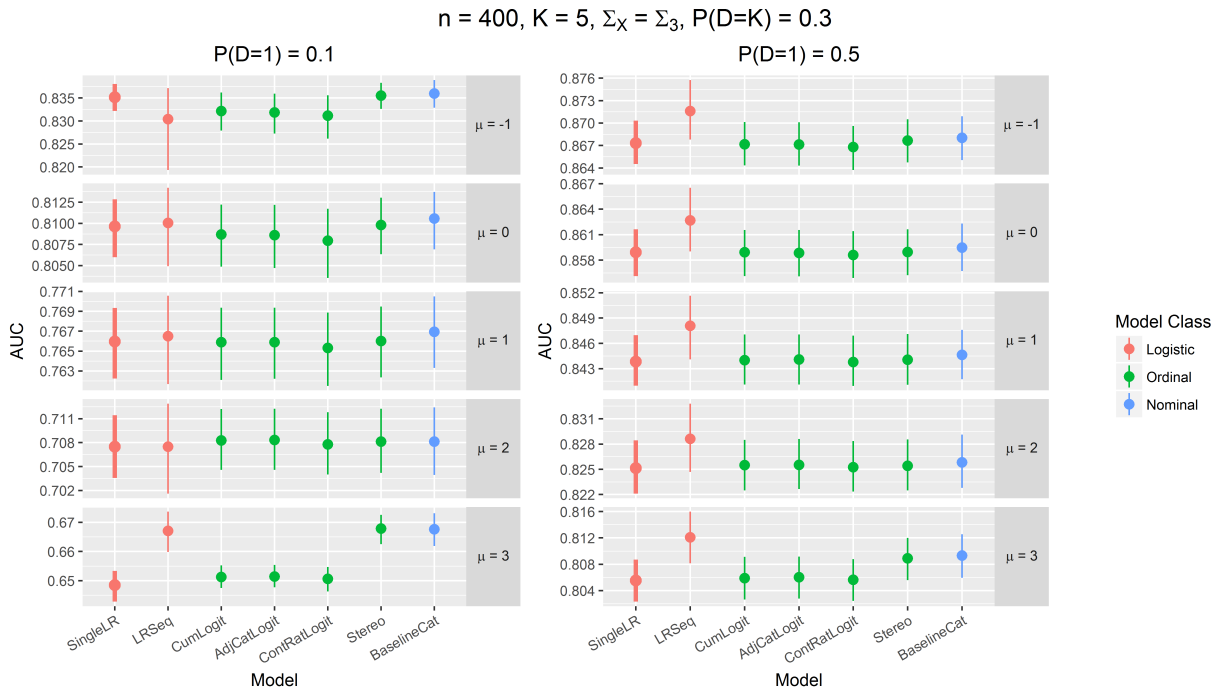

Figure 3.8: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.3$ and $\Sigma_X = \Sigma_3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

Now we consider data where the cumulative logit model with proportional odds holds. Again, we present the results for a training set size of 400 and note that the results for the other sample sizes were generally similar. The results for $P(D = K) = 0.05$ are given below, while the results for $P(D = K) = 0.3$, which showed similar performance across the approaches considered, are presented in Appendix B.2.1.

Figure 3.9 presents the results for $K = 3$. Here we see that although the ordinal modeling approaches offer some gains, as would be expected, the performance of the "standard" method is generally comparable, with differences in the median AUC of less than 0.01. Figure 3.10 presents the results for $K = 5$, and we see similar patterns. Thus, even when the

data are generated by an ordinal model, the standard approach offers similar results in terms of predictive capacity.

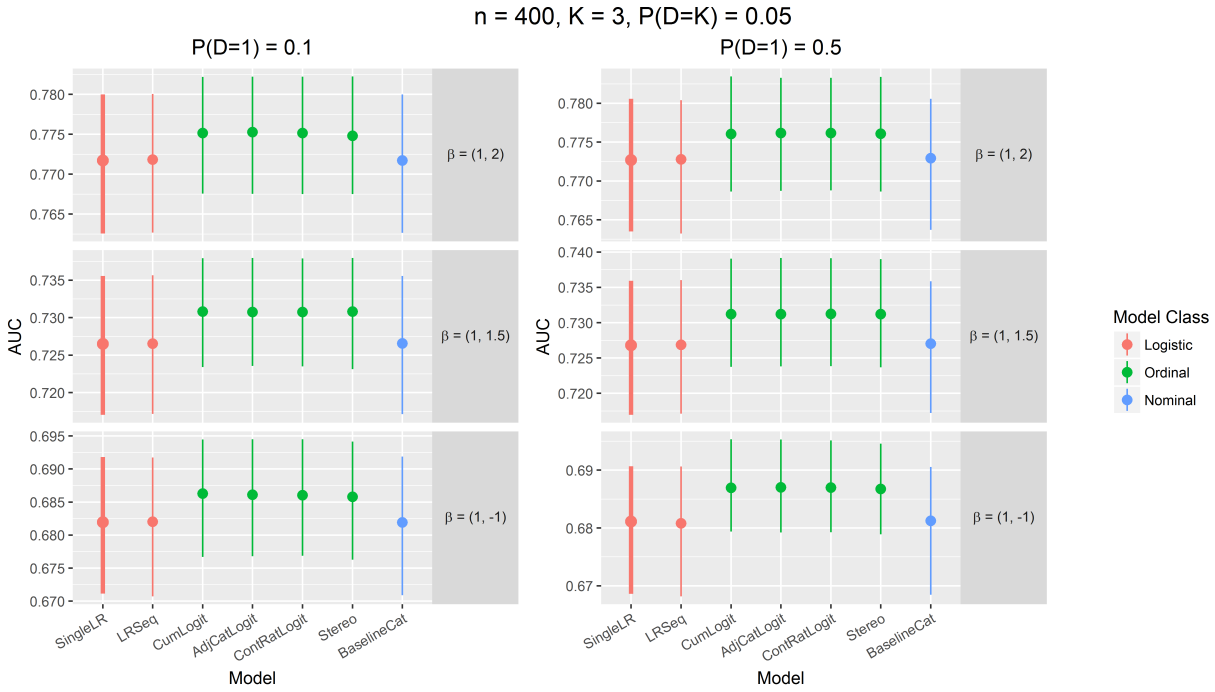

Figure 3.9: Simulation results for $K = 3$ when the cumulative logit model with proportional odds holds and $P(D = 3) = 0.05$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\beta}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

For small to moderate sample sizes, several of the approaches had issues with convergence. When the training set had 200 observations, the single binary logistic model approach failed to converge in up to 3.1% of simulations, the sequential binary logistic approach failed to converge in up to 38% of simulations, the stereotype model failed to converge in up to 2.6% of simulations, and the baseline-category logit model failed to converge in up to 1.4% of simulations. For training data with 400 observations, the sequential binary logistic approach failed to converge in up to 7% of simulations. The proportion of convergence failures was

below 0.2% for larger sample sizes.

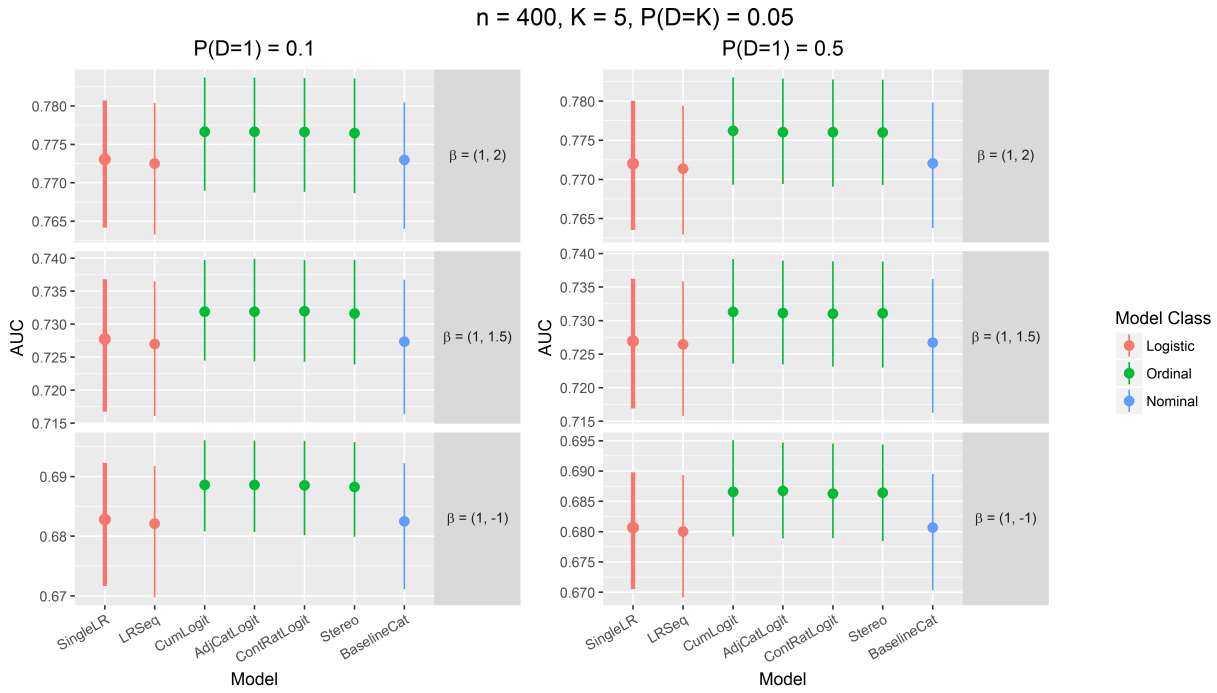

Figure 3.10: Simulation results for $K = 5$ when the cumulative logit model with proportional odds holds and $P(D = 5) = 0.05$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\beta}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

3.4.2 Combination Selection

Simulations

Figures 3.11 and 3.12 present the results for Example 1 and Example 4, respectively, for the proposed combination selection method. The results for Examples 2, 3, and 5 show similar patterns and are presented in Appendix B.2.2. The results in Figures 11 and 12 demonstrate some benefit to using the additional information available in a multilevel outcome to select

a biomarker combination for single-level prediction, both in terms of the degree of model selection bias (the optimism), and the ability of the chosen combination to discriminate $D = 3$ from $D < 3$ in test data.

There were no issues with the logistic regression model failing to converge in Example 1, eight simulations (out of 500) with convergence issues in Example 2, and one simulation with convergence issues in each of Examples 3, 4, and 5.

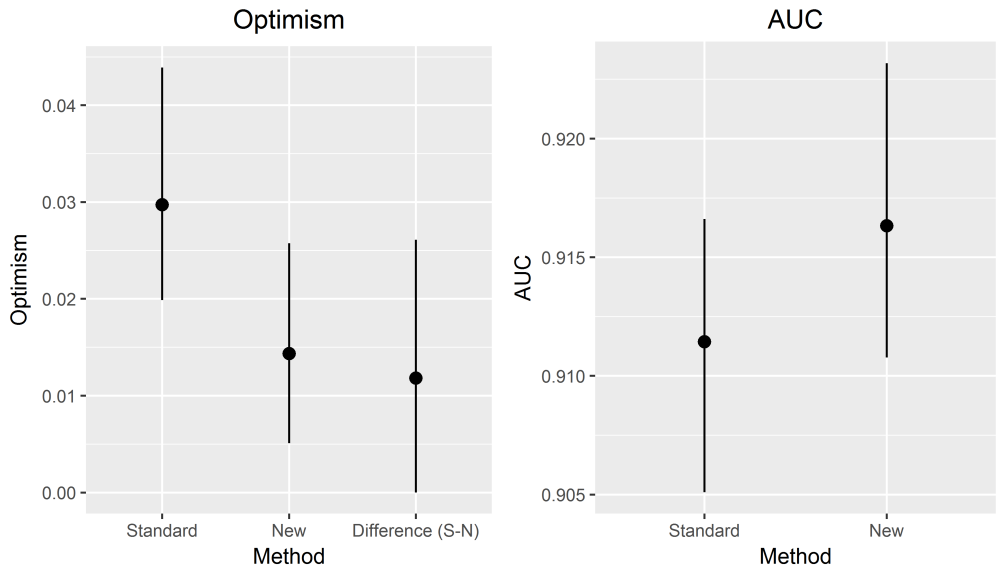

Figure 3.11: Results for the proposed combination selection method for simulation Example 1, where the cumulative logit model with proportional odds holds. The plots give the results for the standard approach, that is, choosing the combination based on the estimated AUC (corrected for optimism due to resubstitution bias) for $D = 3$ vs. $D < 3$, and the results for the new approach, that is, choosing the combination based on the AUC for $D = 3$ vs. $D < 3$ and the AUC for $D = 2$ vs. $D = 1$. The plot on the left gives the median and interquartile range for the estimated optimism due to model selection bias (the difference between the estimated AUC, corrected for optimism due to resubstitution bias, and the AUC in test data) for the selected combinations and the difference in the estimated optimism between the two approaches. The plot on the right gives the $D = 3$ vs. $D < 3$ AUC in test data for the combinations selected by the two approaches.

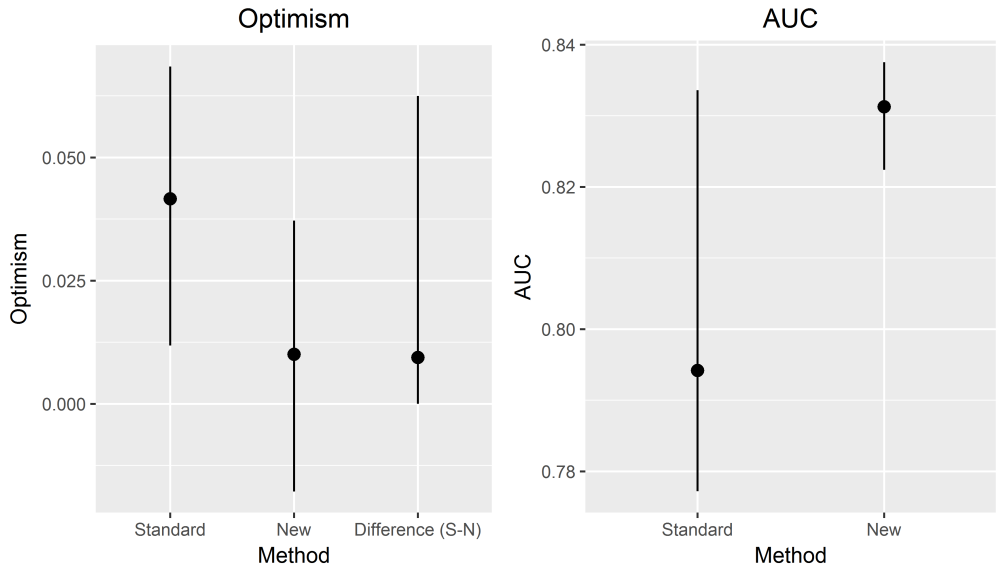

Figure 3.12: Results for the proposed combination selection method for simulation Example 4, where the cumulative logit model with proportional odds does not hold. The plots give the results for the standard approach, that is, choosing the combination based on the estimated AUC (corrected for optimism due to resubstitution bias) for $D = 3$ vs. $D < 3$, and the results for the new approach, that is, choosing the combination based on the AUC for $D = 3$ vs. $D < 3$ and the AUC for $D = 2$ vs. $D = 1$. The plot on the left gives the median and interquartile range for the estimated optimism due to model selection bias (the difference between the estimated AUC, corrected for optimism due to resubstitution bias, and the AUC in test data) for the selected combinations and the difference in the estimated optimism between the two approaches. The plot on the right gives the $D = 3$ vs. $D < 3$ AUC in test data for the combinations selected by the two approaches.

Application to TRIBE-AKI

We applied our proposed method for combination selection to data from the TRIBE-AKI study, which also served as the motivation to develop this method. As noted above, the outcome of greatest interest in the TRIBE-AKI study, acute kidney injury, is a multilevel outcome, as patients may be diagnosed with no, mild, or severe AKI. Furthermore, nearly two dozen biomarkers were measured in the study, though it is believed that only a subset are likely to be useful for early diagnosis. In particular, there was interest in considering all possible pairs of biomarkers, as it was thought that using additional biomarkers would offer

only modest gains. Thus, we applied our combination selection method to this problem.

The TRIBE-AKI study is a multicenter study, but we will restrict attention to the largest center in order to avoid issues related to center differences. We considered 14 biomarkers measured immediately after surgery, and removed observations missing any of these measurements. This left 465 observations (61 with mild AKI, 30 with severe AKI). We also log-transformed the biomarker measurements. As in the simulations, we applied our proposed method with 50 bootstrap replications.

The results for the ten best combinations in terms of the AUC for severe vs. no or mild AKI are given in Table 3.1. The combination with the highest AUC for severe vs. no/mild AKI, which would be selected by the "standard" approach, includes urine IL-18 and plasma PRO-BNP. The estimated AUCs (corrected for optimism due to resubstitution bias) for this combination were 0.8575 for severe vs. no/mild AKI and 0.6125 for mild vs. no AKI. The combination with the highest combined rank for the severe vs. no/mild AKI AUC and the mild vs. no AKI AUC, which would be selected by our proposed method, included plasma h-FABP and plasma IL6. The estimated AUCs (corrected for optimism due to resubstitution bias) for this combination were 0.8365 for severe vs. no/mild AKI and 0.6757 for mild vs. no AKI. Thus, the AUC for severe AKI for this second combination is slightly lower, but the AUC for mild AKI is substantially higher. It may be reasonable to expect that the estimated severe AKI AUC for the second combination (0.8365) is less affected by model selection bias than is the estimated severe AKI AUC for the first combination (0.8575), which may motivate choosing to validate the second combination in independent data instead of the first.

3.5 Discussion

When there is interest in using biomarker combinations for single-level prediction and a multilevel outcome is available, common practice is to dichotomize the outcome for both combination construction and selection. We have considered whether the information in a multilevel outcome variable could be more fully leveraged in the construction and selection of biomarker combinations.

Table 3.1: The ten best biomarker pairs in the TRIBE-AKI study as measured by the AUC for severe AKI vs. no/mild AKI. The AUC for severe AKI vs. no/mild AKI and the AUC for mild AKI vs. no AKI are presented. Both estimates are corrected for optimism due to resubstitution bias.

Biomarkers		AUC (Severe)	AUC (Mild)
Urine IL-18	Plasma PRO-BNP	0.8575	0.6125
Plasma h-FABP	Urine IL-18	0.8495	0.6394
Plasma h-FABP	Plasma BNP	0.8464	0.6403
Plasma h-FABP	Plasma PRO-BNP	0.8459	0.6329
Urine IL-18	Plasma BNP	0.8414	0.6168
Plasma h-FABP	Urine KIM-1	0.8410	0.6400
Plasma h-FABP	Plasma IL6	0.8365	0.6757
Plasma h-FABP	Plasma IL10	0.8342	0.6405
Plasma h-FABP	Plasma CKMB	0.8271	0.6558
Urine KIM-1	Plasma TNTHS	0.8253	0.6005

In the context of constructing biomarker combinations, we used simulations to compare seven regression-based approaches: two binary logistic regression approaches, four ordinal regression approaches, and one nominal regression approach. We considered a variety of data-generating scenarios and found that when some separation in the biomarker distributions for $D = K$ and $D < K$ exits (i.e., $\boldsymbol{\mu} < \mathbf{2}$ in our first simulation scenario) or when the cumulative logit model with proportional odds holds, the standard approach based on dichotomizing the outcome tends to work fairly well in terms of the ability of the resulting combinations to predict $D = K$. More sophisticated regression methods had the most potential to improve over the standard approach when the separation in the biomarker distributions for $D = K$ and $D < K$ was reduced.

When many candidate biomarker combinations exist and there is interest in single-level prediction, we have proposed a method that utilizes the multilevel nature of the outcome in selecting a combination, as opposed to selecting a combination based on its ability to narrowly predict the targeted level. Simulations provide evidence that the proposed method may lead to less model selection bias and potentially result in selection of combinations with

greater predictive capacity. We applied this method to data from the TRIBE-AKI study, where we demonstrated how the method could be used to select a combination from among all possible pairs of 14 biomarkers. This method is expected to be most useful when there is some ordering in the biomarkers by the levels of D. It is important to study this method further in order to fully elucidate the settings in which it could be beneficial.

In using this method for selection, it is generally informative to look at the results for the candidate combinations, as we have done in Table 3.1 for the top ten pairs in the TRIBE-AKI study. If there is a clear "winner" in terms of the AUC for $D = 3$ vs. $D < 3$, that is, if this AUC is substantially higher for one candidate combination, it is probably reasonable to select that combination, regardless of the AUC for $D = 2$ vs. $D = 1$. This is because it is unlikely that such a markedly higher AUC is due to model selection bias. On the other hand, if several combinations have fairly similar performance in terms of AUC for $D = 3$ vs. $D < 3$, it may be worth using the AUC for $D = 2$ vs. $D = 1$ to aid in selection. One possible extension of this method could involve using a weighted average of ranks for the two AUCs, rather than the sum; additionally, using a weighted average of the AUC values themselves (as opposed to their ranks) could be considered. Moving beyond the likelihood framework, it may be possible to use a multilevel outcome to simultaneously construct and select a combination by optimizing a weighted sum of multiple AUCs while penalizing complexity.

When a multilevel outcome is available and there is interest in using biomarker combinations to predict a single level of the outcome, the common approach of dichotomizing the outcome necessarily discards some information. We have described when and how this information might be usefully recovered to advance the goal of single-level prediction, thereby providing insight into how best to use the data at hand.

Chapter 4

BIOMARKER COMBINATIONS FOR RISK PREDICTION IN MULTICENTER STUDIES: PRINCIPLES AND METHODS

Abstract

Many investigators are interested in combining biomarkers to predict an outcome of interest or detect underlying disease. This endeavor is complicated by the fact that many biomarker studies involve data from multiple centers. Depending upon the relationship between center, biomarkers, and the target of prediction, care must be taken when constructing and evaluating combinations of biomarkers. We introduce a taxonomy to describe the role of center, and consider how the biomarker combination should be constructed and evaluated. We show that ignoring center, which is frequently done by clinical researchers, is often not appropriate. The limited statistical literature proposes using random intercept logistic models, which we demonstrate is often inadequate or misleading. We instead propose using fixed intercept logistic regression, which appropriately accounts for center without relying on untenable assumptions. After constructing the biomarker combination, we recommend using performance measures that account for the multicenter nature of the data, namely the center-adjusted area under the receiver operating characteristic curve. We apply these methods to data from a multicenter study of acute kidney injury after cardiac surgery. Appropriately accounting for center, both in construction and evaluation, may increase the likelihood of identifying clinically useful biomarker combinations.

4.1 Introduction

Biomedical investigations are often conducted in multiple centers (e.g., hospitals, clinics, providers). For etiologic and therapeutic studies, there is a substantial literature on the challenges of a multicenter study design. These challenges include correlations among observations from the same center and the impact of differences across centers (Localio et al., 2001). The literature on multicenter studies is especially extensive for randomized trials, where the need for careful design and analysis of multicenter studies is widely acknowledged (Localio et al., 2001).

Multicenter biomarker studies are increasingly common as investigators seek to increase power and generalizability (e.g., Degos et al. (2010); Feldstein et al. (2009); Nickolas et al. (2012)). However, in contrast to randomized trials, the literature on multicenter biomarker studies, where interest is often in using biomarkers for diagnosis and prognosis, is small. As a cause or consequence of this, the challenges and issues posed by a multicenter design appear not to be widely appreciated among biomarker researchers. Furthermore, most biomarker studies measure many biomarkers. Since biomarkers often have only modest individual performance, investigators are usually interested in constructing biomarker combinations to aid in diagnosis or prognosis. A multicenter study design can have implications for both the construction of biomarker combinations and their evaluation.

Center plays a unique role in biomarker studies, where the goal is generally classification (diagnostic setting) or prediction (prognostic setting). Center may be associated with the outcome one wants to predict, yet it cannot be used as a predictor. The reason is that center does not generalize to patients from centers not in the study, so a prediction instrument that used center as a predictor would not be broadly applicable. Recognizing this situation, it seems many biomarker investigators decide to simply ignore the fact that their data come from multiple centers. However, as we will demonstrate, ignoring center can produce misleading or undesirable results. Although center cannot be used as a predictor, it generally must be accounted for.

We will consider the role that center can play in multicenter biomarker studies, including proposing a taxonomy that distinguishes different ways that center can be important and providing guidance to researchers on identifying the role center plays in their studies. We assess the impact of ignoring center and evaluate existing approaches for accounting for center in biomarker studies. Finally, we propose methods for constructing and evaluating biomarker combinations when data come from a multicenter study. We restrict attention to biomarkers that will be used to identify individuals likely to have (in a diagnostic setting) or develop (in a prognostic setting) some clinical outcome; such biomarkers are sometimes referred to as "prognostic" or "diagnostic" biomarkers, as opposed to biomarkers used to predict response to treatment, which are often called "predictive" biomarkers.

As an example of a multicenter biomarker study, we consider the Translational Research Investigating Biomarker Endpoints in Acute Kidney Injury (TRIBE-AKI) study. The TRIBE-AKI study involves 1219 cardiac surgery patients at six centers in North America (Parikh et al., 2011). The participants were followed for diagnosis of post-operative acute kidney injury (AKI). For each patient, blood and urine were collected at multiple time points pre- and post-operatively, 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 (Parikh et al., 2011). One goal of the study is to identify combinations of post-operative biomarkers that can provide an earlier diagnosis of AKI.

4.2 Background

We first introduce several general concepts and methods which we will use throughout this paper. These include the idea of omitted variable bias, random intercept logistic regression, fixed intercept logistic regression, issues related to asymptotics and risk prediction in clustered data, conditional methods for evaluating predictive capacity, and some of the challenges associated with developing biomarker combinations for diagnosis and prognosis in general. These concepts will provide context for the rest of the paper, which will draw on the concepts

introduced below to provide novel insights and propose appropriate methods for constructing and evaluating biomarker combinations for diagnosis and prognosis using data from multiple centers.

4.2.1 Omitted Variables

The problem of omitted variable bias in logistic regression has been considered extensively in both the statistics and econometrics literatures (Begg and Lagakos, 1990; Cramer, 2005; Gail et al., 1984; Lee, 1982; Mood, 2010; Neuhaus and Jewell, 1993). In the case of linear regression, bias will not occur if the omitted variable is independent of the included predictor. However, for the logit link, if a variable orthogonal to the included predictor is omitted, bias in the estimate of the included predictor may result (Lee, 1982).

Suppose the true model can be written as

$$\text{logit}\left\{P(D=1|X,Z)\right\} = \theta_0 + \theta_X X + \boldsymbol{\theta}_Z^\top \mathbf{Z}, \tag{4.1}$$

where $\boldsymbol{\theta}_Z$ is $(k-1)$-dimensional and $\mathbf{Z}$ is a collection of $(k-1)$ dummy variables, representing a discrete variable Z with k levels. If X and Z are independent conditional on D, the estimated coefficient for X will be consistent for θ_X even if Z is omitted (Lee, 1982). This is related to the concept of collapsibility: essentially, for general (possibly multidimensional) X and discrete Z, if model (4.1) holds, θ_X is collapsible if (i) the marginal logistic regression $\text{logit}\left\{P(D=1|X)\right\} = \theta_0^* + \theta_X^* X$ holds and (ii) $\theta_X = \theta_X^*$ (Guo and Geng, 1995). Guo and Geng (1995) prove that θ_X is collapsible if (i) $Z \perp X|D$ or (ii) $Z \perp D|X$. The second condition follows directly from the fact that if $Z \perp D|X$, $\boldsymbol{\theta}_Z = \mathbf{0}$; in this case, θ_0 and θ_X are *both* correctly estimated even if Z is excluded. In general, θ_X^* may differ from θ_X and the marginal logit, $\text{logit}\left\{P(D=1|X)\right\}$, may not be linear.

The issue of omitted variables comes up frequently, often related to discussions of "unobserved heterogeneity" (Cramer, 2005; Dieleman and Templin, 2014; Gardiner et al., 2009), random intercept models (Brumback et al., 2010; Gardiner et al., 2009; Heagerty, 1999; Her-

nan et al., 2011; Hu et al., 1998; Localio et al., 2001; Neuhaus and Jewell, 1993; Neuhaus and Kalbfleisch, 1998; Neuhaus et al., 1991, 2014; Seaman et al., 2014; Ten Have et al., 1995) and exogeneity/endogeneity (Antonakis et al., 2010; Skrondal and Rabe-Hesketh, 2014).

4.2.2 Random Intercept Logistic Regression

Random effects models have been studied extensively in a number of literatures (e.g., statistics, econometrics, epidemiology and sociology). Here we will focus on random intercept logistic regression (RILR).

Random effects models are often used when there is some sort of clustering of the data. The standard notation for clustered data uses i to index clusters, $i = 1, ..., m$, and j to index observation within a cluster, $j = 1, ..., n_i$, where n_i is the number of observations in the i^{th} cluster. We will consider D_{ij} to be an indicator of a binary outcome of interest and $\mathbf{X}_{ij}$ a vector of predictors. When considering the collection of observations within cluster i, we will write $\mathbf{D}_i$ and $\mathbf{X}_i$. We will often differentiate between predictors that are constant for all observations in a cluster (often called cluster-level, cluster-constant, or between-cluster predictors) and those that vary across observations in a cluster (called cluster-varying or within-cluster predictors).

The RILR model can be written as follows:

$$\text{logit}\left\{P(D_{ij} = 1|\mathbf{X}_{ij}, b_i)\right\} = b_i + \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_{ij}, \quad b_i \overset{iid}{\sim} F, \tag{4.2}$$

where b_i, the random intercept for cluster i, has distribution F. Typically, it is assumed that $b_i \sim N(0, \sigma^2)$; if we assume $b_i \sim N(0, \sigma^2)$, σ^2 is an additional parameter in this model. If we view the random intercept b_i as σz_i, where $z_i \sim N(0, 1)$, σ is the regression coefficient for this standardized omitted (cluster-level) predictor (Heagerty and Zeger, 2000). In that sense, b_i is generally "interpreted as the combined effects of omitted cluster-level predictors" (Skrondal and Rabe-Hesketh, 2014). In the model formulation given above, $b_i + \beta_0$ can be thought of as the center-specific offset.

RILR has been used extensively by researchers and is intuitively appealing: the notion of a random intercept corresponds to the idea that the clusters under observation are a sample from some larger population of clusters. However, as noted by Localio et al. (2001), these models are "attractive but challenging," and we will explore some of these difficulties. Despite challenges, these models remain popular in epidemiology and other health-related fields.

Assumptions

The assumptions made by RILR are not always fully appreciated by practitioners, though they can have an impact on the estimates produced by the model. The key assumptions typically made by the RILR model given in equation (4.2) are (Gardiner et al., 2009; Seaman et al., 2014):

(A1) $b_i \perp \mathbf{X}_i$

(A2) $b_i \overset{iid}{\sim} N(0, \sigma^2)$

(A3) Conditional on $(\mathbf{X}_i, b_i)$, the $D_{i1}, ...D_{in_i}$ are independent

Assumptions (A1) and (A3) can be extended to include cluster size n_i (Seaman et al., 2014); however, we will not consider the role of cluster size here. Assumption (A2) is not strictly necessary, but is generally assumed (it is required, though, that $E(b_i)$ be a known constant if the fixed intercept β_0 is also included) (Kahan, 2014; Seaman et al., 2014). The model formulation given in equation (4.2) and the assumptions given above imply that any dependence of D_{ij} on the values of (cluster-varying) predictors for other cluster members must be through a cluster-level summary of these values, and the effect of this summary must be constant for all members of the cluster and captured by b_i (Seaman et al., 2014). Furthermore, the random intercept model implies a compound symmetric covariance structure for the responses $\mathbf{D}_i$ within a cluster; thus, negative correlations among outcomes in

a cluster cannot be accommodated by this model (Hu et al., 1998; Ten Have et al., 1995). However, in many cases of interest, it is reasonable to assume a positive correlation.

Assumption (A1) can be written as $f(b_i|\mathbf{X}_i) = f(b_i)$, which is a fairly strong assumption in the non-randomized setting (Heagerty and Zeger, 2000). This assumption is often implausible when the distribution of the predictors varies by cluster. Also, implicit in assumption (A2) is the assertion that σ is constant (Heagerty and Kurland, 2001; Heagerty and Zeger, 2000). Violations of these assumptions are discussed below.

Estimates

It is important to distinguish between marginal and conditional estimation approaches: the conditional (or cluster-specific) approach involves modeling the probability distribution of D as a function of predictors and cluster-specific parameters (e.g., random cluster-specific intercepts), while the marginal (or population-averaged) approach involves modeling the marginal expectation of D as a function of predictors (Neuhaus et al., 1991). RILR is an example of a conditional approach. Due to the inclusion of cluster-specific parameters, parameter interpretation under the conditional approach is with respect to cluster (Neuhaus et al., 1991). For cluster-varying predictors, conditional approaches are often more relevant than marginal approaches, such as generalized estimating equations (GEE) (Neuhaus et al., 1991).

Predictors frequently have both a between- and within-cluster component; that is, they vary both within and between clusters (Neuhaus and Kalbfleisch, 1998). Estimates obtained via conditional approaches are generally interpreted as estimates of the within-cluster association, i.e., the association within each cluster, averaged across clusters; this is typically what researchers are trying to estimate when they use these approaches with predictors that vary within clusters (Gunasekara et al., 2014; Localio et al., 2001; Ten Have et al., 1995). This is often of interest since the within-cluster association is unaffected by all (unmeasured) cluster-constant confounders (Gunasekara et al., 2014). However, as discussed below, estimated coefficients obtained from RILR may not actually represent the within-cluster

association: depending upon the nature of the data, the resulting estimates are often a combination of within- and between-cluster comparisons (Gunasekara et al., 2014; Hu et al., 1998; Localio et al., 2001; Neuhaus and Kalbfleisch, 1998; Neuhaus et al., 1991; Neuhaus and McCulloch, 2006; Seaman et al., 2014; Ten Have et al., 1996). Between-cluster effects are sometimes called "contextual effects" (Greenland, 2002; Rabe-Hesketh and Skrondal, 2010) and are likely to include the effects of cluster-constant confounders (Gunasekara et al., 2014).

Violations of Assumptions

First we will consider (A1); specifically, we will consider violations of the assumption $b_i \perp \mathbf{X}_i$. In the context of a randomized multicenter clinical trial, the assumption holds if randomization is stratified, i.e., done within each center, since in this situation the distribution of treatment is the same across centers (Localio et al., 2001; Neuhaus and Kalbfleisch, 1998). However, as noted above, it is generally the case that the distribution of a cluster-varying predictor varies across clusters; in other words, most predictors are not purely within-cluster and have both a between-cluster component and a within-cluster component (Berlin et al., 1999; McCulloch and Neuhaus, 2011; Neuhaus and Kalbfleisch, 1998). When such predictors are included in a RILR model, the assumption that $b_i \perp \mathbf{X}_i$ may not hold, leading to distortions of the association of interest (Berlin et al., 1999; Neuhaus and Kalbfleisch, 1998).

As a concrete example, suppose the following model holds for the cluster-varying predictor x:

$$\begin{aligned}\text{logit}\left\{P(D_{ij}=1|x_{ij},b_i')\right\} &= b_i' + \gamma_0 + \gamma_1 h(\mathbf{X}_i) + \gamma_2(x_{ij} - h(\mathbf{X}_i)) \\ &= b_i' + \gamma_0 + (\gamma_1 - \gamma_2)h(\mathbf{X}_i) + \gamma_2 x_{ij}, \end{aligned} \tag{4.3}$$

where $h(\cdot)$ is some cluster-level summary of $\mathbf{X}_i$ such that $x_{ij} - h(\mathbf{X}_i)$ has the same distribution across clusters and $b_i' \sim N(0, \sigma^2)$. Here, $x_{ij} - h(\mathbf{X}_i)$ represents the within-cluster component of x and $h(\mathbf{X}_i)$ represents the between-cluster component of x. If the distribution of the predictor x is the same across clusters, then $(\gamma_1 - \gamma_2)h(\mathbf{X}_i)$ will be constant in large samples,

and can be combined with the fixed intercept γ_0. However, if the distribution of the predictor varies across clusters such that $h(\mathbf{X}_i)$ varies and the RILR model given in equation (4.2) is fit to the data, $b_i = b_i' + (\gamma_1 - \gamma_2)h(\mathbf{X}_i) \not\perp \mathbf{X}_i$ if $\gamma_1 \neq \gamma_2$, violating assumption (A1).

When model (4.3) holds, and model (4.2) is fit to the data, the cluster-level variable $h(\mathbf{X}_i)$ is omitted. As described by Greenland et al. (1999), omitting $h(\mathbf{X}_i)$ leads to correlation between b_i and $\mathbf{X}_i$, which has the effect of confounding γ_2. In particular, confounders are "now covariates that 'explain' the correlation between" b_i and $\mathbf{X}_i$; in particular, this includes $h(\mathbf{X}_i)$ (Greenland et al., 1999). Others have referred to this as between-cluster confounding, since it is caused by omission of cluster-level variables that are correlated with the included variables (Skrondal and Rabe-Hesketh, 2014). Results from research on omitted variable bias reveal that when (4.3) holds, and (4.2) is fit to the data, $\hat{\gamma}_2$ will generally be a combination of the within- and between-effects (Neuhaus and Kalbfleisch, 1998). Thus, in general, the RILR model given in equation (4.2) may not be appropriate if cluster-varying predictors have different distributions across clusters.

Importantly, the effect that is estimated by this misspecified RILR model, that is, the combination of within- and between-effects, is not of substantive interest and in general lacks clinical relevance (Localio et al., 2001; Neuhaus and Kalbfleisch, 1998). Even in situations where it is thought that the between- and within-cluster effects are reasonably close to one another, there is the potential for differential confounding at the between- versus within-cluster level; thus, using both within- and between-cluster comparisons to estimate the within-cluster effect is generally problematic (Gunasekara et al., 2014; Schildcrout and Heagerty, 2008).

This issue arises very frequently in practice; in particular, Graubard and Korn note that cluster-varying predictors that are balanced across clusters are "difficult to find in observational studies" (Graubard and Korn, 1994). If cluster-level factors are associated with predictors, as is often the case in observational studies, the distribution of the predictors is likely to vary across clusters, which may in turn lead to correlation between the random intercepts and predictors (Neuhaus and McCulloch, 2006; Skrondal and Rabe-Hesketh, 2014).

Many papers in the statistical literature have warned that assumption (A1) may not hold outside of the randomized trial setting (Berlin et al., 1999; Dieleman and Templin, 2014; Gardiner et al., 2009; Graubard and Korn, 1994; Greenland et al., 1999; McCulloch and Neuhaus, 2011; Rabe-Hesketh and Skrondal, 2010; Schildcrout and Heagerty, 2008; Seaman et al., 2014; Ten Have et al., 1995, 1996). However, the fact that this assumption is often untenable is frequently overlooked by investigators, who generally implicitly assume $b_i \perp \mathbf{X}_i$ when implementing RILR models (Berlin et al., 1999; Graubard and Korn, 1994; Neuhaus and Kalbfleisch, 1998; Skrondal and Rabe-Hesketh, 2014).

This problem is often called "confounding by cluster" since the within-cluster association, γ_2, is distorted by the between-cluster association, γ_1 (Berlin et al., 1999; Localio et al., 2002, 2001; Pavlou et al., 2015; Seaman et al., 2014); in the econometrics literature, it is called the "endogenous covariates problem" (Skrondal and Rabe-Hesketh, 2014). As described above, the inclusion of the cluster-level variable(s) responsible for this distortion ($h(\mathbf{X}_i)$ in the example above) allows estimation of the effect of interest, i.e., the within-cluster effect; thus, by the somewhat tautological definition of confounding given by Hernan et al. (2011), since including these cluster-level variables eliminates the confounding of the within-cluster effect, these variables can be considered to be confounders of this effect. In general, the inclusion of cluster-level factors may reduce the potential for confounding by cluster (Localio et al., 2002). Purely cluster-varying predictors, which have no between-cluster effect, are orthogonal to all cluster-level variables, including random intercepts, and thus are not susceptible to confounding by cluster (Heagerty and Zeger, 2000; Localio et al., 2002).

Consider assumption (A2): $b_i \overset{iid}{\sim} N(0, \sigma^2)$. This requires that the random cluster-specific intercepts be independently and identically distributed according to a normal distribution with mean zero and variance σ^2; this is the "mixture distribution." Violations of RILR model assumptions, including normality and homoscedasticity of b_i (and $b_i \perp \mathbf{X}_i$), can generally be cast as misspecifications of the mixture distribution (Neuhaus and McCulloch, 2006). Broadly speaking, misspecifications of the random intercept distribution, including non-normality or heteroscedasticity, may lead to bias in the estimate of the fixed intercept but

typically do not have a large effect on the estimates for cluster-varying predictors (Heagerty and Kurland, 2001; Heagerty and Zeger, 2000; McCulloch and Neuhaus, 2011; Neuhaus et al., 1992). Such misspecifications may also lead to biased estimates of the variance of the random intercept (Lukociene and Vermunt, 2008).

Decomposing Predictors

One solution that has been proposed to address violations of assumption (A1) is to decompose the cluster-varying predictor(s) into a between-cluster component and a within-cluster component (Begg and Parides, 2003; Brumback et al., 2012; Heagerty and Zeger, 2000; McCulloch and Neuhaus, 2011; Neuhaus and Kalbfleisch, 1998; Neuhaus and McCulloch, 2006; Schildcrout and Heagerty, 2008; Seaman et al., 2014; Ten Have et al., 1996). In the context of the model given in equation (4.3), this would mean fitting a model with $h(\mathbf{X}_i)$ and x_{ij} as predictors. This method has been used in settings where there is a single predictor, often when the predictor is binary; this simplifies the choice of $h(\cdot)$, and the cluster mean is commonly used (Neuhaus and Kalbfleisch, 1998). When $h(\mathbf{X}_i) = \bar{\mathbf{X}}_i$, the method is called the "poor man's" method (Neuhaus and Kalbfleisch, 1998). However, using the cluster mean may be overly simplistic as it only addresses confounding by cluster when the cluster mean fully captures all the unmeasured cluster-level characteristics responsible for the confounding (Berlin et al., 1999). More flexible methods have been proposed based on modeling b_i as a function of $\mathbf{X}_i$ (Brumback et al., 2010). Of course, these methods require that the model for b_i be correctly specified (Brumback et al., 2010, 2012; Neuhaus and McCulloch, 2006). Thus, the idea of decomposing predictors typically replaces one set of assumptions with another (Brumback et al., 2012).

Efficiency

RILR is often touted as being more efficient than alternative methods. This efficiency arises in part from assuming some (parametric) distribution for the random intercepts (Neuhaus and Lesperance, 1996; Ten Have et al., 1995). In addition, RILR uses both between- and

within-cluster comparisons to estimate the effect of predictors that vary within and between clusters, which allows it to use more information in estimating these effects (Kahan, 2014; Seaman et al., 2014; Ten Have et al., 1995). Thus, under the full assumptions of RILR, the resulting estimates are efficient relative to common alternatives (Gardiner et al., 2009). If a variant of RILR that does not assume equal between- and within-cluster effects is used (e.g., the poor man's method), and the model is correctly specified, estimates may be more efficient than those obtained from non-RILR methods (Seaman et al., 2014). Some studies have found reduced efficiency when the distribution of the random intercept is not normal (Lukociene and Vermunt, 2008).

4.2.3 Fixed Intercept Logistic Regression

Fixed intercept models are a special case of generalized linear models, where a fixed intercept for each cluster is used to model clustered data. We consider two variants of fixed intercept logistic regression (FILR): conditional (cFILR) and unconditional (uFILR). Both cFILR and uFILR have the same form:

$$\text{logit}\left\{P(D_{ij}=1|\mathbf{X}_{ij})\right\} = \beta_{0i} + \boldsymbol{\beta}^{\top}\mathbf{X}_{ij}, \tag{4.4}$$

where β_{0i} represents a cluster-specific intercept. cFILR and uFILR differ in their approach to estimation: uFILR relies on the full likelihood, while cFILR uses a conditional likelihood, conditioning on the number of cases in each cluster ($\sum_j D_{ij}$). The development of cFILR was motivated by the "incidental parameters problem," which occurs when the number of parameters grows with the sample size (Neyman and Scott, 1948).

The use of FILR has been advocated as a reasonable (and often preferred) alternative to other models for clustered data (Brumback et al., 2012; Gardiner et al., 2009; Gunasekara et al., 2014; Heagerty and Zeger, 2000; Localio et al., 2001; Neuhaus and McCulloch, 2006; Neuhaus et al., 2014; Skrondal and Rabe-Hesketh, 2008, 2014; Whittemore and Halpern, 2003). Here we discuss FILR and provide some comparisons with RILR.

Assumptions

In the econometrics literature, the distinction between RILR and FILR is based not on whether the cluster-specific intercepts are fixed or random, but whether they are independent of the predictors (Gardiner et al., 2009). Thus, the key assumption for FILR is (Gardiner et al., 2009; Seaman et al., 2014):

(B1) Conditional on $\mathbf{X}_i$, the $D_{i1}, ...D_{in_i}$ are independent

If assumption (B1) holds, cFILR will yield consistent estimates of $\boldsymbol{\beta}$. If, additionally, the cluster size increases faster than the number of clusters, uFILR will yield consistent estimates of $\boldsymbol{\beta}$ (Ten Have et al., 1995). If the β_{0i} are random, then (B1) must also condition on β_{0i} and the β_{0i} must be independent across clusters (Gardiner et al., 2009; Seaman et al., 2014). When there is reason to suspect that the predictors are associated with b_i, FILR is generally recommended over RILR (Gardiner et al., 2009).

Estimates

FILR consistently estimates the within-cluster effect of predictors that vary within clusters, provided (B1) is satisfied and (4.4) holds (Berlin et al., 1999; Hu et al., 1998; Neuhaus and Kalbfleisch, 1998). Thus, this method avoids the issue of confounding by cluster; in fact, the resulting estimates are not affected by confounding from any unmeasured cluster-constant variables (Berlin et al., 1999; Gunasekara et al., 2014). This is because the cluster-specific intercepts that are included in the FILR model need not be independent of $\mathbf{X}_i$ and so can include the effect of any cluster-constant variables, such as $h(\mathbf{X}_i)$ in the example given above.

As noted above, uFILR maximizes the full data likelihood to obtain parameter estimates, while cFILR maximizes the conditional likelihood. For both methods, since only within-cluster comparisons are used to estimate the parameters, clusters for which all observations have $D = 1$ or all observations have $D = 0$ (what we will call "concordant clusters") do not contribute any information to the estimation of the within-cluster effect (without additional assumptions) and thus are not used in estimation (Gardiner et al., 2009; Ten Have et al.,

1995). This is also true of clusters that are concordant on the predictors, though this situation is generally unlikely in when there are multiple and/or continuous predictors.

Efficiency

Many investigators are hesitant to use FILR since the exclusion of concordant clusters could reduce efficiency (Neuhaus and Kalbfleisch, 1998). Indeed, the efficiency of estimates from cFILR relative to those from RILR improves as the probability of concordance decreases, indicating the role of concordance in considerations of efficiency (Neuhaus and Lesperance, 1996; Ten Have et al., 1995). However, previous research has shown that cFILR provides estimates that are efficient relative to RILR for predictors that predominantly vary within-clusters (Neuhaus et al., 2014; Schildcrout and Heagerty, 2008). Indeed, as pointed out by Neuhaus and Kalbfleisch (1998), for predictors with between- and within-cluster components, the increased efficiency of estimates from RILR that is sometimes observed (particularly when the clusters are not very small) is often largely due to the assumption of common within- and between-cluster effects.

If these effects are indeed equal, there will be some efficiency gain from using RILR since this method uses both within- and between-cluster variations to estimate the predictor effect (Schildcrout and Heagerty, 2008). However, using both types of variation in estimation is generally not recommended due to the potential for differential confounding (Schildcrout and Heagerty, 2008). Importantly, concordant clusters contribute to the between-cluster variation and often exhibit strong between-cluster effects; these can heavily distort the estimated effect for within-cluster predictors if RILR is used when the within- and between-cluster effects are not equal (Hu et al., 1998; Ten Have et al., 1996). Furthermore, if there is no between-cluster variation in the predictor (that is, the distribution of the predictor is the same in each cluster), RILR ignores concordant clusters (Localio et al., 2001). When the random intercept is associated with predictors, and this association is correctly modeled via RILR with a modified mixture distribution, the resulting estimates are only slightly more efficient than estimates from cFILR (even with fairly small clusters) (Neuhaus and McCulloch,

2006).

4.2.4 Asymptotics for Clustered Data

When the number of clusters is fixed, but the cluster size is increasing, the estimates from uFILR and cFILR are asymptotically equal (Hauck, 1984). However, when the number of clusters is growing, uFILR is susceptible to the incidental parameters problem (Hauck, 1984). On the other hand, the estimates from cFILR will be consistent in this setting. For RILR, asymptotic results generally correspond to the setting where the number of clusters is growing (Huang, 2009). If $b_i \not\perp \mathbf{X}_i$, estimates from RILR are expected be consistent when the cluster size is growing if the within-cluster variation of the predictors is sufficiently large since in this situation, within-cluster comparisons will dominate between-cluster comparisons asymptotically (Brumback et al., 2010, 2012; Gunasekara et al., 2014; Skrondal and Rabe-Hesketh, 2014).

4.2.5 Risk Prediction and Clustered Data

In the risk prediction setting, the focus is not on the estimate of a single parameter, as in the etiologic and therapeutic settings, but rather on the combination of several parameter estimates and corresponding predictor values (Bouwmeester et al., 2013). When clustered data are used to develop combinations of biomarkers for diagnosis or prognosis, additional challenges emerge. First, a decision must be made regarding whether interest is in conditional or marginal estimates; often this choice will be dictated by the nature of the clustering variable. For example, if the clusters are centers, researchers are typically interested in the conditional association because allowing center to be predictive limits generalizability and interpretability. Challenges also arise in the application of biomarker combinations for diagnosis or prognosis in the clustered data setting. In order to generate conditional predicted probabilities, an estimate of the cluster-specific intercept must be available. This will generally not be the case for clusters not used in constructing the combination (Bouwmeester et al., 2013; Wynants et al., 2015).

4.2.6 Evaluating Performance

Suppose we have a generic predictor Z and are interested in evaluating its performance. In the context of risk prediction, interest generally centers on evaluating discrimination and calibration. Although both aspects of performance are essential to determining whether a risk prediction instrument should be used clinically, we will primarily focus on discrimination, since determining the discriminative ability of a predictor is often the first step in developing a clinically useful diagnostic or prognostic tool. Discrimination is the ability of Z to separate cases (individuals who have or will experience the outcome, $D = 1$) and controls (individuals who do not have or will not experience the outcome, $D = 0$) and is most commonly assessed via the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve plots the true positive rate, the proportion of correctly classified cases, versus the false positive rate, the proportion of incorrectly classified controls, over the range of possible thresholds for Z; thus, it exists on the unit square (Pepe, 2003). 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 the AUC for such a predictor is 1 (Pepe, 2003). The AUC has a probabilistic interpretation: it is the probability that, for a randomly chosen case-control pair, the value of Z for the case is higher than the value of Z for the control, assuming that higher values of Z are more indicative of D (Pepe, 2003).

As described above in the context of constructing combinations, when data come from multiple centers, it is generally important to avoid allowing center to be predictive, so conditional approaches to constructing combinations are appropriate. Likewise, in order to prevent center from contributing to the assessment of the discriminatory ability of Z, the center-adjusted AUC, a conditional measure, should generally be used to evaluate performance; this is analogous to the center-adjusted odds ratio in multicenter association studies (Janes and Pepe, 2008).

The center-adjusted ROC (aROC) and corresponding center-adjusted AUC (aAUC),

proposed by Janes and Pepe (2009) for general covariate adjustment, can be written as $aROC_Z(t)$ and $aAUC_Z$, respectively, where

$$\begin{aligned} aAUC_Z &= \int_0^1 aROC_Z(t)dt \\ &= \int_0^1 \int ROC_{Z|C=c}(t)dP_D(c)dt \\ &= \sum_c w_c AUC_{Z|C=c}, \end{aligned} \tag{4.5}$$

where t is the false positive rate, $ROC_{Z|C=c}$ and $AUC_{Z|C=c}$ are the center-specific ROC and AUC, respectively, and $P_D(c)$ is the distribution of center among cases. When the center-specific AUC is constant across centers, the adjusted AUC is simply that center-specific AUC; in general, the aAUC is a weighted average of the center-specific AUCs, where the weights correspond to the proportion of cases in each center (Janes et al., 2009; Janes and Pepe, 2009).

When the ROC curve varies by a covariate (a type of effect modification), it is generally recommended that a separate ROC curve be estimated for each value of the covariate (Janes et al., 2009). In the case of center, where only a fraction of the existing centers are observed, this is not possible. However, it is reasonable to assess the heterogeneity in the center-specific AUCs, as this provides some indication of how the predictor may perform in a new center (Bouwmeester et al., 2013).

The marginal AUC (the AUC calculated without regard to center) would be appropriate if between-center heterogeneity were to be used in making decisions, which is not typically the case (van Klaveren et al., 2014). The marginal AUC can be thought of as a summary of the data at hand; that is, the ability of Z to discriminate between cases and controls in a particular set of centers (Janes and Pepe, 2008; van Oirbeek and Lesaffre, 2010). In general, some summary of the conditional, or center-specific, AUCs should be used to avoid allowing center to contribute to the discriminatory ability of the predictor. However, different authors have proposed different means of combining the center-specific AUCs into an overall center-

adjusted measure. Several authors have proposed using a simple average of center-specific AUCs (Bouwmeester et al., 2013; van Oirbeek and Lesaffre, 2010; Wynants et al., 2015) or fixed- and random-effects meta-analysis methods (Riley et al., 2015; Snell et al., 2016; van Klaveren et al., 2014).

The measure defined in equation (4.5) is compelling because it corresponds to the area under the ROC curve given by the weighted average of center-specific true-positive rates, holding the center-specific false-positive rates constant (Janes and Pepe, 2009):

$$aROC_Z(t) = P\Big(Z > g(t|c)\Big|D = 1\Big),$$

where $g(t|c)$ is the center-specific threshold giving a false positive rate of t in each center. In other words, the center-adjusted ROC curve, on which the aAUC is based, is the ROC curve corresponding to the true and false positive rates based on center-specific thresholds; these center-specific thresholds are chosen such that the false positive rate is the same in each center (Janes and Pepe, 2009). When other weights are used, this interpretation of aROC does not apply.

Previous work has considered the relationship between the marginal ROC, ROC_Z, and the conditional ROC for a covariate C, $ROC_{Z|C}$ (Kerr and Pepe, 2011). If the distribution of X among controls varies across values of C, but C is independent of the outcome, ROC_Z will be attenuated relative to $ROC_{Z|C}$ (Janes et al., 2009; Janes and Pepe, 2008). In this situation, C is said to "calibrate" Z (Janes and Pepe, 2008; Kerr and Pepe, 2011). In general, if C is associated with Z and the outcome, ROC_Z will differ from $ROC_{Z|C}$ since ROC_Z includes part of the discriminatory ability of C (Janes and Pepe, 2008).

When the same data are used to construct a combination and evaluate its performance, the resulting estimate of performance is optimistic (Copas and Corbett, 2002). This type of optimistic bias is often called "resubstitution bias" (Kerr et al., 2015). Typically, when the data are not clustered, this is addressed by using a bootstrapping procedure to estimate the degree of optimism (Copas and Corbett, 2002; Harrell, 2013). Bootstrapping assumes

exchangeability of the observations, which may not hold in the multicenter setting due to the correlation among observations in the same center (Bouwmeester et al., 2013). Thus, it is often recommended that bootstrapping preserve the effect of clustering by resampling by center (Janes et al., 2009; Localio et al., 2001; van Oirbeek and Lesaffre, 2010). However, Bouwmeester et al. (2013) found similar results for the average of cluster-specific AUC estimates whether resampling was done on clusters or individual observations.

4.2.7 Biomarker Combinations

Many biomarker assays are now capable of measuring multiple biomarkers, and the cost of biomarker assays in general continues to decline. As a result, it is becoming increasingly common to collect data on an array of biomarkers and, consequently, there is great interest in constructing combinations of these biomarkers for diagnosis or prognosis.

For a collection of biomarkers $\mathbf{X}$, the risk score, $P(D = 1|\mathbf{X})$, is optimal in terms of maximizing the true positive rate at each false positive rate among all possible combinations of the biomarkers $\mathbf{X}$ (McIntosh and Pepe, 2002). Thus, to the extent that the linear logistic model $P(D = 1|\mathbf{X}) = \text{expit}(\boldsymbol{\beta}^\top \mathbf{X})$ holds, the combination $\boldsymbol{\beta}^\top \mathbf{X}$ will be optimal. As the linear logistic model may not hold, methods have been developed to optimize the AUC within the class of linear combinations $\boldsymbol{\beta}^\top \mathbf{X}$ without relying on this model (Pepe et al., 2006). However, logistic regression is quite robust and, although there are no guarantees, in many cases will provide a reasonably good fit to the data, and so is often used to construct biomarker combinations (Fong et al., 2016; Lin et al., 2011; Ma and Huang, 2007; Pepe et al., 2006). Likewise, including biomarkers as linear terms may not reflect the underlying data-generating distribution (Pepe and Thompson, 2000), but it is often a reasonable choice and has intuitive appeal for clinical collaborators.

Methods have also been developed to identify combinations of biomarkers that maximize AUC while accommodating covariates (Liu and Zhou, 2013; Schisterman et al., 2004). However, the method proposed by Liu and Zhou (2013) is computationally challenging when there are more than two biomarkers to be combined, and the method proposed by Schister-

man et al. (2004) relies on an assumption of multivariate normality of the biomarkers and specification of the relationship between the covariates (e.g., center) and the biomarkers.

Previous research has found that combining biomarkers does not always result in better performance (Bansal and Pepe, 2013). In general, when considering biomarker combinations, it is recommended that the performance of the combinations be evaluated, rather than deciding which biomarkers to combine based on marginal performance and/or correlation among biomarkers (Bansal and Pepe, 2013).

4.3 Methods

When the data come from a single center, common practice is to first construct a (linear) combination of the biomarkers, often using logistic regression, and evaluate its performance using measures such as the AUC. With more than one center, it is important to consider how to appropriately accommodate center in both the construction and evaluation of biomarker combinations. As with the center-adjusted odds ratio in multicenter etiologic studies or the center-adjusted treatment effect in multicenter randomized trials, we propose using conditional approaches in the construction and evaluation of biomarker combinations, in particular, using FILR to construct combinations and the center-adjusted AUC to evaluate them.

Throughout, we focus on constructing a single biomarker combination, as opposed to fitting center-specific combinations; that is, we do not allow the relationship between the biomarkers and the outcome to vary across centers. In the clinical trial setting, assessing treatment-by-center interactions is usually not part of the primary analysis (Kahan, 2014). Analogously, in the risk prediction setting, it is preferable to give a single combination that is not center-specific, as this would make development highly localized. Of course, this may mean that we do not identify the combination that is "optimal" in every center. However, our primary goal is to identify a single combination and evaluate its performance across centers, though as we discuss below, it is generally informative to assess the degree of heterogeneity in performance across centers for a given biomarker combination. In this section and the simulations that follow, we consider situations where there is no effect modification

by center (either in terms of the "true" biomarker combination or its performance); we then examine the potential impact of effect modification by center in the Discussion. We focus on constructing linear combinations via the logistic regression framework. While this may seem restrictive, Pepe et al. (2006) noted that the class of linear combinations is actually quite large (taking into consideration possible biomarker transformations and interactions) and, as mentioned above, the logistic form is fairly robust.

We consider a setting where we have data from m centers (where the population consists of $M \in [m, \infty]$ centers) with n_c ($c = 1, ..., m$) observations in each, a p-dimensional vector of biomarkers $\mathbf{X}$, and a binary outcome D. In general, let $\mathbf{X}$ denote the vector of biomarkers for an arbitrary individual. Let C indicate center and D denote the binary outcome, where $D = 1$ or the subscript D indicate cases, and $D = 0$ or the subscript $\bar{D}$ indicate controls. In each center there are n_D^c cases and $n_{\bar{D}}^c$ controls, and there are n_D total cases and $n_{\bar{D}}$ total controls. Throughout, we will assume a non-trivial center-specific disease prevalence; that is, $P(D = 1|C = c) := \gamma_c \in [1/V, 1 - 1/V]$, $c = 1, ..., M$, for $V \in (2, \infty)$.

4.3.1 The Role of Center

As indicated above, biomarkers may be either diagnostic or prognostic: diagnostic biomarkers represent some underlying disease or disease process (i.e., $D \rightarrow \mathbf{X}$), while prognostic biomarkers cause some future outcome (that is, $\mathbf{X} \rightarrow D$).

We consider the role of center in the context of two sets of characteristics:

1. Characteristics affecting the prevalence of D: differences in the populations served by each center could affect the prevalence of D.

2. Characteristics affecting biomarker measurements: center-level factors, including storage and handling of specimens and hospital-level practices, could lead to variations in biomarker measurements unrelated to D.

In other words, we consider two ways in which center differences may arise: (1) through differences in prevalence and (2) through differences in biomarker measurements unrelated

to D. Others have similarly noted the potential for center differences to arise through differences in the patients at each center and/or through differences in the centers themselves (for example, differences in protocols and practices) (Janes and Pepe, 2008; Kahan, 2014).

We focus on three possibilities for the role of center. We call center a *confounder* when center separately affects both the prevalence of D and biomarker measurements, a *case mix variable* when center affects only the prevalence of D, and a *calibration variable* when center affects only the biomarker measurements, separate from D. This taxonomy helps researchers to understand and communicate the role of center in their studies, highlights the possible causes of center differences, and provides insight into the potential impact of different decisions made during data analysis.

In the TRIBE-AKI study, where the goal is to use biomarkers to diagnose AKI, it may be the case that certain centers serve particularly unhealthy communities and that this results in differences in biomarker levels; however, these differences may simply reflect true underlying biology. If factors such as storage and handling of biomarkers and surgical practices are standardized, such that the distribution of biomarkers is similar across centers, conditional on case status, center would be a case mix variable. If, however, these factors vary across centers (e.g., in some centers surgeons use different protocols for fluid administration) in addition to variability in disease prevalence, center would be a confounder. On the other hand, if the populations served by each center are relatively similar in terms of underlying AKI risk, but factors such as surgical protocols vary across centers and lead to variations in biomarker measurements, center would be a calibration variable.

We can also consider an example with prognostic biomarkers. Suppose carotid intima-media thickness is used to predict which patients will experience stroke. If certain centers tend to serve less healthy populations (i.e., those at greater risk for stroke), but practices for measuring carotid intima-media thickness are standardized across centers, center would be a case mix variable. If the make-up of patients in terms of underlying risk of stroke is similar across centers, but different protocols or imaging tools are used at different centers such that the distribution of intima-media thickness measurements varies across centers, center would

be a calibration variable. If both the composition of patients and intima-media thickness measurement practices vary across centers, center would be a confounder.

In Figure 4.1, we present graphical and probabilistic depictions of center as a case mix variable, a calibration variable, and a confounder for a set of diagnostic biomarkers $\mathbf{X}$. Likewise, in Figure 4.2, we present graphical and probabilistic depictions of center as a case mix variable, a calibration variable, and a confounder for a set of prognostic biomarkers $\mathbf{X}$. While the graphical depictions for prognostic biomarkers are complementary to those for diagnostic markers, the probabilistic depictions include important differences. We make a distinction between diagnostic and prognostic biomarkers (and invoke causality in discussing them) because this allows application of probabilistic notions of dependence, which will in turn provide insights into the role of center and the repercussions of applying different methods to multicenter data.

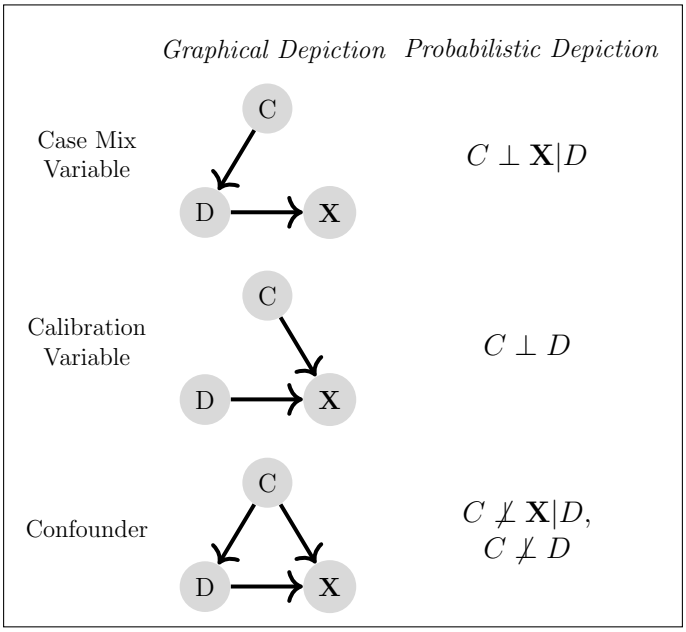

Figure 4.1: Select potential roles of center in studies of diagnostic markers.

It is important to distinguish center as a confounder, as described above and defined in Figures 4.1 and 4.2, from "confounding by cluster" in the context of a RILR model. The

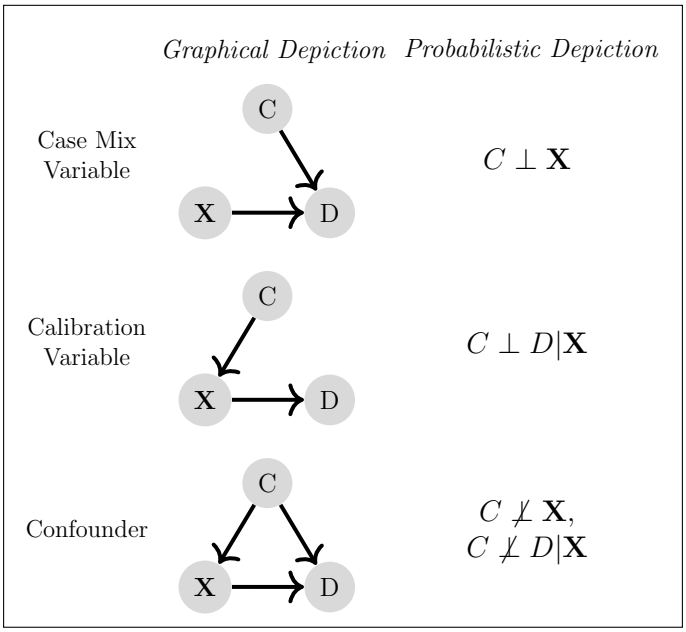

Figure 4.2: Select potential roles of center in studies of prognostic markers.

definition of "confounding" in Figures 4.1 and 4.2 is in line with standard epidemiological notions of confounding. The idea of "confounding by cluster" for RILR models, on the other hand, is specific to the RILR framework: "confounding by cluster" occurs when the random intercepts and the predictor(s) are not independent, leading to an omitted cluster-level covariate that is associated with the predictor(s). In order to address violations of assumption (A1), and in so doing resolve the distortion of the within-cluster effect caused by this omitted covariate, the covariate must be included in the RILR model. As we will see, when the biomarkers are diagnostic and center is either a case mix variable or a calibration variable, the random intercepts and the biomarkers may not be independent, so in the context of the RILR model, we are susceptible to "confounding by cluster," even though center is not a confounder by the definition in Figure 4.1.

4.3.2 Ignoring Center

Clinical researchers frequently ignore center in the construction and/or evaluation of diagnostic or prognostic biomarker combinations (e.g., Shapiro et al. (2009); Vuilleumier et al. (2008)). This is likely due to the fact that investigators acknowledge the unique role center can play, in the sense that it could be predictive of the outcome yet should not naïvely be included as a predictor, but are not familiar with methods for accommodating center or the repercussions of ignoring it. In general, attention must be paid to role of center in both the construction and evaluation of biomarker combinations for diagnosis or prognosis.

Construction

Suppose the linear-logistic model holds:

$$\text{logit}\left\{P(D=1|\mathbf{X},C=c)\right\} = \beta_0^c + \boldsymbol{\beta}^\top \mathbf{X}. \tag{4.6}$$

Such a model could arise from the following data-generating model for two biomarkers, $\mathbf{X} = (X_1, X_2)$:

$$\left(\begin{array}{c} X_1 \\ X_2 \end{array} \middle| \, D=d, C=c \right) \sim N\left(\left(\begin{array}{c} f_{X_1}(c) + \mu_{X_1} d \\ f_{X_2}(c) + \mu_{X_2} d \end{array} \right), \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right) \right) \tag{4.7}$$

where μ_{X_1} and μ_{X_2} are related to the AUC for each marker ($\mu_{X_1} = \sqrt{2}\Phi^{-1}(AUC_{X_1|C})$ and $\mu_{X_2} = \sqrt{2}\Phi^{-1}(AUC_{X_2|C})$). We consider constant center-specific AUCs for the individual markers, and thus allow for center effects on biomarker levels via conditional mean shifts. Equation (4.7) gives

$$\text{logit}\left\{P(D=1|\mathbf{X},C=c)\right\} = \beta_0^c + \beta_1 X_1 + \beta_2 X_2.$$

where β_0^c is a center-specific offset and, as shown in Appendix A.2.1,

$$\beta_0^c = \frac{-\mu_{X_1}^2 - \mu_{X_2}^2}{2(1-\rho^2)} + \frac{\rho\mu_{X_1}\mu_{X_2} + \rho\mu_{X_1}f_{X_2}(c) + \rho\mu_{X_2}f_{X_1}(c)}{1-\rho^2} - \frac{\{\mu_{X_1}f_{X_1}(c) + \mu_{X_2}f_{X_2}(c)\}}{1-\rho^2} + \log\left(\frac{\gamma_c}{1-\gamma_c}\right),$$

and

$$\beta_1 = \frac{\mu_{X_1} - \rho\mu_{X_2}}{1-\rho^2}, \;\; \beta_2 = \frac{\mu_{X_2} - \rho\mu_{X_1}}{1-\rho^2}.$$

Thus, the risk score $P(D = 1|\mathbf{X}, C)$ can be written in linear-logistic form where center is included as a nominal adjustment (stratification) variable. This conditional distribution leads to a biomarker combination $(\beta_1 X_1 + \beta_2 X_2)$ that is the same across centers.

Returning to the general linear-logistic model given in (4.6), suppose that the model holds, but β_0^c is not allowed to vary across centers. That is, suppose we fit the following (potentially misspecified) model to the data pooled across centers:

$$\text{logit}\left\{P(D = 1|\mathbf{X})\right\} = \alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X}. \tag{4.8}$$

Based on results for misspecified regression models, $(\alpha_0, \boldsymbol{\alpha})$ minimize the Kullback-Leibler divergence between the marginal model (4.8) and the true data-generating model (Akaike, 1998). That is, the estimates $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}})$ obtained from equation (4.8) will converge to the values $(\alpha_0, \boldsymbol{\alpha})$ that minimize the Kullback-Leibler divergence. When $C \perp D|\mathbf{X}$ or $C \perp \mathbf{X}|D$, and model (4.6) holds, we have collapsibility (Guo and Geng, 1995). This means that the conditional and marginal coefficients are the same $(\boldsymbol{\alpha} = \boldsymbol{\beta})$ and the marginal logit, $\text{logit}\{P(D = 1|\mathbf{X})\}$, is still linear. Therefore, in these situations, the relationship between the biomarkers and the outcome is the same whether we condition on center or marginalize over it. Furthermore, in model (4.6), when $C \perp D|\mathbf{X}$, β_0^c will not vary across centers, so $\alpha_0 = \beta_0^c$. This means that for a calibration variable in the prognostic setting, there is no impact of ignoring center on any of the parameters (i.e., the intercept and the biomarker

coefficients) under model (4.6). Thus, under model (4.6), the biomarker coefficients will not be affected by ignoring center in some special situations.

However, when 4.6 holds, $C \not\perp D|\mathbf{X}$, and $C \not\perp \mathbf{X}|D$, center may not be able to be ignored even if we are only interested in the biomarker coefficients, as will be seen below. Furthermore, the linear-logistic model (4.6) may not hold, in which case the collapsibility results will no longer be expected to apply. More generally, ignoring center in the construction of the biomarker combination potentially allows center to be predictive; that is, part of the effect of center may be included in the estimates of the biomarker coefficients when center is omitted. As noted above, allowing center to be predictive limits generalizability and interpretability.

Evaluation

We focus on evaluating the discrimination of biomarker combinations; that is, how well the combination separates cases and controls. This is because we view our methods as most applicable in the early stages of biomarker combination development, where interest is in identifying promising combinations of biomarkers for further study and refinement, including modifying assays and practices related to sample collection and storage with the goal of reducing or eliminating center differences. Specifically, we focus on the AUC as a measure of discrimination.

As with biomarker combination construction, center is often ignored in the evaluation of combinations (Janes and Pepe, 2008). Suppose we have two biomarkers, X_1 and X_2, and we have a linear combination: $L_{\boldsymbol{\theta}}(\mathbf{X}) = \boldsymbol{\theta}^\top \mathbf{X} = \theta_1 X_1 + \theta_2 X_2$. $L_{\boldsymbol{\theta}}(\mathbf{X})$ can be thought of as a "super marker." When center is ignored in the evaluation of $L_{\boldsymbol{\theta}}(\mathbf{X})$, the data are pooled across centers. This gives the marginal AUC, $AUC(\boldsymbol{\theta}) = P(L_{\boldsymbol{\theta}}(\mathbf{X}_D) > L_{\boldsymbol{\theta}}(\mathbf{X}_{\bar{D}}))$, where $\mathbf{X}_D$ is the vector of biomarkers for an arbitrary case and $\mathbf{X}_{\bar{D}}$ is the vector of biomarkers for an

arbitrary control. In practice, $AUC(\boldsymbol{\theta})$ is estimated empirically:

$$\hat{AUC}(\boldsymbol{\theta}) = \frac{\sum_{i:D_i=1, j:D_j=0} 1(\theta_1 X_{1i} + \theta_2 X_{2i} > \theta_1 X_{1j} + \theta_2 X_{2j})}{n_D n_{\bar{D}}},$$

where D_i is the outcome for the i^{th} observation, X_{1i} is the value of X_1 for the i^{th} case and X_{1j} is the value of X_1 for the j^{th} control, and X_{2i} and X_{2j} are defined analogously. When the biomarkers $\mathbf{X}$ are associated with center such that $L_{\boldsymbol{\theta}}(\mathbf{X})$ is also associated with center, the marginal AUC includes part of the discriminatory accuracy of center, even if center is taken into account when constructing the combination (Janes and Pepe, 2008).

This is in contrast to the conditional, or center-specific, AUC, AUC_c. This corresponds to evaluating combinations by stratifying by center and is written $AUC_c(\boldsymbol{\theta}) = P(L_{\boldsymbol{\theta}}(\mathbf{X}^c_D) > L_{\boldsymbol{\theta}}(\mathbf{X}^c_{\bar{D}}))$, where $\mathbf{X}^c_D$ is the vector of biomarkers for an arbitrary case in center c and $\mathbf{X}^c_{\bar{D}}$ is the vector of biomarkers for an arbitrary control in center c. Of course, it is possible for $AUC_c(\boldsymbol{\theta})$ of a given combination to vary across centers. In practice, $AUC_c(\boldsymbol{\theta})$ is estimated empirically:

$$\hat{AUC}_c(\boldsymbol{\theta}) = \frac{\sum_{i:D^c_i=1, j:D^c_j=0} 1(\theta_1 X^c_{1i} + \theta_2 X^c_{2i} > \theta_1 X^c_{1j} + \theta_2 X^c_{2j})}{n^c_D n^c_{\bar{D}}},$$

where D^c_i is the outcome for the i^{th} subject in center c, X^c_{1i} is the value of X_1 for the i^{th} case in center c and X^c_{1j} is the value of X_1 for the j^{th} control in center c, and X^c_{2i} and X^c_{2j} are defined analogously. Note that AUC_c can only be estimated in discordant centers (those with at least one case and one control).

4.3.3 Accounting for Center

Often, multicenter studies of association account for center in some way, typically by estimating a center-adjusted measure of association. In multicenter randomized trials, randomization is often stratified by center and the target of estimation is then the center-adjusted treatment effect (Kahan, 2014). This idea can be extended to the construction and evaluation

of biomarker combinations for diagnosis and prognosis.

Indeed, the limited statistical literature related to center effects in risk prediction generally recognizes that ignoring center is problematic. That is, when the goal is to apply biomarker combinations to individuals from new centers, it is important to stratify by center when constructing the combination (i.e., the combination should be constructed conditional on center) and evaluating the combination (using a summary of conditional performance) (Bouwmeester et al., 2013; van Klaveren et al., 2014). Therefore, we focus on methods that stratify (condition) on center for both construction and evaluation. As a consequence of focusing on conditional AUC, we do not need an estimate of the intercept in each center to evaluate the combination, as the center-specific AUC is a rank-based measure and so would be unaffected by such estimates.

Construction

We can accommodate center in the construction of biomarker combinations by stratifying by center. In the standard regression framework, this means that center will be included in the model with the biomarkers in some form. If the goal is to construct a single biomarker combination, there are two main ways the logistic regression framework could be used:

1. Random intercept model (including the "poor man's" method)
2. Fixed intercept model (including conditional regression methods)

Specifically, we will consider RILR and FILR; for FILR, we will consider both cFILR and uFILR. For concreteness, we consider $p = 2$ biomarkers in the discussion below. In discussing RILR and FILR, we suppress some of the subscript notation used above in introducing these models. Recall that $\mathbf{X}$ is used to denote the vector of biomarkers for an arbitrary individual.

Construction: RILR

To the limited extent that the literature has acknowledged the potential role of center in the prediction setting, RILR models are often the chosen method for estimation (Bouwmeester

et al., 2013). This model can be written as

$$\begin{aligned} \text{logit}\left\{P(D=1|\mathbf{X},b_c)\right\} &= b_c + \tau_0 + \tau_1 X_1 + \tau_2 X_2, \\ b_c &\overset{iid}{\sim} F(0,\sigma^2), \end{aligned} \tag{4.9}$$

where the distribution of the random center-specific intercepts, F, is typically assumed to be normal. The model makes three key assumptions, (A1)-(A3). Assumption (A1) is satisfied if $C \perp \mathbf{X}$. In general, when the distribution of X_1 or X_2 varies by center, assumption (A1) may not hold and the corresponding estimates $(\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2)$ may not be meaningful.

For biomarkers measured in multiple centers, it is generally unreasonable to expect the distribution of the biomarkers to be the same across centers. The "poor man's" method may be useful in addressing violations of (A1), and can be written

$$\begin{aligned} \text{logit}\left\{P(D=1|X_1,X_2,b_c^*)\right\} = {} & b_c^* + \tau_0^* + \tau_1^W(X_1 - \bar{X}_{1c}) + \tau_2^W(X_2 - \bar{X}_{2c}) \\ & + \tau_1^B \bar{X}_{1c} + \tau_2^B \bar{X}_{2c}, \\ b_c^* \sim {} & F(0,\sigma^{*2}), \end{aligned}$$

where $\bar{X}_{1c}$ and $\bar{X}_{2c}$ are the center-specific means of X_1 and X_2, respectively, b_c^* and τ_0^* represent the random center-specific and overall (fixed) intercepts, respectively, τ_1^W and τ_2^W represent the within-center effects of the biomarkers, and τ_1^B and τ_2^B represent the between-center effects of the biomarkers. This method has been proposed for the case where there is one predictor and may not fully address violations of (A1) in the multivariable setting. Importantly, this method relies on a RILR model, so assumptions (A1)-(A3) are still required. If the differences in the distribution of biomarkers across center are captured by mean shifts, b_c^* and $(X_1 - \bar{X}_{1c}, X_1 - \bar{X}_{1c})$ would be independent in large samples.

Construction: FILR

An option that has been discussed at length in the literature on multicenter randomized trials, but has been largely (if not entirely) neglected in the prediction literature is FILR (Berlin et al., 1999; Localio et al., 2001; Neuhaus and Kalbfleisch, 1998). We propose using uFILR when the number of centers is modest, and cFILR when the number of centers is large. The FILR model can be written as

$$\text{logit}\left\{P(D=1|\mathbf{X},C=c)\right\}=\beta_0^c+\beta_1X_1+\beta_2X_2.$$

If the β_0^c are not random, this model relies on assumption (B1). When the number of centers is large relative to the total sample size or the number of centers is expected to grow with the total sample size, cFILR is preferred to uFILR in order to avoid the incidental parameters problem.

RILR vs. FILR in Diagnostic and Prognostic Research

Random intercept models are, at first glance, appealing for data from multicenter studies in the context of prediction: these models are thought to represent a situation where there exists a large population of centers, and the data at hand constitute a random draw of centers from that population. This intuition may make investigators more comfortable with generalizing their results to centers not included in their data, typically the goal of prediction research, and thus more likely to use RILR. However, since the key distinction between random and fixed intercept models is not necessarily whether the center-specific intercepts are random or fixed, but rather whether they are associated with the biomarkers, thinking about center-specific intercepts as random as opposed to fixed generally offers little meaningful benefit (Gardiner et al., 2009).

Researchers may also be drawn to RILR since it gives an estimate of the overall intercept τ_0 and the center-specific intercepts b_c are typically assumed to be normally distributed with mean 0; this leads researchers to believe that they can provide predicted probabilities for

patients in new centers not used in model fitting via $\hat{\tau}_0 + \hat{\tau}_1 X_1 + \hat{\tau}_2 X_2$. However, assuming $b_c = 0$ in new centers generally leads to poor calibration; that is, it does not provide useful estimates of risk (Pavlou et al., 2015). Even if a valid estimate of b_c for a new center is available, the estimate of τ_0 from RILR can be badly biased if the random intercept distribution is misspecified. Thus, the fact that RILR seemingly provides a way to estimate predicted probabilities does not justify its use.

The "poor man's" method has been proposed as an alternative to standard RILR. Even if the distributions of the mean-centered predictors are the same across centers (which would help to address violations of assumption (A1)), this method is not particularly compelling in the prediction setting since application of the model to new centers requires estimates of the center-specific biomarker means; such reliance on information from the new center to update the model makes external validation and clinical application more challenging. In addition, since the "poor man's" method relies on a RILR model, the estimate of the fixed intercept will potentially face the same challenges as the standard RILR model, discussed above. Thus, the "poor man's" method will often not provide useful predicted probabilities.

The idea behind the poor man's method is that subtracting the center-specific mean transforms the biomarkers into predictors that have the same distribution across centers (or are at least independent of b_c). Essentially, this is an attempt to force the model to estimate the within-center effect of the biomarkers, as opposed to a combination of the within- and between-center effects. However, FILR estimates this effect with no further assumptions or transformations of the data. It is important to estimate the within-center effect of the biomarkers in order to avoid allowing center to be predictive: in the presence of between-center differences in the biomarkers, where the between- and within-center effects of the biomarkers differ, RILR may allow center to be predictive since the fitted combination will in part reflect the between-center comparisons. Thus, the use of FILR in multicenter biomarker studies is compelling as estimates of biomarker associations that are unaffected by center differences are most useful in identifying promising combinations for further development.

Conversely, an obvious criticism of uFILR is that it does not allow predicted probabilities

to be estimated in new centers, since it only provides estimates of the intercepts for the centers used in construction. In addition, due to the form of the likelihood, cFILR cannot provide estimates of the center-specific intercepts for the centers used in construction, nor can it provide such estimates for new centers. We agree that these are limitations of FILR; however, as discussed above, RILR does not necessarily solve this problem. Furthermore, the biomarker combination can still be useful without intercept estimates, for example, to stratify patients within each center according to likelihood of having or developing the outcome. In addition, the center-adjusted performance of combinations (in terms of discrimination) can be evaluated without an intercept, allowing for identification of promising combinations of biomarkers for further development.

Evaluation: Center-Adjusted ROC and AUC

One way to account for center in the evaluation of a biomarker combination is via the center-adjusted AUC. As described earlier, the center-adjusted AUC (aAUC) is a stratified measure of performance (Janes and Pepe, 2008) and can be written as $aAUC(\boldsymbol{\theta}) = \sum_{c=1}^{M} AUC_c(\boldsymbol{\theta}) w_c$ for a given combination $\boldsymbol{\theta}^\top \mathbf{X}$. The empirical aAUC estimate can be written as

$$a\hat{AUC}(\boldsymbol{\theta}) = \sum_{c=1}^{m} \hat{w}_c \hat{AUC}_c(\boldsymbol{\theta}),$$

where $\hat{w}_c$ is the fraction of observed cases in center c and is the empirical estimate of the weight w_c, that is, $\hat{w}_c = \frac{n_D^c}{\sum_{c=1}^{m} n_D^c}$. If a study involves outcome-dependent sampling, care must be taken to ensure that $\hat{w}_c$ is a valid estimate the distribution of cases across centers. If $\hat{w}_c$ is not a valid estimate of $P(C = c|D = 1)$, $a\hat{AUC}(\boldsymbol{\theta})$ will still correspond to an adjusted measure of performance, but it will no longer estimate $aAUC(\boldsymbol{\theta})$ as defined above.

Asymptotic Properties

Our proposal involves constructing linear combinations of biomarkers by estimating $\boldsymbol{\theta}$ and evaluating the performance of these combinations with the aAUC. We would like to demonstrate consistency of this estimate of performance; that is, $a\hat{A}UC(\hat{\boldsymbol{\theta}}) \xrightarrow{p} aAUC(\boldsymbol{\theta}_0)$ if $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. This is shown by Lemmas 4.1 and 4.2 and Theorem 4.1 below; the proofs of these results can be found in Appendices A.2.3, A.2.3, and A.2.5. Lemma 4.1 provides uniform convergence in probability of $\hat{AUC}_c(\boldsymbol{\theta})$ to $AUC_c(\boldsymbol{\theta})$. Lemma 4.2 provides uniform convergence in probability of $a\hat{A}UC_c(\boldsymbol{\theta})$ to $aAUC_c(\boldsymbol{\theta})$ and Theorem 4.1 provides convergence in probability of $a\hat{A}UC_c(\hat{\boldsymbol{\theta}})$ to $aAUC_c(\boldsymbol{\theta}_0)$ if $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$. First, we describe the conditions required for these results to hold. Let $\mathbf{X}_i^c$ denote the vector of biomarkers for the i^{th} individual in center c and let $\mathbf{X}^c$ denote the vector of biomarkers for an arbitrary individual in center c.

(C1) The m centers are randomly sampled from the population of M centers, and n_c observations are randomly sampled from center c, $c = 1, ..., m$.

(C2) $\sum_{c=1}^{m} |E(\hat{w}_c) - w_c| \to 0$ as $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n_c}/m \to \infty$.

(C3) The centers are independent and within each center, the observations $O_i^c = (D_i^c, \mathbf{X}_i^c)$, $i = 1, ..., n_c$, are independent and identically distributed $(p+1)$-dimensional random vectors with distribution function F_c such that there exists at least one component of $\mathbf{X}^c$, X_k^c for some $k \in \{1, ..., p\}$, with distribution that has everywhere positive Lebesgue density, conditional on the other $\mathbf{X}^c$ components.

(C4) The support of $\mathbf{X}^c$, $c = 1, ..., M$, is not contained in any proper linear subspace of $\mathbb{R}^p$.

(C5) $AUC_c(\boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}_0$ and $||AUC_c'(\boldsymbol{\theta}_0)|| \leq T < \infty$, $c = 1, ..., m$.

(C6) $\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0 \in B = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1, |\theta_k| > 0\}$

Lemma 4.1. *Suppose conditions (C1), (C3), and (C4) hold for a given center c. Then* $\sup_{\boldsymbol{\theta}\in B}\left|\hat{AUC}_c(\boldsymbol{\theta}) - AUC_c(\boldsymbol{\theta})\right| = o_p(1)$ *as* $n_c \to \infty$, *where* $B = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1, |\theta_k| > 0\}$.

Lemma 4.2. *Suppose conditions (C1)-(C4) hold. Then for* $B = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}|| = 1, |\theta_k| > 0\}$,

$$\sup_{\boldsymbol{\theta}\in B}\left|a\hat{AUC}(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})\right| \xrightarrow{p} 0$$

and

$$\sum_{c=1}^{m} |\hat{w}_c - w_c| = o_p(1)$$

as $n_c \to \infty$, $c = 1, ..., m$, *and* $m \to M$ *such that* $\sqrt{n_c}/m \to \infty$.

Theorem 4.1. *Suppose* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ *as* $n_c \to \infty$, $c = 1, ..., m$, *and* $m \to M$ *such that* $\sqrt{n_c}/m \to \infty$. *Further suppose that conditions (C1)-(C6) hold. Then* $a\hat{AUC}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} aAUC(\boldsymbol{\theta}_0)$ *as* $n_c \to \infty$, $c = 1, ..., m$, *and* $m \to M$ *such that* $\sqrt{n_c}/m \to \infty$.

The proof of Lemma 4.1 relies heavily on the proof of a similar result for a statistic related to $\hat{AUC}_c$ given by Han (1987). The proof of Lemma 4.2 uses Lemma 4.1 to demonstrate the first claim and a Taylor approximation to $\hat{w}_c$ to demonstrate the second. Finally, the proof of Theorem 4.1 relies on Lemma 4.2 and a Taylor approximation.

In Lemma 4.1, Lemma 4.2, and Theorem 4.1, we restrict the combinations to have $||\boldsymbol{\theta}|| = 1$ for mathematical ease; since AUC_c is invariant to monotone transformations, this is not restrictive in a practical sense.

4.3.4 Combining Construction and Evaluation

When considering constructing and evaluating biomarker combinations, there are two binary decisions to make with regard to center, giving four possibilities (using the notation of models (4.6) and (4.8)):

1. Pooling the data across centers for both construction and evaluation, giving $AUC(\boldsymbol{\alpha})$
2. Pooling the data across centers for construction, but stratifying by center for evaluation, giving $AUC_c(\boldsymbol{\alpha})$

3. Stratifying by center for construction, but pooling across centers for evaluation, giving $AUC(\boldsymbol{\beta})$

4. Stratifying by center for both construction and evaluation, giving $AUC_c(\boldsymbol{\beta})$

In the discussion below we consider only the true parameter values, rather than the estimates. We will address small-sample variability via simulations.

The two results given below follow directly from Pepe (2003). In particular, they show that the marginal and center-adjusted AUCs of a combination based on some $\boldsymbol{\theta}$ are equivalent if $C \perp L_{\boldsymbol{\theta}}(\mathbf{X})|\bar{D}$. Note that $C \perp \mathbf{X}|D$ implies $C \perp L_{\boldsymbol{\theta}}(\mathbf{X})|\bar{D}$. Thus, if model (4.6) holds and $C \perp \mathbf{X}|D$, then $AUC(\boldsymbol{\beta}) = aAUC(\boldsymbol{\beta}) = aAUC(\boldsymbol{\alpha}) = AUC(\boldsymbol{\alpha})$, since $\boldsymbol{\alpha} = \boldsymbol{\beta}$ by collapsibility.

Proposition 4.1. *Suppose $C \perp L_{\boldsymbol{\theta}}(\boldsymbol{X})|\bar{D}$ for a combination based on some $\boldsymbol{\theta}$ such that $L_{\boldsymbol{\theta}}(\boldsymbol{X})$ has common support among cases and controls. Then $AUC(\boldsymbol{\theta}) = aAUC(\boldsymbol{\theta})$.*

This proposition is presented without proof; the result follows directly from Result 6.2 in ((Pepe, 2003)).

When the prevalence and center-specific AUC do not vary with center and the center-specific ROC curves are concave, the center-specific AUC for a given biomarker combination will be at least as large as the marginal AUC. This is given by Proposition 4.2, stated below. In general, the center-specific ROC curves will be concave if, in each center, increasing $L_{\boldsymbol{\theta}}(\mathbf{X})$ increases the likelihood that $D = 1$ (Copas and Corbett, 2002).

Proposition 4.2. *Suppose $C \perp D$ and for a combination based on some $\boldsymbol{\theta}$, $AUC_c(\boldsymbol{\theta}) = AUC^*(\boldsymbol{\theta}), \quad c = 1, ..., M$ and $ROC_c(\boldsymbol{\theta})$ is concave for $c = 1, ..., M$. Then $AUC_c(\boldsymbol{\theta}) \geq AUC(\boldsymbol{\theta})$, $c = 1, ..., M$.*

This proposition is presented without proof, as the result follows directly from Result 6.1 in Pepe (2003), but here we have a linear combination $L_{\boldsymbol{\theta}}(\mathbf{X})$ instead of a single marker.

When model (4.6) holds, optimality of the risk score $P(D = 1|\mathbf{X}, C)$ implies that the combination based on $\boldsymbol{\beta}$ is optimal within each center, in terms of maximizing center-specific

AUC (Pepe, 2003). Thus, under this model,

$$AUC_c(\boldsymbol{\beta}) \geq AUC_c(\boldsymbol{\theta}),$$

for any $\boldsymbol{\theta}$. Furthermore, by the collapsibility results discussed above, when model (4.6) holds and $C \perp D|\mathbf{X}$, $\boldsymbol{\alpha} = \boldsymbol{\beta}$, giving

$$AUC(\boldsymbol{\alpha}) = AUC(\boldsymbol{\beta})$$
$$AUC_c(\boldsymbol{\alpha}) = AUC_c(\boldsymbol{\beta}).$$

Thus, ignoring center may not affect the results in some special cases. Outside of these, however, ignoring center can give misleading results or yield biomarker combinations with diminished performance.

4.3.5 Identifying the Role of Center

When biomarker data from multiple centers are available, graphical displays and other data summaries may be useful for identifying whether center is a case mix variable, calibration variable, or confounder. We therefore propose investigating:

1. The distribution of the biomarkers across center (to assess whether $C \perp \mathbf{X}$)
2. The distribution of the biomarkers across center stratified by the outcome (to assess whether $C \perp \mathbf{X}|D$)
3. The prevalence of the outcome across centers (to assess whether $C \perp D$)
4. The prevalence of the outcome across centers, stratified by biomarker categories (to assess whether $C \perp D|\mathbf{X}$)

Not all of these tools will be useful in all settings; for example, in the case of diagnostic biomarkers, items (2) and (3) will be most useful, while items (1) and (4) will be most

useful for prognostic markers. Also, implementing item (4) may be challenging as it requires the designation of biomarker categories, which necessarily results in a loss of information. Alternatively, this relationship is could be assessed with a regression model. Some of these suggestions are similar in spirit to those made by Berlin et al. (1999) in the case of a single binary predictor.

It is important to keep in mind that there are no "rules" for interpreting the results of the data summaries we have proposed above. Rather, they should be used as a guide, in conjunction with knowledge about the design and conduct of the study, to assess the role of center.

4.4 Simulations

4.4.1 Ignoring Center

We considered the impact of ignoring center in the construction and/or evaluation of biomarker combinations. We considered diagnostic markers, and allowed center to be a case mix variable, a calibration variable, or a confounder. The two biomarkers X_1 and X_2 were distributed as described in equation (4.7) with $\rho = 0.5$, and $AUC_{X_1|C} = 0.6$ and $AUC_{X_2|C} = 0.65$ in all centers.

Throughout, the center-specific mean offsets in the biomarkers were equal; that is, $f_{X_1}(c) = f_{X_2}(c) = f(c)$. When center was a case mix variable, $\text{logit}(\gamma_c) \sim N(0, \sigma^2_{\gamma_c})$ and $f(c) = 0$. When center was a calibration variable, $\gamma_c = 0.5$ and $f(c) \sim N(0, \sigma^2_{f(c)})$. Finally, when center was a confounder,

$$\begin{pmatrix} \text{logit}(\gamma_c) \\ f(c) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\gamma_c} & \delta\sigma_{\gamma_c}\sigma_{f(c)} \\ \delta\sigma_{\gamma_c}\sigma_{f(c)} & \sigma^2_{f(c)} \end{pmatrix} \right)$$

We considered $\sigma^2_{\gamma_c} = 1$, $\sigma^2_{f(c)} = 5$, and $\delta \in \{-0.75, 0.75\}$.

The combinations were constructed in a training dataset consisting of either 6 centers with 200 observations each or 500 centers with 20 observations each. The first scenario is

intended to be representative of a large cohort study similar to the TRIBE-AKI study, while the second scenario is intended to be representative of a study of small clinics or individual physicians. The combinations were constructed in the training data via logistic regression, where center was either ignored or incorporated using uFILR (in the case of 6 centers) or cFILR (in the case of 500 centers). These estimates correspond to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as defined in equations (4.8) and (4.6), respectively, and were used to construct two combinations:

$$L_{\hat{\boldsymbol{\beta}}}(X_1, X_2) = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2,$$
$$L_{\hat{\boldsymbol{\alpha}}}(X_1, X_2) = \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2.$$

We evaluated the fitted combinations via the conditional AUC, $AUC_c(\cdot)$, in a large test dataset with a single center, and the marginal AUC, $AUC(\cdot)$, in a large test dataset with multiple centers. Because the conditional AUC is constant across centers under our data-generating model (see Appendix A.2.2), $AUC_c(\cdot) = aAUC(\cdot)$. The test set used to evaluate the conditional AUC consisted of a single center with 200,000 observations while the test set used to evaluate the marginal AUC included either 6 centers with 30,000 observations each or 500 centers with 400 observations each, depending on the structure of the training data. The observations in the test data represent subjects from new centers, i.e., not the same centers as used in the training data. The coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $AUC_c(\boldsymbol{\beta})$ were determined analytically for comparison. The simulations were repeated 500 times.

Figure 4.3 presents the results of the simulations with 6 centers. These simulations support the results presented above: that is, when center is a case mix variable, the AUC is not affected by ignoring center in construction and/or evaluation. Likewise, the simulation results when center is a calibration variable support the conclusions given above, that is, $AUC_c(\hat{\boldsymbol{\beta}}) \geq AUC_c(\hat{\boldsymbol{\alpha}})$, $AUC_c(\hat{\boldsymbol{\beta}}) \geq AUC(\hat{\boldsymbol{\beta}})$ and $AUC_c(\hat{\boldsymbol{\alpha}}) \geq AUC(\hat{\boldsymbol{\alpha}})$. Thus, when center is a calibration variable, ignoring center during construction can lead to a biomarker combination with reduced predictive capacity in new centers, and ignoring center during evaluation yields a measure of performance that is lower than the performance of the combination in a

new center.

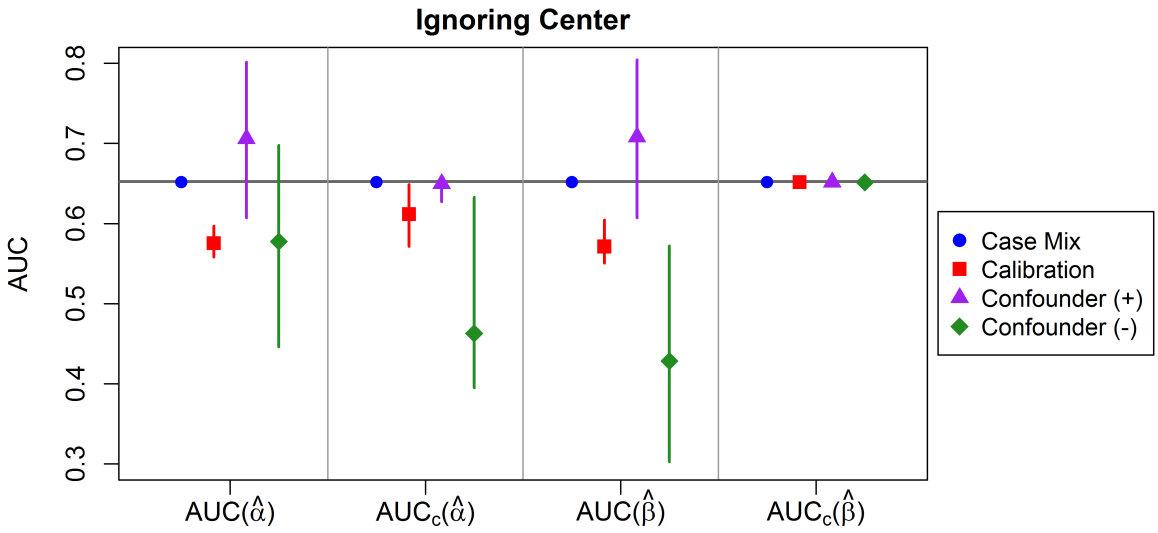

Figure 4.3: Simulation results for training data with 6 centers. The first column is the marginal AUC based on the combination constructed by ignoring center, $AUC(\hat{\boldsymbol{\alpha}})$, the second column is the conditional AUC based on the combination constructed by ignoring center, $AUC_c(\hat{\boldsymbol{\alpha}})$, the third column is the marginal AUC based on the combination constructed by stratifying by center, $AUC(\hat{\boldsymbol{\beta}})$, and the fourth column is the conditional AUC based on the combination constructed by stratifying by center, $AUC_c(\hat{\boldsymbol{\beta}})$. For each, the median and middle 90% of the distribution across simulations are shown. Different colors and shapes correspond to different roles for center: blue circles indicate center is a case mix variable, red squares indicate center is a calibration variable, purple triangles indicate center is a confounder with positive correlation (0.75) between logit(γ_c) and $f(c)$, and green diamonds indicate center is a confounder with negative correlation (-0.75) between logit(γ_c) and $f(c)$. The gray horizontal line represents $AUC_c(\boldsymbol{\beta})$ as determined analytically.

When center is a confounder and logit(γ_c) and $f(c)$ are positively correlated, ignoring center during evaluation yields measures of performance that are somewhat higher than the performance of the combination in a new center ($AUC(\hat{\boldsymbol{\alpha}})$ versus $AUC_c(\hat{\boldsymbol{\alpha}})$ and $AUC(\hat{\boldsymbol{\beta}})$ versus $AUC_c(\hat{\boldsymbol{\beta}})$). When center is a confounder and logit(γ_c) and $f(c)$ are negatively correlated, ignoring center during evaluation yields a measure of performance that is higher or lower than the performance of the combination in a new center, depending upon whether

center was also ignored in the construction of the combination. In particular, if center is ignored during construction (yielding $\hat{\boldsymbol{\alpha}}$), then ignoring center during evaluation tends to give a measure of performance that is higher than the actual performance in a new center (i.e., $AUC(\hat{\boldsymbol{\alpha}})$ is generally larger than $AUC_c(\hat{\boldsymbol{\alpha}})$). On the other hand, if center is included in construction (yielding $\hat{\boldsymbol{\beta}}$), ignoring center during evaluation tends to give a measure of performance that is lower than the performance of the combination in a new center; that is, $AUC(\hat{\boldsymbol{\beta}})$ tends to be smaller than than $AUC_c(\hat{\boldsymbol{\beta}})$. As expected, regardless of the correlation between logit(γ_c) and $f(c)$, ignoring center during construction generally results in a combination with worse performance in new centers than if center were included in construction ($AUC_c(\hat{\boldsymbol{\alpha}})$ vs. $AUC_c(\hat{\boldsymbol{\beta}})$).

These results indicate that when center is a calibration variable or a confounder, ignoring center during construction could lead to a combination with worse performance in a new center ($AUC_c(\hat{\boldsymbol{\alpha}})$ vs. $AUC_c(\hat{\boldsymbol{\beta}})$) and ignoring center during evaluation could yield a measure of performance that does not reflect the performance of the combination in a new center ($AUC(\cdot)$ vs. $AUC_c(\cdot)$). The results were similar when the training data included 500 centers (Figure 4.4). The full results are given in Appendix B.3.1. Notably, those results indicate that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are quite different when center is a calibration variable or a confounder, as would be expected given the results presented above.

4.4.2 Including Center

We conducted simulations to compare combinations constructed by RILR to those constructed by FILR. Both methods were used to construct linear combinations. The set-up of these simulations is similar to those conducted above. We again consider two diagnostic markers X_1 and X_2, and allow center to be a case mix variable, a calibration variable, or a confounder. Thus, $C \not\perp \mathbf{X}$ for all three scenarios. The two biomarkers were distributed as described by equation (4.7) with $\rho = 0.5$ and $AUC_{X_1|C} = 0.6$ and $AUC_{X_2|C} = 0.65$ in all centers.

As before, the center-specific mean offsets in the biomarkers were equal; that is, $f_{X_1}(c) =$

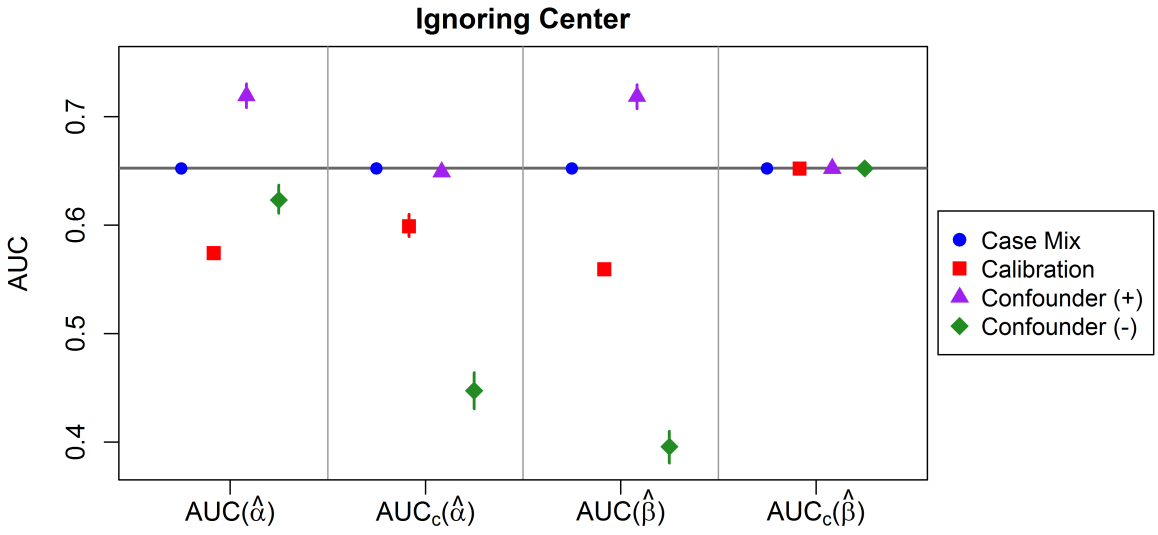

Figure 4.4: Simulation results for training data with 500 centers. The first column is the marginal AUC based on the combination constructed by ignoring center, $AUC(\hat{\boldsymbol{\alpha}})$, the second column is the conditional AUC based on the combination constructed by ignoring center, $AUC_c(\hat{\boldsymbol{\alpha}})$, the third column is the marginal AUC based on the combination constructed by stratifying by center, $AUC(\hat{\boldsymbol{\beta}})$, and the fourth column is the conditional AUC based on the combination constructed by stratifying by center, $AUC_c(\hat{\boldsymbol{\beta}})$. For each, the median and middle 90% of the distribution across simulations are shown. Different colors and shapes correspond to different roles for center: blue circles indicate center is a case mix variable, red squares indicate center is a calibration variable, purple triangles indicate center is a confounder with positive correlation (0.75) between logit(γ_c), and $f(c)$ and green diamonds indicate center is a confounder with negative correlation (-0.75) between logit(γ_c) and $f(c)$. The gray horizontal line represents $AUC_c(\boldsymbol{\beta})$ as determined analytically.

$f_{X_2}(c) = f(c)$. When center was a case mix variable, logit(γ_c) $\sim F$ with mean 0 and variance $\sigma^2_{\gamma_c}$ and $f(c) = 0$. When center was a calibration variable, $\gamma_c = 0.5$ or 0.1 and $f(c) \sim F$ with mean 0 and variance $\sigma^2_{f(c)}$. Finally, when center was a confounder, logit(γ_c) $\sim F$ with mean 0 and variance $\sigma^2_{\gamma_c}$, $f(c) \sim F$ with mean 0 and variance $\sigma^2_{f(c)}$, and Corr(logit(γ_c), $f(c)$) $= \delta$. We varied F, $\sigma^2_{\gamma_c}$, $\sigma^2_{f(c)}$ and δ as described in Table 4.1.

The combinations were constructed in training data and evaluated in a large test dataset. For the training data, two scenarios were considered: 6 centers with 200 observations each

Table 4.1: Simulation parameters.

Simulation Parameter	*Scenarios Considered*
F	Normal, Gumbel, Laplace, Uniform
$\sigma^2_{\gamma_c}$	0.5, 1, 3, (0.5, 1.5), (1, 5)
$\sigma^2_{f(c)}$	1, 5, (2, 8)
δ	-0.5, 0, 0.5

and 500 centers with 20 observations each. The combinations were constructed in the training data via logistic regression, where center was either (i) incorporated using RILR assuming $b_c \sim N(0, \sigma^2)$ or (ii) incorporated using uFILR (in the case of 6 centers) or cFILR (in the case of 500 centers). The fitted biomarker combinations based on these models were evaluated in the test dataset, which consisted of a single new center with 10,000 observations; that is, in all scenarios, we evaluated $AUC_c(\cdot)$. In large samples, this is expected to equal the center-adjusted AUC, $aAUC(\cdot)$, as our simulation set-up yields constant center-specific AUCs.

We considered some settings where the variances of logit(γ_c) and/or $f(c)$ were not constant (those pairs of values in parentheses in Table 4.1); in these scenarios, half of the centers were assigned one value of $\sigma^2_{\gamma_c}$ (or $\sigma^2_{f(c)}$) and the remainder were assigned the other (the single center in the test data was assigned the lower of the two values). When center was a calibration variable, $\gamma_c = 0.5$ in most simulations. To study the impact of concordant centers, we also considered simulations where center was a calibration variable and $\gamma_c = 0.1$. This was also the motivation for including $\sigma^2_{\gamma_c} = 3$ and $\sigma^2_{\gamma_c} = (1,5)$.

The true coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $AUC_c(\boldsymbol{\beta})$ were determined analytically. These simulations were repeated 500 times. In Figure 4.5, we present the results for $m = 500$ centers with $F =$ Normal, $\sigma^2_{\gamma_c} = 1$, $\sigma^2_{f(c)} = 5$, prevalence of 0.5 when center was a calibration variable, and $\delta = -0.5$ when center was a confounder. In all scenarios, the results from FILR are close to the truth. The differences in the parameter estimates when RILR is used are clear, particularly when center is a calibration variable or a confounder. This leads to substantially different conditional AUC values, particularly when center is a calibration

variable. The differences in AUC are small when center is a case mix variable; in this setting, the differences in the coefficient estimates are not as large, and the AUC, which is a rank-based measure, can overcome these more modest perturbations. Furthermore, the difference in AUC is much larger when center is a calibration variable than when it is a confounder (the same is true of the difference in the coefficient estimates). The full results are given in Appendix B.3.2. In general, we see that the differences between RILR and FILR tend to be smaller when there are fewer centers ($m = 6$), $\sigma^2_{f(c)}$ is small, or $\sigma^2_{\gamma_c}$ is large.

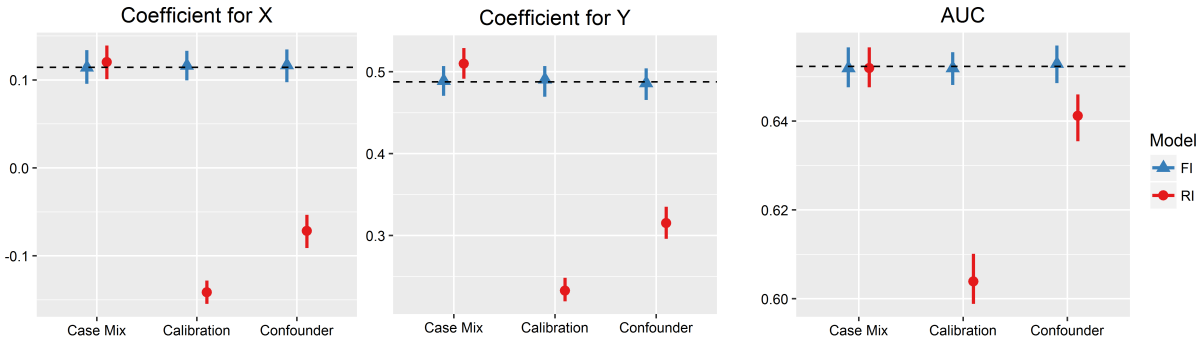

Figure 4.5: Simulation results comparing random and fixed intercept logistic regression for $m = 500$ in the training data, where F = Normal, $\sigma^2_{\gamma_c} = 1$, $\sigma^2_{f(c)} = 5$, $\gamma_c = 0.5$ when center was a calibration variable, and $\delta = -0.5$ when center was a confounder. The median and interquartile ranges across the 500 simulations are reported. The columns in each plot correspond to different roles for center. The results based on FILR are displayed as blue triangles and the results based on RILR are displayed as red circles. The results for the biomarker coefficients are shown in the first two plots, and the results for the (conditional) AUC are shown in the third plot. In each plot, the dashed horizontal line indicates the true value.

The improved performance of FILR persisted even when we considered situations where there were 500 centers and, on average, 7-12% were concordant (induced via large variability in logit(γ_c) when center was a case mix variable or a confounder, or low γ_c when center was a calibration variable). This is supported by the results given in Appendix B.3.2, which show almost no benefit for RILR, even in the presence of concordant centers. In simulations not designed specifically to have high concordance, up to 2% of centers were concordant, on

average.

Finally, we evaluated the degree of bias in the estimate of the overall fixed intercept provided by RILR (i.e., τ_0 in equation (4.9)) that would need to be used to generate predicted probabilities. We found that in many scenarios, this estimate was substantially biased, i.e., absolute biases of more than 20% (Appendix B.3.2).

4.5 Application to the TRIBE-AKI Study

We applied the methods we have discussed to data from the TRIBE-AKI study. Recall that this is a study of 1219 adults undergoing cardiac surgery, and there is interest in using biomarkers to provide an earlier diagnosis of post-operative AKI. We consider three biomarkers: urine NGAL, h-FABP, and plasma TNI. After removing observations with missing values for any of these biomarkers, 962 observations remained. The three biomarkers were log-transformed and we considered the measurements taken immediately after surgery. Thus, we are in the setting of diagnostic biomarkers, since the biomarkers are measured post-operatively and the injury to the kidney occurs during surgery; as a result, these biomarkers are thought to reflect the underlying disease (AKI) process.

First we consider the role of center in this study. Since we are considering diagnostic biomarkers, we evaluated the distribution of the biomarkers in each center among AKI controls and the prevalence of AKI across centers. The biomarker distributions are given in Figure 4.6. There is evidence that the distribution of the biomarker measurements varies across centers among controls. We also see in Table 4.2 that the prevalence of AKI varies quite substantially across centers. Thus, there is evidence that center is a confounder in this study.

In light of these findings, we sought to construct a diagnostic biomarker combination while accounting for center by fitting a FILR model. We evaluated this combination by estimating the center-adjusted AUC. We then corrected this estimate for resubstitution bias by bootstrapping the individual observations to quantify the degree of optimistic bias. This provides an honest assessment of the performance of the combination in the centers used in

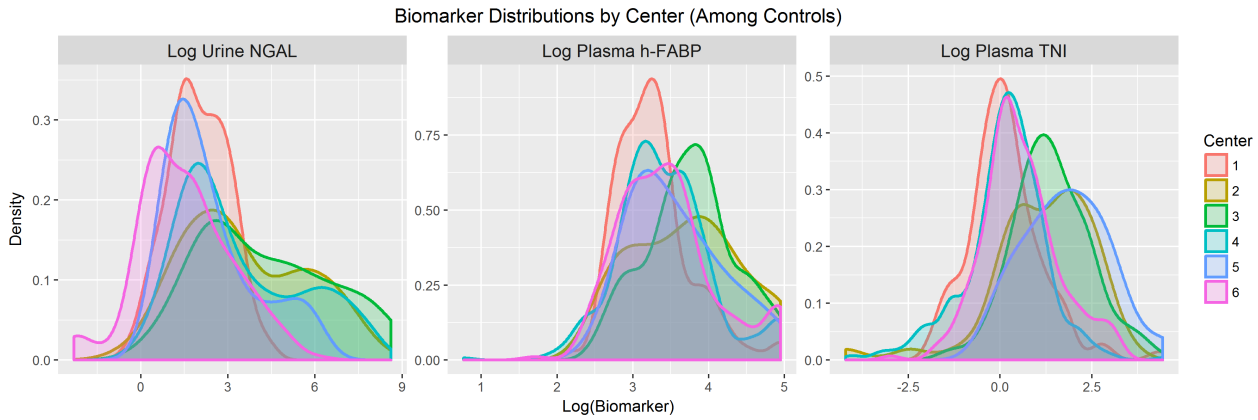

Figure 4.6: Distribution of log urine NGAL, log plasma h-FABP, and log plasma TNI in the TRIBE-AKI study among controls. The biomarker distributions are stratified by center.

Table 4.2: Center-specific AKI prevalence in the TRIBE-AKI study.

Center (n)	*AKI Prevalence (95% CI)*
1 (103)	7.8% (3.4%-14.7%)
2 (53)	17.0% (8.1%-29.8%)
3 (70)	22.9% (13.7%-34.4%)
4 (483)	19.5% (16.0%-23.3%)
5 (27)	22.2% (8.6%-42.3%)
6 (226)	11.1% (7.3%-15.9%)

the TRIBE-AKI study. We compared the combination fitted by FILR to the combinations fitted by RILR and by marginalizing over (i.e., ignoring) center.

The biomarker combination estimated by FILR was

$$0.025 * \log(\text{NGAL}) + 1.103 * \log(\text{h-FABP}) - 0.065 * \log(\text{TNI}).$$

The optimism-corrected center-adjusted AUC for this combination was 0.6823. In contrast, the combination estimated by RILR was

$$0.054 * \log(\text{NGAL}) + 1.096 * \log(\text{h-FABP}) - 0.065 * \log(\text{TNI}).$$

The optimism-corrected center-adjusted AUC for this combination was 0.6806. When center was ignored during construction, the estimated combination was

$$0.081 * \log(\text{NGAL}) + 1.103 * \log(\text{h-FABP}) - 0.094 * \log(\text{TNI}),$$

and the optimism-corrected center-adjusted AUC for this combination was 0.6811.

Thus, in these data, the three fitted combinations were quite similar, and, correspondingly, the gains offered by FILR in terms of the center-adjusted AUC were very modest. The estimate of the adjusted AUC was driven in large part by center 4, which had 59% of the AKI cases. There is some indication of effect modification of associations between the biomarkers and AKI across center: the center-specific coefficient estimates ranged between -0.235 and 0.797 for NGAL, -0.013 and 1.382 for h-FABP, and -0.430 and 0.448 for TNI. The three fitted combinations presented above are essentially weighted averages of these center-specific combinations. For the combination estimated by FILR the center-specific AUCs (unadjusted for optimism) were between 0.639 and 0.716, while for the combination estimated by RILR, they were between 0.628 and 0.716, and when center was ignored, they ranged between 0.617 and 0.729.

4.6 Discussion

We have created a unified framework for constructing and evaluating biomarker combinations in multicenter studies, including a taxonomy to differentiate several roles center might play, tools for identifying the role of center, and methods for constructing biomarker combinations and evaluating their performance. Essentially, by conditioning on center in both the construction and evaluation of biomarker combinations, we obtain combinations and measures of performance that are unaffected by center differences; given that these center differences are often not scientifically relevant and are expected to vary in magnitude from center to center, using conditional approaches to construction and evaluation of biomarker combinations is more likely to yield useful combinations. The concepts and methods we des-

cribe apply to biomarker combinations, and also to combinations of biomarkers and clinical or demographic variables.

As discussed above, an important limitation of using FILR to construct a biomarker combination is that an estimate of the center-specific intercept in new centers is not available. If we do not have an estimate of the center-specific intercept, it is not possible to generate predicted probabilities; in such situations, the biomarker combination is a tool for risk stratification within each center rather than a risk prediction model.

It may be that the center-specific AUC is not the same across centers; in this situation, it is generally informative to evaluate the variability in the center-specific AUCs across center, as we did in the illustration with data from the TRIBE-AKI study. This may provide some indication of how the biomarker combination might be expected to perform in a new center, if the centers included in the evaluation are "similar" to the new centers. However, when evaluating the center-specific AUCs, it is important to keep in mind that AUC estimates from centers with fewer observations may be less reliable.

In general, the performance of a given biomarker combination could differ in a new center due to effect modification of the performance of the biomarker combination by center and/or effect modification of the association of the biomarkers with the outcome by center. Since we are interested in developing a single biomarker combination, when there is effect modification of the combination itself (via effect modification of the biomarker associations), we estimate a weighted averaged of the center-specific combinations. In the event that the center-specific combinations differ and are each optimal in terms of the center-specific AUC, the resulting weighted-average combination is unlikely to be optimal in terms of the center-specific AUC in each center. Additionally, if the center-specific combinations vary but have the same center-specific AUC, the performance of the weighted-average combination may vary across centers. Thus, while the approach of estimating a single biomarker combination is still reasonable in the presence of effect modification of the biomarker associations since our goal is to provide a single fitted combination, doing so could affect the performance of the weighted-average combination, the degree of variability in center-specific performance of the weighted-average

combination, and, potentially, generalizability (if the centers used for construction are not representative of the centers to which we wish to apply the combination).

As we have noted, if the center-specific AUC varies, different sampling schemes could affect the estimated weights $\hat{w}_c$, which could in turn affect the the estimated center-adjusted AUC. The center-specific AUC itself is unaffected by case-control sampling within each center (Pepe et al., 2006) and the center-adjusted AUC is unaffected by center-dependent sampling among controls (Janes and Pepe, 2009), though the asymptotic results we have provided may be affected by certain sampling schemes. If a study involves matching, care must be taken to adjust the AUC for the matching as well as for center (Janes and Pepe, 2008).

Multicenter biomarker studies will continue to grow in popularity, and the possibility of using these data to construct combinations of prognostic or diagnostic biomarkers will be appealing to many investigators. However, as we have demonstrated, using inappropriate methods to construct and/or evaluate biomarker combinations can provide unfavorable or misleading results. These methods include random intercept logistic regression and ignoring center entirely; such approaches will only provide useful results in special cases. In particular, when biomarker distributions vary by center, as will often happen, there is little reason to expect random intercept logistic regression to provide useful biomarker combinations, that is, fitted combinations not influenced by center differences. Constructing combinations of center-specific placement values may be a viable approach. Placement values constitute a transformation of the data, wherein for each observation, the value of each biomarker is replaced by the proportion of controls in the same center that have a larger value of the biomarker (Pepe, 2003). Of course, this requires knowing the distribution of each biomarker among controls in each center, making application of the combination to a new center challenging.

Future research will include methods for other performance measures, including the true positive rate for a specific false positive rate and the partial AUC for a range of false positive rates. It will also be important to consider approaches that do not rely on empirical estima-

tes of the AUC, perhaps by modeling the combination parametrically (e.g., using a model to relate the combination to center among controls); such an approach may be beneficial when there are a large number of very small centers, as might happen when the "centers" are clinicians. In these settings, the empirical AUC estimate may be unreliable, and an alternative estimate may be preferable.

An important contribution of this work is that it demonstrates that methods often applied to multicenter data are generally not appropriate. Biomarkers hold great potential in the risk prediction setting, but have for the most part been relatively disappointing thus far as very few have been adopted clinically. Much of the problem has been blamed on "validation failures"; that is, biomarkers that are found to be quite promising initially, but are never used in clinical practice due to disappointing results in follow-up studies (Ioannidis, 2013). Thus, to the extent possible, it is important to recognize aspects of study design, conduct, and analysis that require special attention when developing biomarker combinations for diagnosis and prognosis. Carefully addressing these issues can increase the likelihood of identifying clinically useful combinations, leading to better patient care.

Chapter 5

DEVELOPING BIOMARKER COMBINATIONS IN MULTICENTER STUDIES VIA DIRECT MAXIMIZATION AND PENALIZATION

Abstract

When biomarker studies involve patients at multiple centers and the goal is to develop biomarker combinations for diagnosis, prognosis, or screening, the predictive capacity of a given combination is typically evaluated by estimating 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 incorporate a penalty for variability in center-specific performance. We demonstrate good theoretical 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. Simulated datasets also 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.

5.1 Introduction

Multicenter studies, where centers could be hospitals, clinics, or providers, have long been used in the therapeutic setting as a way to increase power and improve generalizability, and

are increasingly common in studies of biomarkers (e.g., Degos et al. (2010); Feldstein et al. (2009); Nickolas et al. (2012)). Additionally, it is now feasible to measure many biomarkers on each participant. As the individual performance of these biomarkers is often modest, there is interest in using 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). The participants 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 be used to 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 generally interest in summaries of 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. In addition, our method can be used to construct biomarker combinations with good overall performance and more homogeneous performance across centers by maximizing the aAUC while penalizing variability in center-specific performance.

5.2 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 $\bar{D}$).

5.2.1 Center-adjusted AUC

The ROC curve for a biomarker or biomarker combination 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 biomarkers and biomarker combinations is often summarized via the AUC, a measure of the ability of Z to discriminate between cases and controls. The ROC curve for a useless biomarker or combination lies on the 45-degree line, and the corresponding AUC is 0.5 (Pepe, 2003). The ROC curve for a perfect biomarker or combination 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, assuming that higher values of Z are more indicative of D (Pepe, 2003). The AUC is invariant with respect to monotone transformations of Z (Pepe, 2003).

In the multicenter setting, a biomarker combination 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 marginal measure of performance (i.e., the AUC for Z pooled across centers), we are potentially allowing center to be predictive, severely restricting interpretability and generalizability (Janes and Pepe, 2008). Instead, the 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 (Janes and Pepe, 2008; Kahan, 2014). One such summary measure is the center-adjusted AUC (aAUC).

The center-adjusted ROC (aROC) and corresponding aAUC, proposed by Janes and Pepe (2009), can be written as

$$\begin{aligned} aAUC_Z &= \int_0^1 aROC_Z(t)dt = \int_0^1 \int ROC_{Z|C=c}(t)dP_D(c)dt \\ &= \sum_c AUC_{Z|C=c}P(C=c|D=1) = \sum_c w_c AUC_{Z|C=c}, \end{aligned} \tag{5.1}$$

where C indicates center, $ROC_{Z|C=c}$ is the center-specific ROC curve, $AUC_{Z|C=c}$ is the center-specific AUC, t is the false positive rate, and $P_D(c)$ is the distribution of center among cases. When the center-specific AUCs, $AUC_{Z|C=c}$, are constant across centers, the adjusted AUC is simply that center-specific AUC (Janes et al., 2009). More generally, the aAUC is a weighted average of the center-specific AUCs where the center-specific AUCs are weighted by the proportion of cases in each center (Janes et al., 2009). Weighting by the proportion of cases is appealing because centers with more cases generally 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). For a given biomarker combination, the aAUC provides an estimate of the performance of the biomarker combination in new centers, to the extent that the new centers are similar to those used to evaluate the combination.

The expression for the aAUC given in equation (5.1) corresponds to the area under the ROC curve given by the weighted average of the true positive rates in each center, holding the false positive rates in each center constant (Janes and Pepe, 2009). In other words, the aROC curve corresponds to using center-specific thresholds, chosen such that the false positive rate is constant across centers, to determine the true positive rate (Janes et al., 2009; Janes and Pepe, 2009):

$$\begin{aligned} aROC_Z(t) &= \sum_c ROC_{Z|C=c}(t) P(C = c|D = 1) \\ &= P\Big(Z > g(t|c)\Big|D = 1\Big), \end{aligned}$$

where $g(t|c)$ gives a false positive rate of t in center c. The adjusted ROC curve defined here represents one way of combining the center-specific ROC curves, that is, by averaging the curves vertically; these curves could be combined in other ways (Janes and Pepe, 2009).

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). That is, if the same combination were applied

to independent data, the performance is expected to be diminished. 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 optimism and correct the apparent performance estimate (Copas and Corbett, 2002; Harrell, 2013). 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, as opposed to resampling observations, has been suggested (Bouwmeester et al., 2013; Janes et al., 2009; Localio et al., 2001; van Oirbeek and Lesaffre, 2010). However, similar results in terms of bias have been found for the average of cluster-specific AUC estimates (where in our case, 'cluster' is center) whether resampling was done on clusters or individual observations (Bouwmeester et al., 2013).

5.2.2 Biomarker Combinations

Many biomarker assays are now relatively affordable and/or can be used to measure multiple biomarkers at once. This has increased investigators' ability to measure many biomarkers in each individual, leading to growing interest in developing biomarker combinations for diagnosis, prognosis, or screening. We will consider linear biomarker combinations, as they are often a reasonable choice and have intuitive appeal for clinical collaborators.

For a collection of biomarkers $\mathbf{X}$, the risk score, $P(D = 1|\mathbf{X})$, 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 optimize the AUC among linear combinations of biomarkers 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 et al., 2004). However, implementation of the method proposed by Liu and Zhou (2013) is computationally challenging with more than two biomarkers. The method proposed by Schisterman et al.

(2004) assumes that the biomarkers have a multivariate normal distribution and requires specification of the relationship between the covariates (i.e., center) and the biomarkers.

5.2.3 Smooth AUC Approximations

As alluded to above, an appealing alternative to logistic regression is to construct biomarker combinations by directly maximizing the AUC. Methods for fitting logistic regression models use the logistic likelihood as the objective function. However, we are interested in using fitted combinations for diagnosis, prognosis, or screening (Pepe et al., 2006; Pepe and Thompson, 2000), which motivates maximizing measures of predictive capacity, i.e., matching the objective function to the measure of interest. A benefit of directly maximizing the AUC is that the resulting combination is reasonable regardless of whether the linear logistic model holds (Pepe et al., 2006). Furthermore, the AUC of a combination constructed by targeting the AUC will be at least as large as the AUC for the individual biomarkers; this may not be true when the combination is constructed by optimizing another objective function (Pepe and Thompson, 2000).

For any given vector of coefficients $\boldsymbol{\theta}$, since the true AUC is unknown, we are limited to maximizing an estimate of the AUC. The empirical AUC,

$$\hat{AUC}(\boldsymbol{\theta}) = \frac{1}{n_D n_{\bar{D}}} \sum_{i:D_i=1, j:D_j=0} 1(\boldsymbol{\theta}^\top \mathbf{X}_i > \boldsymbol{\theta}^\top \mathbf{X}_j),$$

where $\mathbf{X}_i$ denotes the biomarker vector for the i^{th} case and $\mathbf{X}_j$ denotes the biomarker vector for the j^{th} control, involves indicator functions, making direct maximization challenging. However, smooth approximations to the empirical AUC estimate have been proposed. Lin et al. (2011) used the probit approximation to estimate $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} R_n(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{1}{n_D n_{\bar{D}}} \sum_{i:D_i=1, j:D_j=0} \Phi\left\{\boldsymbol{\theta}^\top (\mathbf{X}_i - \mathbf{X}_j)/h\right\} \right],$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}||_2 = 1\}$, 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 (Lin et al., 2011). The tuning parameter h is actually a sequence of numbers such that $\lim_{n\to\infty} h = 0$ (Lin et al., 2011). Lin et al. (2011) note 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 $\tilde{\boldsymbol{\theta}}$. The constraint $||\boldsymbol{\theta}||_2 = 1$ in $\boldsymbol{\Theta}$ is required for identifiability. Other identifiability constraints have been proposed. Ma and Huang (2007) suggested fixing the coefficient for the first parameter, though Lin et al. (2011) point out that, in practice, it is difficult to know whether the first biomarker has a non-zero coefficient. Lin et al. (2011) suggest constraining $||\boldsymbol{\theta}||_1 = 1$, while Fong et al. (2016) argue that instead constraining $||\boldsymbol{\theta}||_2 = 1$ allows for a more uniformly accurate approximation to the AUC across the parameter space.

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.

5.3 Methods

Suppose we have a p-dimensional biomarker vector $\mathbf{X}$, a binary outcome D, and data from m centers with n_c $(c = 1, ..., m)$ observations in each, where the total sample size is $n = \sum_{c=1}^{m} n_c$. In each center there are n_D^c cases and $n_{\bar{D}}^c$ controls $(n_c = n_D^c + n_{\bar{D}}^c)$. There are n_D total cases and $n_{\bar{D}}$ total controls $(n_D = \sum_{c=1}^{m} n_D^c$, $n_{\bar{D}} = \sum_{c=1}^{m} n_{\bar{D}}^c)$. The biomarkers in for an arbitrary observation in center c are denoted $\mathbf{X}^c$; those for an arbitrary case in center c are denoted $\mathbf{X}_D^c$ and those for an arbitrary control in center c are denoted $\mathbf{X}_{\bar{D}}^c$. We are interested in the

center-adjusted AUC (aAUC) for a combination of the biomarkers defined by $\boldsymbol{\theta}$:

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

where M is the number of centers in the population (where $M \in [m, \infty]$) and $w_c = P(C = c|D = 1)$. The empirical aAUC, $a\hat{AUC}$, is based on empirical estimates of the center-specific AUCs, $\hat{AUC}_c$, and empirical estimates of the weights, $\hat{w}_c$:

$$a\hat{AUC}(\boldsymbol{\theta}) = \sum_{c=1}^{m} \hat{w}_c \hat{AUC}_c(\boldsymbol{\theta})$$
$$\hat{AUC}_c(\boldsymbol{\theta}) = \frac{1}{n_D^c n_{\bar{D}}^c} \sum_{i:D_i^c=1, j:D_j^c=0} 1(\boldsymbol{\theta}^\top \mathbf{X}_i^c > \boldsymbol{\theta}^\top \mathbf{X}_j^c)$$
$$\hat{w}_c = \frac{n_D^c}{\sum_{c=1}^{m} n_D^c},$$

where D_i^c denotes the outcome for observation i in center c, $\mathbf{X}_i^c$ denotes the biomarkers for the i^{th} case in center c, and $\mathbf{X}_j^c$ denotes the biomarkers for the j^{th} control in center c.

5.3.1 Direct Maximization

Previous research has considered methods to construct biomarker combinations by directly maximizing smooth approximations to the AUC as an alternative to logistic regression. We propose a distribution-free method that extends this idea to the center-adjusted AUC.

We can consider

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} aAUC(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{c=1}^{M} w_c AUC_c(\boldsymbol{\theta}).$$

As with the unadjusted AUC, we are limited to maximizing the empirical estimate, $a\hat{AUC}$, in practice. Of course, $a\hat{AUC}$ is a function of $\hat{AUC}_c$, 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 $aA\hat{U}C$.

As described above, several smooth approximations to $A\hat{U}C$ have been proposed, and these can be applied to $A\hat{U}C_c$. We use the probit function approximation in light of its accuracy, stability, and ease of implementation. In particular, we propose the following SaAUC estimate:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} aR_n(\boldsymbol{\theta}), \tag{5.2}$$

where

$$\begin{aligned} aR_n(\boldsymbol{\theta}) &= \sum_{c=1}^{m} \hat{w}_c R^c_{n_c}(\boldsymbol{\theta}) \\ R^c_{n_c}(\boldsymbol{\theta}) &= \frac{1}{n^c_D n^c_{\bar{D}}} \sum_{i:D^c_i=1, j:D^c_j=0} \Phi\left\{\boldsymbol{\theta}^\top(\mathbf{X}^c_i - \mathbf{X}^c_j)/h_c\right\} \\ \hat{w}_c &= \frac{n^c_D}{\sum_{c=1}^m n^c_D}, \end{aligned}$$

and h_c is a tuning parameter such that $h_c \to 0$ as $n_c \to \infty$.

In the above definition, each center has its own tuning parameter h_c. We propose choosing 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 defined in (5.2) is a sum of smooth functions, and is therefore also smooth. In order to incorporate the $||\boldsymbol{\theta}||_2 = 1$ constraint suggested by Fong et al. (2016), we use Lagrange multipliers. Existing software can be used to estimate $\boldsymbol{\theta}$ (e.g., the `Rsolnp` package in `R`). Asymptotic results for this method are given in Section 5.3.2.

5.3.2 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 patient characteristics that affect discrimination. 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:

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

where λ is a fixed penalty parameter, $\lambda \geq 0$. Using the ℓ_2-norm in the penalty function allows for some variation in performance across centers; that is, small deviations around $aR_n(\boldsymbol{\theta})$ will be made smaller by squaring, reducing their contribution to the size of the penalty term, $\sum_{c=1}^{m} \hat{w}_c \left(R^c_{n_c}(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta}) \right)^2$. 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 $\boldsymbol{\theta}_\lambda$.

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. In order to incorporate the $||\boldsymbol{\theta}||_2 = 1$ constraint, we use Lagrange multipliers.

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 maximiza-

tion of aR_n without penalization. We have previously demonstrated (Lemmas 4.1 and 4.2) that, under certain conditions, $\hat{AUC}_c(\boldsymbol{\theta})$ converges uniformly in probability to $AUC_c(\boldsymbol{\theta})$ and $a\hat{AUC}(\boldsymbol{\theta})$ converges uniformly in probability to $aAUC(\boldsymbol{\theta})$, and we use these results in the proof of the theorem. Throughout, we will assume that the center-specific disease prevalence is non-trivial; that is, $P(D = 1|C = c) \in (0, 1)$, $c = 1, ..., M$. Let

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

$$Q(\boldsymbol{\theta}; \lambda) = aAUC(\boldsymbol{\theta}) - \lambda \sum_{c=1}^{M} w_c \left(AUC_c(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta}) \right)^2,$$

and let $\mathbf{X}^c_i$ denote the collection of biomarkers for the i^{th} observation in center c. 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 with distribution function F_c 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) Both the maximum of $\tilde{Q}_n(\boldsymbol{\theta}; \lambda)$ and the maximum of $Q(\boldsymbol{\theta}; \lambda)$ over $B = \{\boldsymbol{\theta} \in \mathbb{R}^p : ||\boldsymbol{\theta}||_2 = 1, |\theta_k| > 0\}$ are attained.

Theorem 5.1. *Fix $\lambda \geq 0$ and suppose conditions (A1)-(A5) hold. Then* $\max_{\boldsymbol{\theta} \in B} Q(\boldsymbol{\theta}; \lambda) = Q(\hat{\boldsymbol{\theta}}_\lambda; \lambda) + o_p(1)$ *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 Appendix A.3.1. The proof of the theorem 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 using results from Chapter 4. The proof then uses previous results for R_n (Ma and Huang, 2007) 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.

Choosing 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 (that is, measures of performance for a model that are based on the same data used to fit the model) will tend to be optimistic (Hastie et al., 2016). Thus, selecting a value of λ based on apparent performance may result in substantially diminished performance in new data. Cross-validation is one method for avoiding 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 in general is 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 in terms 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 proposing an automated procedure, as the trade-offs involved in using a larger or smaller value of λ may depend on the individual investigator and/or the specific context. As with any resampling procedure, the LOCOCV procedure we propose will be most useful when the centers available for estimation are in some sense representative of the population of centers we wish to consider.

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

5.4 Results

5.4.1 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:

1. $M = 50, N_c = 5,000, m = 6, n_c = 200$ ("m=6")

2. $M = 500, N_c = 500, m = 50, n_c = 50$ ("m=50")

3. $M = 5000, N_c = 200, m = 500, n_c = 20$ ("m=500")

Fong et al. (2016) note 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 X_1 and X_2 and considered four scenarios.

- Scenario I:

$$\left(\begin{array}{c} X_1 \\ X_2 \end{array} \middle| C \right) = \{(1 - \Delta) \times Z_0\} + \{\Delta \times Z_1\}$$

$$(D|X_1, X_2, C) \sim \text{Bernoulli}\left[\text{expit}\{\theta_0^C + 4X_1 - 3X_2 - (X_1 - X_2)^3\}\right]$$

$$\theta_0^C \sim \text{Uniform}(-1, 1)$$

- Scenario II:

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

$$\nu_C \sim \text{Uniform}(-0.1, 0.1)$$

$$(D|X_1, X_2, C) \sim \text{Bernoulli}\left[\text{expit}\{\theta_0^C + 4X_1 - 3X_2 - (X_1 - X_2)^3\}\right]$$

$$\theta_0^C \sim \text{Uniform}(-1, 1)$$

- Scenario III:

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

$$\nu_C \sim \text{Uniform}(-0.1, 0.1)$$

$$(D|X_1, X_2, C) \sim \text{Bernoulli}\left[\text{expit}\{\theta_0^C + 4\omega_C X_1 - 3X_2 - (\omega_C X_1 - X_2)^3\}\right]$$

$$\theta_0^C \sim \text{Uniform}(-1, 1)$$

$$\omega_C \sim \text{Uniform}(0.75, 1)$$

- Scenario IV:

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

$$\nu_C \sim \text{Uniform}(-0.1, 0.1)$$

$$(D|X_1, X_2, C) \sim \text{Bernoulli}\left[\text{expit}\{\theta_0^C + 4\omega_C X_1 - 3X_2 - (\omega_C X_1 - X_2)^3\}\right]$$

$$\theta_0^C \sim \text{Uniform}(-1, 1)$$

$$\omega_C \sim \text{Uniform}(1, 1.1)$$

In each scenario,

$$Z_0 \sim N\left(\mathbf{0}, 0.2 \times \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$$
$$Z_1 \sim N\left(\mathbf{0}, 2 \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),$$

where Z_0 and Z_1 were independent and $\Delta \sim$ Bernoulli(π), where $\pi = 0.05$ when outliers were simulated and $\pi = 0$ otherwise (independent of Z_0 and Z_1).

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.

We present some key results in the plots below; the full results are given in Appendix B.4.1. Figure 5.1 shows the results for $m = 6$ under Scenario I with outliers. Clearly, the proposed method outperformed both standard and robust logistic regression, both in terms of the center-adjusted AUC and the center-specific AUCs. The performance of the combination estimated by robust logistic regression was also considerably more variable than the performance of the combinations estimated by standard logistic regression or the proposed SaAUC method. Figure 5.2 shows the results for the same scenario with $m = 50$; here, estimates from robust logistic regression were used as starting values (in general, we found that this gave very similar results as when estimates from conditional logistic regression were used). Again, we see that the proposed method clearly outperformed both forms of logistic

regression. Finally, we see similar results in Figure 5.3, which presents the results for the same scenario with $m = 500$. As was observed in Fong et al. (2016) for the AUC, when outliers are not present, all three approaches yielded similar results (Appendix B.4.1).

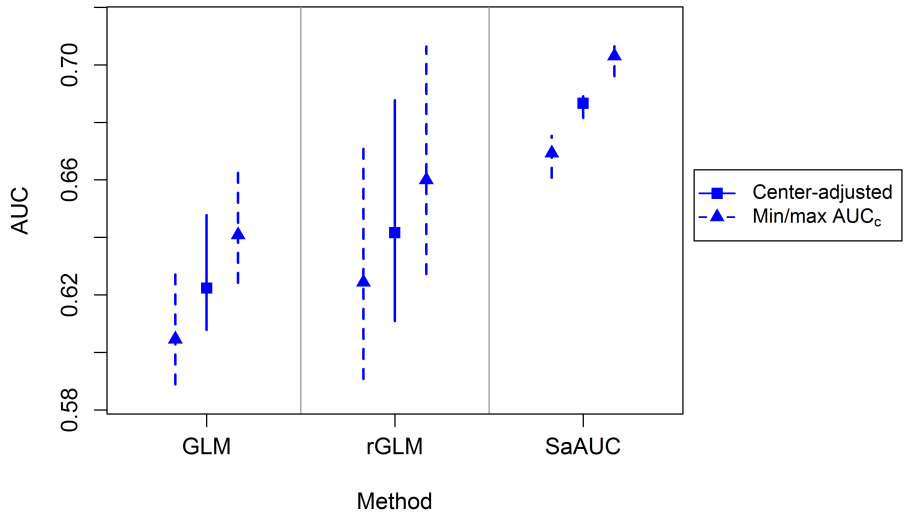

Figure 5.1: Simulation results for training data with 6 centers, under Scenario I with outliers. The starting values for the SaAUC maximization routine were the robust logistic regression estimates. This plot shows the median and middle 90% of the distribution (across simulations) of the center-adjusted AUC and the minimum and maximum center-specific AUCs, calculated in test data.

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 centers were concordant and were removed from the analysis. Up to 11% of simulations had one or two concordant training

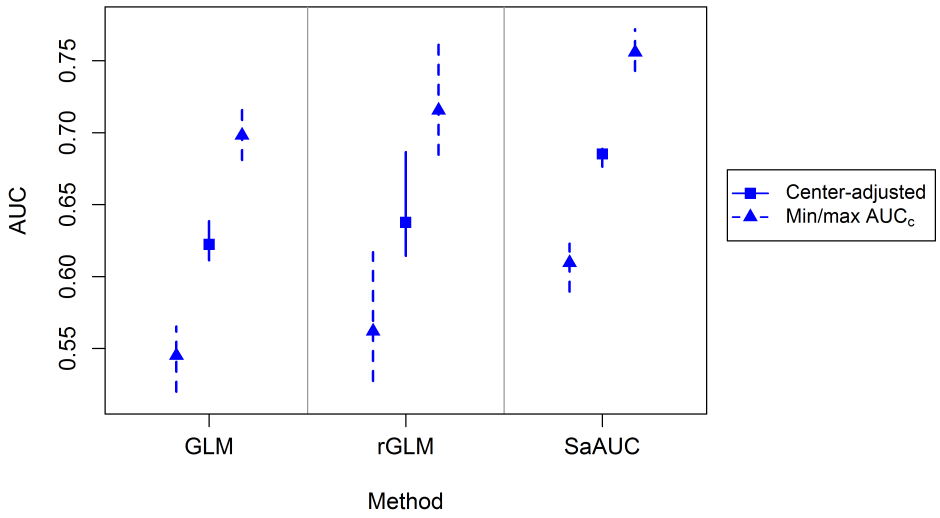

Figure 5.2: Simulation results for training data with 50 centers, under Scenario I with outliers. The starting values for the SaAUC maximization routine were the robust logistic regression estimates. This plot shows the median and middle 90% of the distribution (across simulations) of the center-adjusted AUC and the minimum and maximum center-specific AUCs, calculated in test data.

centers.

5.4.2 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. Essentially, our goal was to establish "proof of principle."

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. In particular, we first sampled $m = 6$ centers from a population of $M = 50$ centers with $N_c = 5,000$ observations in each. Then, within each of the m sampled centers, we sampled $n_c = 200$ observations. These observations formed the training data, in which the combinations were constructed. The fitted combinations were then evaluated in independent test data, which

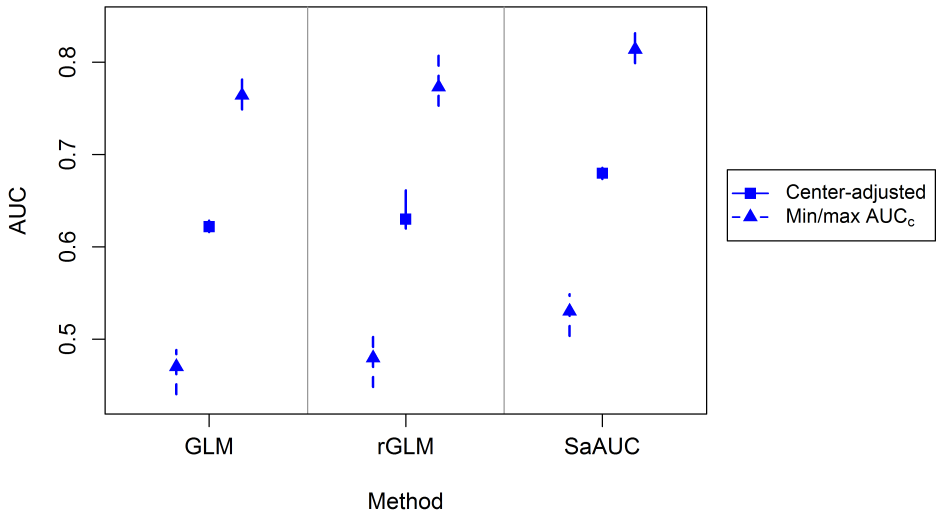

Figure 5.3: Simulation results for training data with 500 centers, under Scenario I with outliers. The starting values for the SaAUC maximization routine were the robust logistic regression estimates. This plot shows the median and middle 90% of the distribution (across simulations) of the center-adjusted AUC and the minimum and maximum center-specific AUCs, calculated in test data.

consisted of the $N_c = 5,000$ observations in each of the $M - m = 44$ centers not used in the training 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 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 present a handful of examples here, and include several more in Appendix B.4.2. In these examples, we considered 50 values of λ equally-spaced (on the log scale) between 0.1 and 200. This range of values was chosen somewhat arbitrarily. 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 method. The key point is that the range of λ values used here is meant to be illustrative, not prescriptive.

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, $\log_{10}\lambda$ is presented on the x-axis, and there are two y-axes. The y-axis on the left side plots 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 y-axis on the right side plots the variability on the standard deviation scale and corresponds to the red lines (variability relative to the aAUC estimate in the training centers) and blue lines (variability relative to the aAUC estimate in the test centers). For example, in the test data, for a combination $\hat{\boldsymbol{\theta}}$ estimated in the training data, we would have

$$\text{Variability relative to training: } \sum_{c=1}^{M-m} w_c \left(AUC_c(\hat{\boldsymbol{\theta}}) - a\hat{AUC}(\hat{\boldsymbol{\theta}}) \right)^2$$

$$\text{Variability relative to test: } \sum_{c=1}^{M-m} w_c \left(AUC_c(\hat{\boldsymbol{\theta}}) - aAUC(\hat{\boldsymbol{\theta}}) \right)^2,$$

where w_c are the weights in the test centers, $a\hat{AUC}$ denotes the estimated aAUC in training, and $aAUC$ and AUC_c denote the aAUC and AUC_c, respectively, in the test data (where the centers in the test data are assumed to be so large that these are close to the population values). Finally in the training and test data results, the dashed lines represent the standard logistic regression results, and the dot-dashed lines represent the robust logistic regression results.

One example is presented in Figure 5.4. We consider each of the three plots individually.

- **Training data:** The results in the training data indicate that the center-specific AUCs based on the penalized estimation method (that is, based on the combination fitted by the penalized estimation method; gray lines) are approaching a common value as λ increases: the "low" center-specific AUC curves are increasing, and the "high" center-specific AUC curves are decreasing. Thus, we see a drop in the variability of the center-specific AUCs based on the penalized estimation method (red solid line) as λ increases. Since for larger values of λ the "high" center-specific AUCs decrease more rapidly than the "low" center-specific AUCs increase, there is a drop in the center-adjusted AUC based on the penalized estimation method (black solid line) for $\lambda > 10$. For all values of λ, the variability in the center-specific AUCs based on the penalized estimation method (red solid line) is lower than for standard and robust logistic regression (red dashed and dot-dashed lines). For small values of λ, the adjusted AUC based on the penalized estimation method (black solid line) is slightly higher than the adjusted AUCs for standard and robust logistic regression (black dashed and dot-dashed lines). For larger values of λ, the adjusted AUC based on the penalized estimation method dips below those based on standard and robust logistic regression.

- **LOCOCV:** This plot only includes the results for the penalized estimation procedure. Here we see a pattern that mimics what was observed in the training data, with some differences. As λ increases, the center-specific AUCs (gray lines) become more similar, resulting in a decrease in variability in center-specific AUCs relative to both the adjusted AUC estimated in the "training" centers (the centers used by the LOCOCV procedure for estimation; red line) and the adjusted AUC in the "test" centers (the centers held out by the LOCOCV procedure for evaluation; blue line). This decrease in variability is seen for $\lambda < 10^{1.25}$, beyond which there is a small increase in variability due to the continued decrease in the center-specific AUCs in two centers. Likewise, we see that the center-adjusted AUC is relatively flat for $\lambda < 10^{1.25}$. This plot might lead us to choose

$\lambda \approx 10 - 10^{1.25}$.

- **Test data:** These results generally mirror the patterns observed in both the training data and the LOCOCV results. The center-specific AUCs based on the penalized estimation method (gray lines) become more similar and the center-adjusted AUC based on the penalized estimation method (black solid line) is relatively flat for $\lambda < 10$. The variability in center-specific AUCs relative to the adjusted AUC in both training (red solid line) and test (blue solid line) for the penalized estimation method are decreasing for $\lambda < 10^{1.25}$. For standard and robust logistic regression, the variability in center-specific AUCs relative to both the adjusted AUC in the training centers (red dashed and dot-dashed lines) and the adjusted AUC in the test centers (blue dashed and dot-dashed lines) is higher than the corresponding variability for the penalized estimation method (red and blue solid lines). On the other hand, the adjusted AUCs for standard and robust logistic regression (black dashed and dot-dashed lines) are similar to the adjusted AUC based on the penalized estimation method for $\lambda < 10^{0.5}$, beyond which the adjusted AUC for the penalized estimation method begins to decrease. Thus, the test data, which would not be observable in practice, support the choice of $\lambda \approx 10 - 10^{1.25}$. For λ values in this range, we see a decrease in the center-adjusted AUC from approximately 0.725 to 0.705 and a decrease in variability of more than 50%.

Figures 5.5 and 5.6 present examples where the LOCOCV procedure does a particularly nice job of mimicking the patterns in the test data. Figure 5.7 presents an example where we see a clear benefit to penalization in terms of a substantial reduction in variability in center-specific performance, with little decrease in overall (center-adjusted) performance. Figure 5.8 provides an example where the LOCOCV procedure gives results that are inconclusive or difficult to interpret. When this occurs, it may be best to err on the side of caution and choose smaller values of λ or not penalize at all.

We encountered some datasets where the penalized estimation procedure did not work as well. For instance, Figure 5.9 presents an example where the center-adjusted AUC decre-

ased more quickly with increasing λ in the test data than was suggested by the LOCOCV procedure and the training data. This could lead to a choice of λ that gives slightly worse overall performance than anticipated, although the problem is not severe. Figure 5.10 presents another example where the penalized estimation procedure did not work as well. Here, the variance increases with increasing λ in test data, despite the patterns seen in the training data and the LOCOCV results. 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 increase in variability is fairly small.

Problems with convergence were not common in our simulations. Out of nearly 400 examples considered and 50 values of λ, fewer than 6% of the examples encountered any convergence issues. This generally only occurred with the more extreme examples we considered. The issues with convergence were primarily convergence failures in the cross-validation procedure. In practice, this may require modification of the range of λ values considered. None of the results included here had any convergence failures.

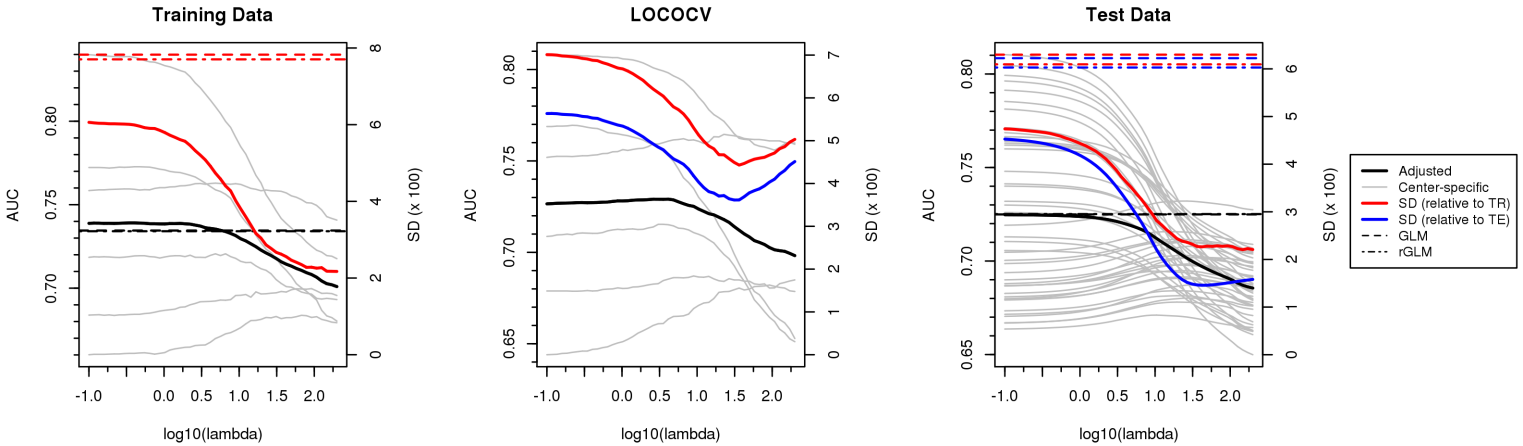

Figure 5.4: Penalized estimation example 1. These results are described in detail in the text.

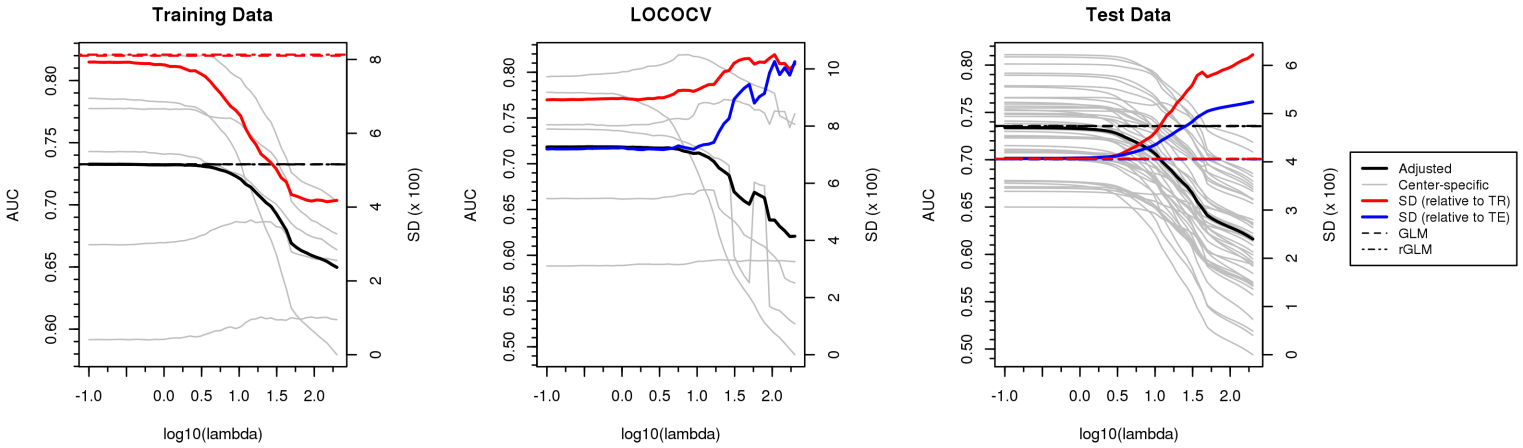

Figure 5.5: Penalized estimation example 2. This is an example where the LOCOCV procedure does very well in mimicking the patterns seen in the test data.

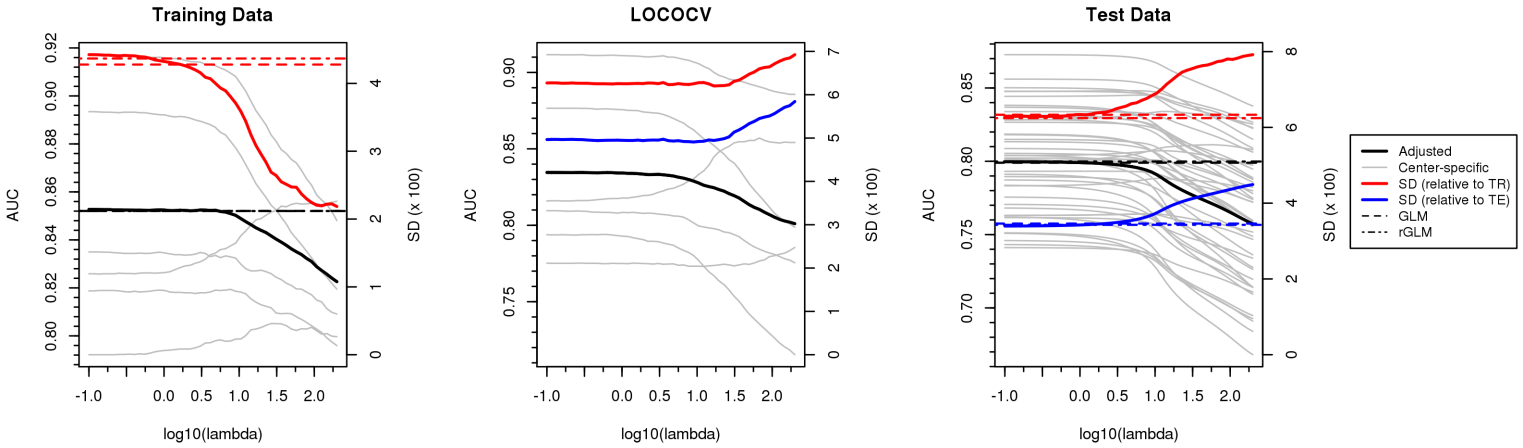

Figure 5.6: Penalized estimation example 3. This is an example where the LOCOCV procedure does very well in mimicking the patterns seen in the test data.

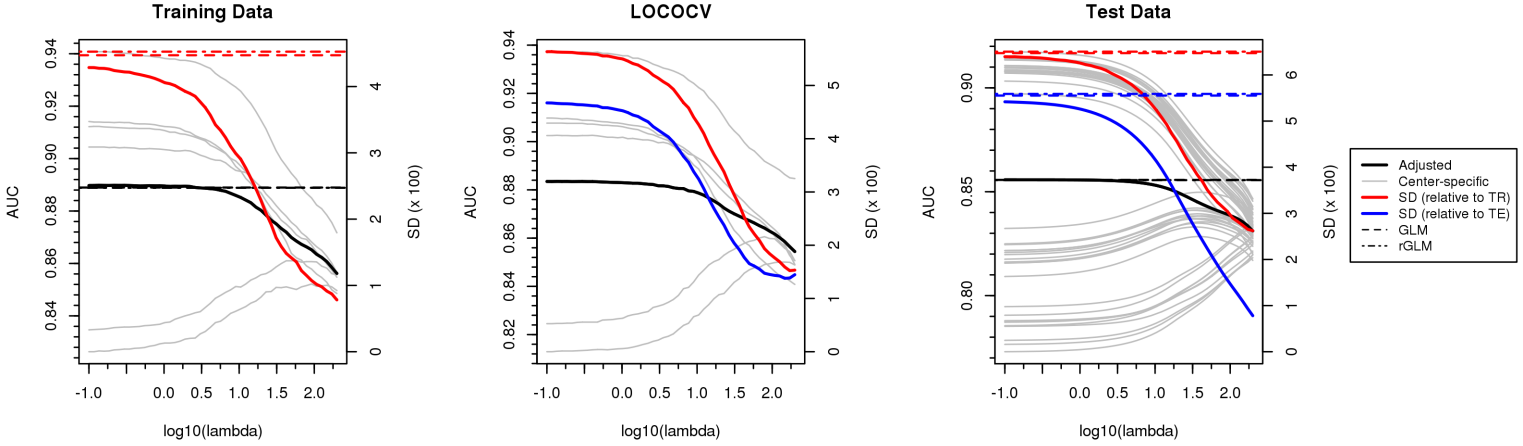

Figure 5.7: Penalized estimation example 4. This example illustrates a setting where there is a clear benefit to penalizing in terms of reduction in variability in performance with little loss in overall performance.

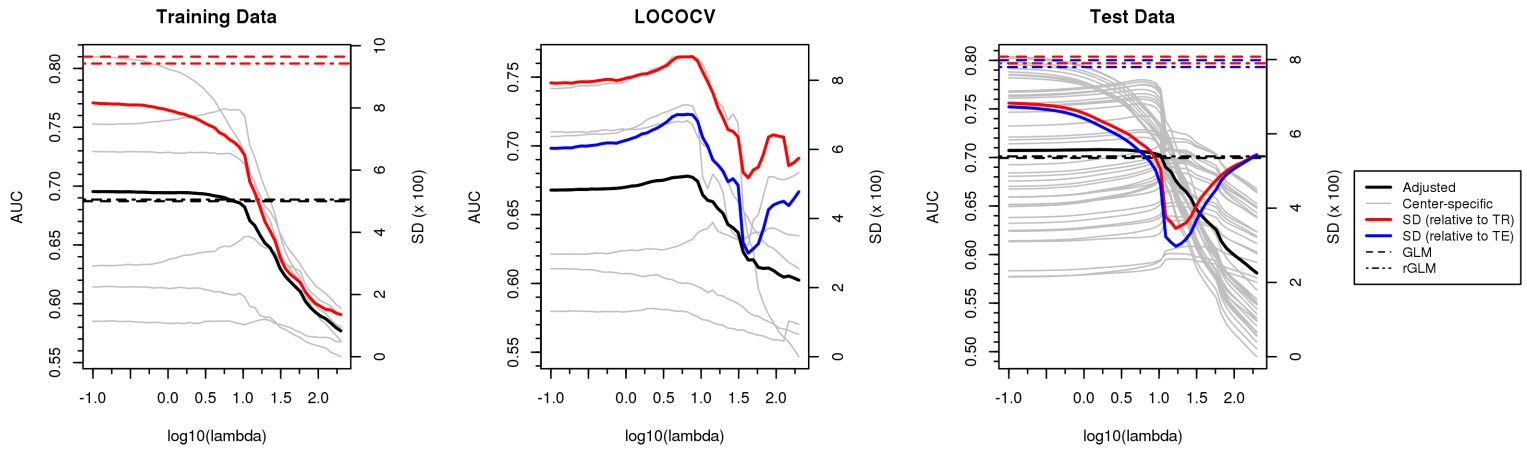

Figure 5.8: Penalized estimation example 5. This is an example where the LOCOCV results are inconclusive in terms of which value of λ to choose.

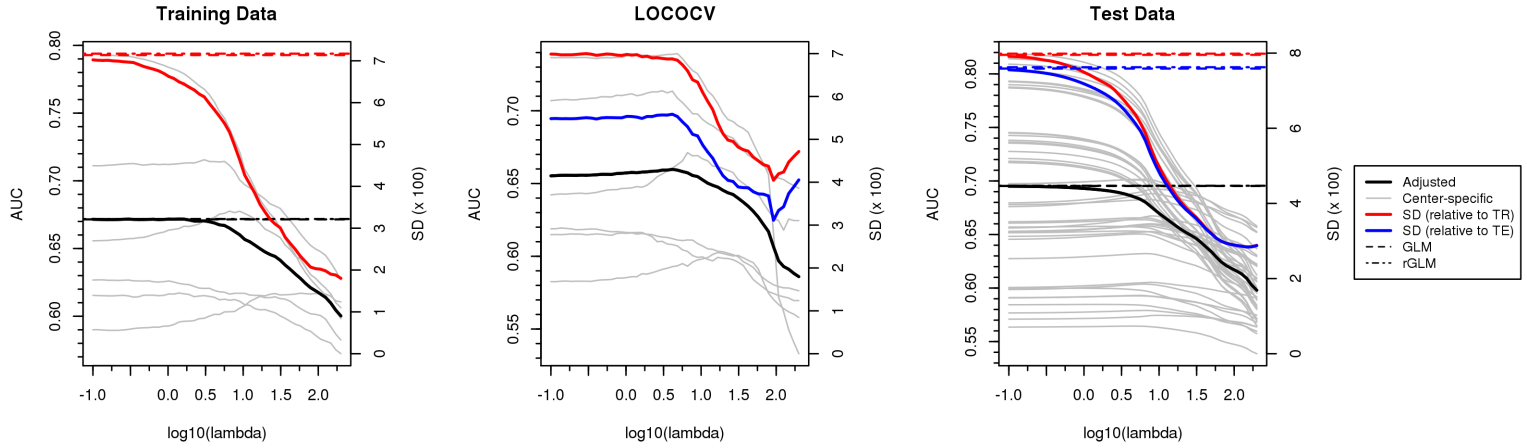

Figure 5.9: Penalized estimation example 6. This is an example where the penalization procedure does not work as well, since the aAUC decreased more quickly with increasing λ in the test data than was indicated by the LOCOCV results and the training data.

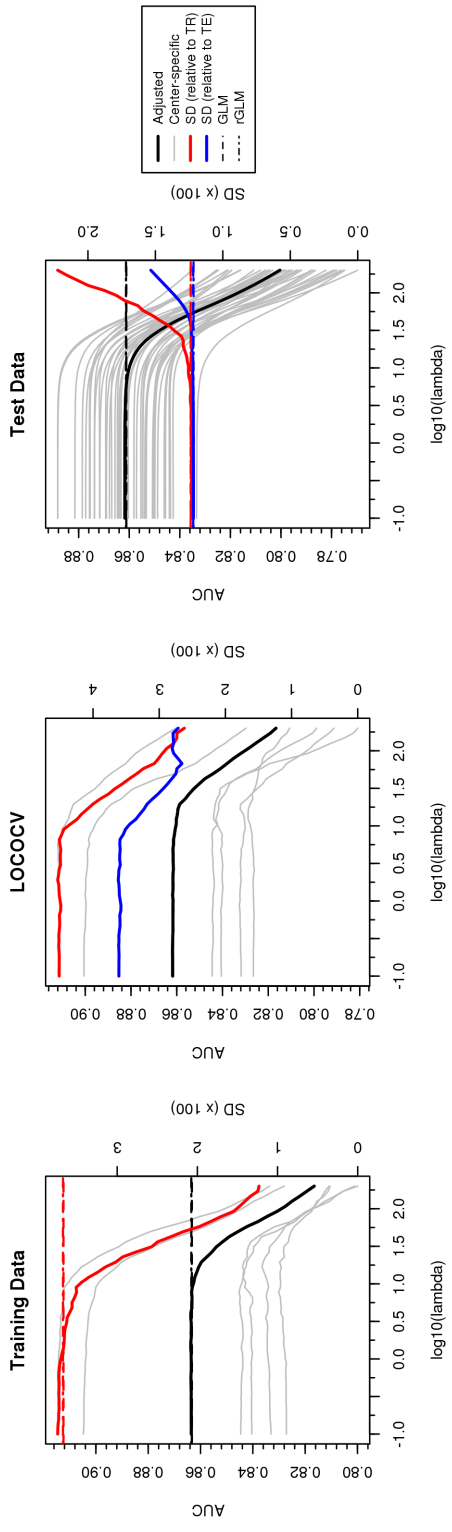

Figure 5.10: Penalized estimation example 7. This is an example where the penalization procedure does not work as well, since the variability increases with increasing λ, despite the patterns seen in the training data and the LOCOCV procedure. However, the increase in variability in performance is fairly small.

Table 5.1: Center-specific AKI prevalence in the TRIBE-AKI study.

Center (n)	*AKI Prevalence (95% CI)*
1 (103)	7.8% (3.4%-14.7%)
2 (53)	17.0% (8.1%-29.8%)
3 (70)	22.9% (13.7%-34.4%)
4 (483)	19.5% (16.0%-23.3%)
5 (27)	22.2% (8.6%-42.3%)
6 (226)	11.1% (7.3%-15.9%)

5.4.3 TRIBE-AKI Data

To illustrate the methods we have developed, we applied them to data from the TRIBE-AKI study and constructed combinations of three biomarkers measured immediately after surgery: urine NGAL, plasma h-FABP, and plasma TNI. We removed observations with missing values for any of these three biomarkers, leaving 962 observations, and log-transformed the biomarker values. The biomarker distributions among AKI controls (stratified by center) are given in Figure 5.11. Table 5.1 shows the center-specific prevalences of AKI.

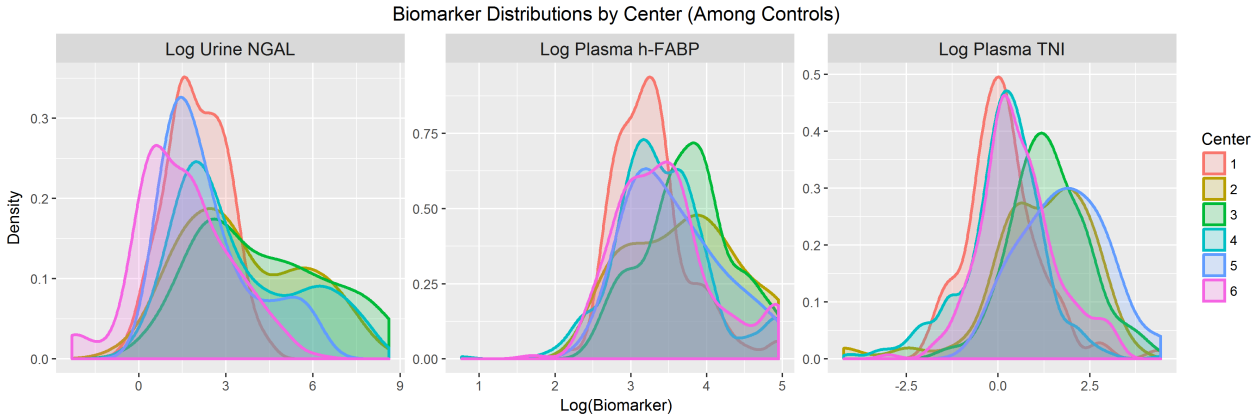

Figure 5.11: Distribution of log urine NGAL, log plasma h-FABP, and log plasma TNI in the TRIBE-AKI study among controls. The biomarker distributions are stratified by center.

We applied standard logistic regression ("GLM"), robust logistic regression ("rGLM"), and the proposed SaAUC method to the TRIBE-AKI study data, after scaling the three biomarkers to have equal variance. The fitted combinations (with normalized coefficients) for the three methods were:

$$\text{GLM: } 0.0720 * \log(\text{NGAL}) + 0.9917 * \log(\text{h-FABP}) - 0.1068 * \log(\text{TNI})$$
$$\text{rGLM: } 0.0720 * \log(\text{NGAL}) + 0.9917 * \log(\text{h-FABP}) - 0.1068 * \log(\text{TNI})$$
$$\text{SaAUC: } 0.0107 * \log(\text{NGAL}) + 0.9585 * \log(\text{h-FABP}) - 0.2849 * \log(\text{TNI}).$$

These combinations had apparent center-adjusted AUCs of 0.6878, 0.6878 and 0.6918, respectively. After optimism correction, the AUC estimates were 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 seeks to optimize a smooth approximation to this estimate.

Finally, we applied the proposed penalized estimation method to the TRIBE-AKI study data, again using the scaled biomarkers (Figure 5.12). 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 $\lambda \approx 0.1$, the center-specific AUC estimates were between 0.6042 and 0.7250, but when $\lambda = 10^{1.5}$, the center-specific AUC estimates were between 0.6270 and 0.6986. Using $\lambda = 10^{1.5}$ to fit the combination in the full TRIBE-AKI study dataset yielded the combination

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

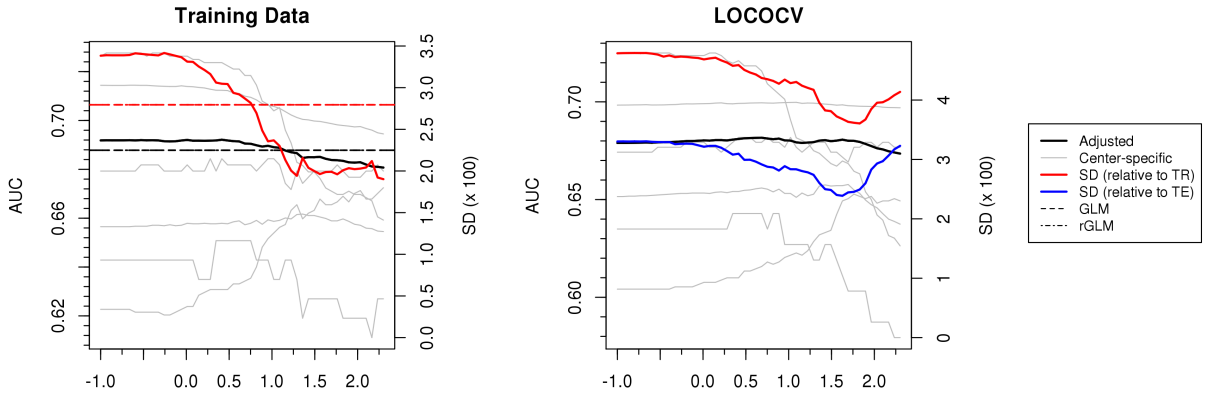

Figure 5.12: Penalized estimation procedure applied to the TRIBE-AKI study data. The results from the LOCOCV procedure support choosing $\lambda \approx 10^{1.5}$, which is expected to give a reduction in variability in center-specific performance of about 25-30%.

5.5 Discussion

We have developed a method to estimate biomarker combinations by maximizing the center-adjusted AUC (aAUC). This method is directly applicable to the covariate-adjusted AUC for any discrete covariate, and so could be applied beyond the multicenter setting. Furthermore, our method includes 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 from a study of biomarker measurements taken after cardiac surgery to construct diagnostic biomarker combinations for acute kidney injury, demonstrating the feasibility of our methods.

An important limitation of the methods we have proposed is that they cannot be used to generate predicted probabilities, as they do not relate the biomarkers to the probability of $D = 1$. As a result, a fitted biomarker combination provided by our method is a tool for risk stratification within each center rather than a risk prediction model. 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 used could also affect the validity of the asymptotic results we have provided. Finally, 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 degree of heterogeneity in the center-specific AUCs, as this provides some insight into how the combination can be expected to perform in a new center. In addition, differences in performance across centers may be scientifically meaningful and merit further investigation. However, when assessing the 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 unreliable. 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. One benefit of direct maximization of the aAUC is that if a single optimal combination exists in the sense of maximizing the center-specific AUC in each center, our method will identify it (in large samples) and if there is no such single optimal combination, we will still succeed in identifying a combination with the optimal aAUC (again, in large samples).

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 using 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, could also be explored.

Chapter 6

CONCLUSION

This dissertation has addressed challenges related to constructing, evaluating, and selecting biomarker combinations. For constructing combinations, this included methods both within and beyond the traditional maximum likelihood framework. In particular, when a multilevel outcome is available and there is interest in single-level prediction, we explored whether using regression methods for multilevel outcomes to construct biomarker combinations offered improvements over the traditional approach of binary logistic regression. We also considered methods for constructing combinations in the multicenter setting; specifically, we assessed the implications of constructing combinations by (i) ignoring center by fitting a marginal logistic regression model, or (ii) accounting for center by using random intercept logistic regression. We illustrated problems with these standard approaches, and proposed using fixed intercept logistic regression to accommodate center in the construction of biomarker combinations. Likelihood-based methods are appealing as they are typically well-understood, well-developed, and broadly accessible.

Outside of the maximum likelihood framework, we proposed a novel method to construct biomarker combinations by directly maximizing a smooth approximation to the empirical true positive rate while constraining a smooth approximation to the empirical false positive rate. In the context of multicenter data, we developed a method for constructing biomarker combinations by directly maximizing a smooth approximation to the empirical center-adjusted AUC; indeed, this method can be applied to maximize the covariate-adjusted AUC for any discrete covariate. Furthermore, this method allows for penalization of variability in the covariate-specific AUC, which may be useful for discrete nuisance covariates beyond center. These maximization methods may be preferable to likelihood-based methods since

they directly target a measure of predictive capacity, and so may provide combinations with better performance.

For evaluating biomarker combinations, we considered the impact of ignoring center in the evaluation of biomarker combinations when the data come from multiple centers. In particular, we discussed the importance of using conditional, rather than marginal, measures of performance in the context of multicenter data, and demonstrated the repercussions of ignoring center during evaluation.

We also considered the issue of combination selection in the setting where a multilevel outcome is available and there is interest in single-level prediction. We proposed an algorithm for combination selection that leverages the additional information available in the multilevel outcome, and showed that this procedure may be preferable to the standard approach of selecting combinations on the basis of the candidate combinations' ability to narrowly predict the target outcome level.

There are many possible extensions to the work we have done. Perhaps most pressing is the extension of these methods to the high-dimensional setting. As the number of biomarkers available in a given dataset continues to increase, it will be important to incorporate biomarker selection into combination construction. This has been considered in the context of maximizing the AUC, where methods have been proposed that include an ℓ_1 penalty function to encourage biomarker selection concurrent with combination construction (Ma and Huang, 2005). A similar approach could be taken to extend our direct maximization methods to the high-dimensional setting.

Likewise, it would be straightforward to build on the work we have already done to develop a method that directly maximizes the covariate-adjusted true positive rate while constraining each of the covariate-specific false positive rates at some clinically acceptable level. In particular, the methods we proposed to directly maximize the true positive rate while constraining the false positive rate and those we have proposed to directly maximize the covariate-adjusted AUC could be extended to the covariate-adjusted true and false positive rates for a discrete covariate. It may also be possible to allow for penalization when the

covariate is a nuisance variable, as we did with the center-adjusted AUC.

It would also be possible to apply the ideas we have discussed to related, but distinct, questions. For instance, decision analytic measures of performance, such as the net benefit, are gaining in popularity (Vickers and Elkin, 2006). The net benefit combines true and false positives, weighted by the benefits and harms of an intervention. If the assumptions of the framework hold, this measure is useful for deciding whether a particular model or test should be used, and can also be used to evaluate the performance of a biomarker for treatment selection. Since the true and false positive rates are components of the net benefit, it may be possible to extend the methods we have developed for maximizing the true positive rate while constraining the false positive rate to the net benefit, yielding a method that constructs biomarker combinations by maximizing the net benefit.

In this dissertation, we have considered current approaches to and proposed novel methods for constructing, evaluating, and selecting biomarker combinations for diagnosis, prognosis, and screening. All of the methods we have developed are available as `R` packages, facilitating their application. We expect widespread interest in biomarkers to continue, commensurate with continued emphasis on developing new and better tools to inform patients about their risks of having or experiencing some clinical outcome. Furthermore, as the feasibility of measuring large numbers of biomarkers continues to increase, and the associated costs continue to decrease, interest in developing biomarker combinations will grow. This work provides novel insights and rigorous methods to aid researchers in developing biomarkers combinations for diagnosis, prognosis, and screening.

BIBLIOGRAPHY

Agresti, A. (2013). *Categorical Data Analysis* (3 ed.). John Wiley & Sons.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Ananth, C. and D. Kleinbaum (1997). Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology 26*(6), 1323–33.

Anderson, J. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological) 46*(1), 1–30.

Anderson, T. and R. Bahadur (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics*, 420–31.

Antonakis, J., S. Bendahan, P. Jacquart, and R. Lalive (2010). On making causal claims: a review and recommendations. *The Leadership Quarterly 21*, 1086–1120.

Armstrong, B. and M. Sloan (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology 129*(1), 191–204.

Baker, S. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics 56*(4), 1082–7.

Bansal, A. and M. Pepe (2013). When does combining markers improve classification performance and what are implications for practice? *Statistics in Medicine 32*, 1877–92.

Bartfay, E., A. Donner, and N. Klar (1999). Testing the equality of twin correlations with multinomial outcomes. *Annals of Human Genetics 63*(4), 341–9.

Begg, C. and R. Gray (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika 71*(1), 11–18.

Begg, M. and S. Lagakos (1990). On the consequence of model misspecification in logistic regression. *Environmental Health Perspectives 87*, 69–75.

Begg, M. and M. Parides (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine 22*, 2591–602.

Bender, R. and U. Grouven (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology 51*(10), 809–16.

Berlin, J., S. Kimmel, T. Ten Have, and M. Sammel (1999). An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics 55*, 470–6.

Bernau, C., T. Augustin, and A.-L. Boulesteix (2013). Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics 69*(3), 693–702.

Bianco, A. and V. Yohai (1996). Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods*. Springer.

Biesheuvel, C., Y. Vergouwe, E. Steyerberg, D. Grobbee, and K. Moons (2008). Polytomous logistic regression analysis could be applied more often in diagnostic research. *Journal of Clinical Epidemiology 61*(2), 125–34.

Blankenberg, S., T. Zeller, O. Saarela, A. Havulinna, F. Kee, H. Tunstall-Pedoe, K. Kuulasmaa, J. Yarnell, R. Schnabel, P. Wild, T. Münzel, K. Lackner, L. Tiret, A. Evans, and V. Salomaa (2010). Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts. *Circulation 121*(22), 2388–97.

Boulesteix, A.-L. and C. Strobl (2009). Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Medical Research Methodology 9*(1), 85.

Bouwmeester, W., K. Moons, T. Kappen, W. Van Klei, J. Twisk, M. Eijkemans, and Y. Vergouwe (2013). Internal validation of risk models in clustered data: a comparison of bootstrap schemes. *American Journal of Epidemiology 177*(11), 1209–17.

Bouwmeester, W., J. Twisk, T. Kappen, W. van Klei, and K. Moons (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Medical Research Methodology 13*(19).

Brumback, B., A. Dailey, L. Brumback, M. Livingston, and Z. He (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters 80*, 1650–4.

Brumback, B., A. Dailey, and H. Zheng (2012). Adjusting for confounding by neighborhood using a proportional odds model for complex survey data. *American Journal of Epidemiology 175*(11), 1133–41.

Bull, S. and A. Donner (1993). A characterization of the efficiency of individualized logistic regressions. *Canadian Journal of Statistics 21*(1), 71–8.

Bünger, S., T. Laubert, U. Roblick, and J. Habermann (2011). Serum biomarkers for improved diagnostic of pancreatic cancer: a current overview. *Journal of Cancer Research and Clinical Oncology 137*(3), 375–89.

Campbell, M. and A. Donner (1989). Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association 84*(406), 587–91.

Cawley, G. and N. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research 11*, 2079–107.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A 158*(3), 419–66.

Coca, S., S. Singanamala, and C. Parikh (2012). Chronic kidney disease after acute kidney injury: a systematic review and meta-analysis. *Kidney International 81*(5), 442–8.

Copas, J. and P. Corbett (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika 89*(2), 315–31.

Cramer, J. (2005). Omitted variables and misspecified disturbances in the logit model. Technical report, Amsterdam School of Economics.

Degos, F., P. Perez, B. Roche, A. Mahmoudi, J. Asselineau, H. Voitot, and P. Bedossa (2010). Diagnostic accuracy of FibroScan and comparison to liver fibrosis biomarkers in chronic viral hepatatis: a multicenter prospective study (the FIBROSTIC study). *Hepatology 53*(6), 1013–21.

Dieleman, J. and T. Templin (2014). Random-effects, fixed-effects and the within-between specification for clustered data in observational health studies: a simulation study. *PLoS One 9*(10).

Ding, Y., S. Tang, S. Liao, J. Jia, S. Oesterreich, Y. Lin, and G. Tseng (2014). Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics 30*(22), 3152–8.

Dupuy, A. and R. Simon (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute 99*(2), 147–57.

Feldmann, U. and I. Steudel (2000). Methods of ordinal classification applied to medical scoring systems. *Statistics in Medicine 19*(4), 575–86.

Feldstein, A., A. Wieckowska, A. Lopez, Y.-C. Liu, N. Zein, and A. McCullough (2009). Cytokeratin-18 fragement levels as noninvasive biomarkers for nonalcoholic steatohepatitis: a multicenter validation study. *Hepatology 50*(4), 1072–8.

Fong, Y., S. Yin, and Y. Huang (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. *Statistics in Medicine 35*(21), 3792–809.

Gail, M., S. Wieand, and S. Piantadosi (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika 71*(3), 431–44.

Gao, F., C. Xiong, Y. Yan, K. Yu, and Z. Zhang (2008). Estimating optimum linear combination of multiple correlated diagnostic tests at a fixed specificity with receiver operating characteristic curves. *Journal of Data Science 6*(1), 105–23.

Gardiner, J., Z. Luo, and L. Roman (2009). Fixed effects, random effects and GEE: what are the differences? *Statistics in Medicine 28*, 221–39.

Gevaert, O., F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics 22*(14), e184–e190.

Graubard, B. and E. Korn (1994). Regression analysis with clustered data. *Statistics in Medicine 13*, 509–22.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine 21*, 389–95.

Greenland, S., J. Robins, and J. Pearl (1999). Confounding and collapsibility in causal inference. *Statistical Science 14*(1), 29–46.

Gruenewald, T., T. Seeman, C. Ryff, A. Karlamangla, and B. Singer (2006). Combinations

of biomarkers predictive of later life mortality. *Proceedings of the National Academy of Sciences 103*(38), 14158–63.

Guisan, A. and F. Harrell (2000). Ordinal response regression models in ecology. *Journal of Vegetation Science 11*(5), 617–26.

Gunasekara, F., K. Richardson, K. Carter, and T. Blakely (2014). Fixed effects analysis of repeated measures data. *International Journal of Epidemiology 43*(1), 264–9.

Guo, J. and Z. Geng (1995). Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society: Series B 57*(1), 263–7.

Han, A. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics 35*(2), 303–16.

Harrell, F. (2013). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media.

Harrell, F., P. Margolis, S. Gove, K. Mason, E. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. Eichenwald (1998). Development of a clinical prediction model for an ordinal outcome: the world health organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Statistics in Medicine 17*(8), 909–44.

Hastie, T., R. Tibshirani, and J. Friedman (2016). *The elements of statistical learning*. Springer Series in Statistics.

Hauck, W. (1984). A comparative study of conditional maximum likelihood estimation of a common odds ratio. *Biometrics 40*(4), 1117–23.

Heagerty, P. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics 55*, 688–98.

Heagerty, P. and B. Kurland (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika 88*(4), 973–85.

Heagerty, P. and S. Zeger (2000). Marginalized multilevel models and likelihood inference. *Statistical Science 15*(1), 1–26.

Hernan, M., D. Clayton, and N. Keiding (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology 40*, 780–5.

Hsu, M.-J. and H.-M. Hsueh (2013). The linear combinations of biomarkers which maximize the partial area under the ROC curves. *Computational Statistics 28*(2), 647–66.

Hu, F., J. Goldberg, D. Hedeker, B. Flay, and M. Pentz (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology 147*(7), 694–703.

Huang, X. (2009). Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics 65*, 361–8.

Hwang, K.-B., B.-Y. Ha, S. Ju, and S. Kim (2013). Partial AUC maximization for essential gene prediction using genetic algorithms. *BMB Reports 46*(1), 41.

Ioannidis, J. (2013). Biomarker failures. *Clinical Chemistry 59*(1), 202–4.

Janes, H., G. Longton, and M. Pepe (2009). Accommodating covariates in ROC analysis. *Stata Journal 9*(1), 17–39.

Janes, H. and M. Pepe (2008). Adjusting for covariates in studies of diagnostic, screening or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology 168*(1), 89–97.

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

Jelizarow, M., V. Guillemot, A. Tenenhaus, K. Strimmer, and A.-L. Boulesteix (2010). Over-optimism in bioinformatics: an illustration. *Bioinformatics 26*(16), 1990–8.

Kahan, B. (2014). Accounting for centre-effects in multicentre trials with a binary outcome - when, why and how? *BMC Medical Research Methodology 14*(20).

Kerr, K., A. Meisner, H. Thiessen-Philbrook, S. Coca, and C. Parikh (2015). Rigor: reporting guidelines to address common sources of bias in risk model development. *Biomarker Research 3*(1), 1.

Kerr, K. and M. Pepe (2011). Joint modeling, covariate adjustment and interaction: contrasting notions in risk prediction models and risk prediction performance. *Epidemiology 22*, 805–12.

Komori, O. and S. Eguchi (2010). A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics 11*(1), 1.

Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer.

Lee, L.-F. (1982). Specification error in multinomial logit models. *Journal of Econometrics 20*, 197–209.

Lin, H., L. Zhou, H. Peng, and X.-H. Zhou (2011). Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics 39*(2), 324–43.

Liu, A., E. Schisterman, and Y. Zhu (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in medicine 24*(1), 37–47.

Liu, D. and X. Zhou (2013). ROC analysis in biomarker combination with covariate adjustment. *Academic Radiology 20*(7), 874–82.

Liu, I. and A. Agresti (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *TEST 14*(1), 1–73.

Localio, A., J. Berlin, and T. Ten Have (2002). Confounding due to cluster in multicenter studies - causes and cures. *Health Services & Outcomes Research Methodology 3*, 195–210.

Localio, A., J. Berlin, T. Ten Have, and S. Kimmel (2001). Adjustments for center in multicenter studies: an overview. *Annals of Internal Medicine 135*, 112–23.

Lukacs, P., K. Burnham, and D. Anderson (2010). Model selection bias and freedmans paradox. *Annals of the Institute of Statistical Mathematics 62*(1), 117–25.

Lukociene, O. and J. Vermunt (2008). A comparison of multilevel logistic regression models with parametric and nonparametric random intercepts. Technical report, Tilburg University, Department of Methodology and Statistics.

Lunt, M. (2005). Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statistics in Medicine 24*(9), 1357–69.

Ma, S. and J. Huang (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics 21*(24), 4356–62.

Ma, S. and J. Huang (2007). Combining multiple markers for classification using ROC. *Biometrics 63*(3), 751–7.

Maas, A., E. Steyerberg, A. Marmarou, G. McHugh, H. Lingsma, I. Butcher, J. Lu, J. Weir, B. Roozenbeek, and G. Murray (2010). Impact recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics 7*(1), 127–34.

Manor, O., S. Matthews, and C. Power (2000). Dichotomous or categorical response? analysing self-rated health and lifetime social class. *International Journal of Epidemiology 29*(1), 149–57.

McCulloch, C. and J. Neuhaus (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science 26*(3), 388–402.

McHugh, G., I. Butcher, E. Steyerberg, A. Marmarou, J. Lu, H. Lingsma, J. Weir, A. Maas, and G. Murray (2010). A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the impact project. *Clinical Trials 7*(1), 44–57.

McIntosh, M. and M. Pepe (2002). Combining several screening tests: optimality of the risk score. *Biometrics 58*(3), 657–64.

Mood, C. (2010). Logistic regression: why we cannot do what we think we can do, and what we can do about it. *European Sociological Review 26*(1), 67–82.

Moore, R., A. Brown, M. Miller, S. Skates, W. Allard, T. Verch, M. Steinhoff, G. Messerlian, P. DiSilvestro, C. Granai, and R. Bast (2008). The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. *Gynecologic Oncology 108*(2), 402–8.

Neuhaus, J., W. Hauck, and J. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika 79*(4), 755–62.

Neuhaus, J. and N. Jewell (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika 80*(4), 807–15.

Neuhaus, J. and J. Kalbfleisch (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics 54*, 638–45.

Neuhaus, J., J. Kalbfleisch, and W. Hauck (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review 59*(1), 25–35.

Neuhaus, J. and M. Lesperance (1996). Estimation efficiency in a binary mixed-effects model setting. *Biometrika 83*(2), 441–6.

Neuhaus, J. and C. McCulloch (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society: Series B 68*(5), 859–72.

Neuhaus, J., A. Scott, C. Wild, Y. Jiang, C. McCulloch, and R. Boyland (2014). Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics 70*(1), 44–52.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32.

Nickolas, T., K. Schmidt-Ott, P. Canetta, C. Forster, E. Singer, M. Sise, A. Elger, O. Maarouf, D. Sola-Del Valle, M. O'Rourke, E. Sherman, P. Lee, A. Geara, P. Imus, A. Guddati, A. Polland, W. Rahman, S. Elitok, N. Malik, J. Giglio, S. El-Sayegh, P. Devarajan, S. Hebber, S. Saggi, B. Hahn, R. Kettritz, F. Luft, and J. Barasch (2012). Diagnostic and prognostic stratification in the emergency department using urinary biomarkers of nephron damage. *Journal of the American College of Cardiology 59*(3), 246–55.

Norris, C., W. Ghali, L. Saunders, R. Brant, D. Galbraith, P. Faris, M. Knudtson, and A. Investigators (2006). Ordinal regression model and the linear regression model were superior to the logistic regression models. *Journal of Clinical Epidemiology 59*(5), 448–56.

Parikh, C., S. Coca, H. Thiessen-Philbrook, M. Shlipak, J. Koyner, Z. Wang, C. Edelstein, P. Devarajan, U. Patel, M. Zappitelli, C. Krawczeski, C. Passik, M. Swaminathan, and A. Garg (2011). Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. *Journal of the American Society of Nephrology 22*(9), 1748–57.

Pavlou, M., G. Ambler, S. Seaman, and R. Omar (2015). A note on obtaining correct

marginal predictions from a random intercepts model for binary outcomes. *BMC Medical Research Methodology 15*(59).

Pennings, J., M. Koster, W. Rodenburg, P. Schielen, and A. de Vries (2009). Discovery of novel serum biomarkers for prenatal down syndrome screening by integrative data mining. *PLoS ONE 4*(11).

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

Pepe, M., T. Cai, and G. Longton (2006). Combining predictors for classification using area under the receiver operating characteristic curve. *Biometrics 62*, 221–9.

Pepe, M. and H. Janes (2013). Methods of evaluating prediction performance of biomarkers and tests. In M.-L. Lee, M. Gail, R. Pfeiffer, G. Satten, T. Cai, and A. Gandy (Eds.), *Risk Assessment and Evaluation of Predictions*, pp. 107–42. Springer.

Pepe, M. and M. Thompson (2000). Combining diagnostic test results to increase accuracy. *Biostatistics 1*(2), 123–40.

Rabe-Hesketh, S. and A. Skrondal (2010). Generalized linear mixed models. In P. Peterson, E. Baker, and B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)*. Elsevier.

Ricamato, M. and F. Tortorella (2011). Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recognition 44*(10), 2669–77.

Riley, R., I. Ahmed, T. Debray, B. Willis, J. Noordzij, J. Higgins, and J. Deeks (2015). Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Statistics in Medicine 34*(13), 2081–103.

Risselada, R., H. Lingsma, A. Molyneux, R. Kerr, J. Yarnold, M. Sneade, E. Steyerberg, and M. Sturkenboom (2010). Prediction of two month modified rankin scale with an ordinal

prediction model in patients with aneurysmal subarachnoid haemorrhage. *BMC Medical Research Methodology 10*(1), 86.

Roozenbeek, B., H. Lingsma, P. Perel, P. Edwards, I. Roberts, G. Murray, A. Maas, and E. Steyerberg (2011). The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Critical Care 15*(3), R127.

Roukema, J., R. van Loenhout, E. Steyerberg, K. Moons, S. Bleeker, and H. Moll (2008). Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. *Journal of Clinical Epidemiology 61*(2), 135–41.

Schildcrout, J. and P. Heagerty (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics 9*(4), 735–49.

Schisterman, E., D. Faraggi, and B. Reiser (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine 23*(21), 3319–31.

Schuit, E., A. Kwee, M. Westerhuis, H. van Dessel, G. Graziosi, J. van Lith, J. Nijhuis, S. Oei, H. Oosterbaan, N. Schuitemaker, M. Wouters, G. Visser, B. Mol, K. Moons, and R. Groenwold (2012). A clinical prediction model to assess the risk of operative delivery. *BJOG: An International Journal of Obstetrics & Gynaecology 119*(8), 915–23.

Scott, S., M. Goldberg, and N. Mayo (1997). Statistical assessment of ordinal outcomes in comparative studies. *Journal of Clinical Epidemiology 50*(1), 45–55.

Seaman, S., M. Pavlou, and A. Copas (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine 33*, 5371–87.

Shapiro, N., S. Trzeciak, J. Hollander, R. Birkhahn, R. Otero, T. Osborn, E. Moretti, H. Nguyen, K. Gunnerson, D. Milzman, D. Gaieski, M. Goyal, C. Cairns, L. Ngo, and E. Rivers (2009). A prospective, multicenter derivation of a biomarker panel to assess risk

of organ dysfunction, shock, and death in emergency department patients with suspected sepsis. *Critical Care Medicine 37*(1), 96–104.

Skrondal, A. and S. Rabe-Hesketh (2008). Multilevel and related models for longitudinal data. In J. de Leeuw and E. Meijer (Eds.), *Handbook of Multilevel Analysis*. Springer.

Skrondal, A. and S. Rabe-Hesketh (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society: Series C 63*(2), 211–37.

Smith, J., J. Everhart, W. Dickson, W. Knowler, and R. Johannes (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261–5.

Snell, K., H. Hua, T. Debray, J. Ensor, M. Look, K. Moons, and R. Riley (2016). Multivariate meta-analysis of individual participant data helps externally validate the performance and implementation of a prediction model. *Journal of Clinical Epidemiology 69*, 40–50.

Steyerberg, E. (2008). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media.

Steyerberg, E., S. Bleeker, H. Moll, D. Grobbee, and K. Moons (2003). Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology 56*(5), 441–7.

Steyerberg, E., N. Mushkudiani, P. Perel, I. Butcher, J. Lu, G. McHugh, G. Murray, A. Marmarou, I. Roberts, J. Habbema, and A. Maas (2008). Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine 5*(8).

Strömberg, U. (1996). Collapsing ordered outcome categories: A note of concern. *American Journal of Epidemiology 144*(4), 421–4.

Su, J. and J. Liu (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association 88*(424), 1350–5.

Ten Have, T., J. Landis, and S. Weaver (1995). Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statistics in Medicine 14*, 413–29.

Ten Have, T., J. Landis, and S. Weaver (1996). Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statistics in Medicine 15*, 1227–9.

van Calster, B., L. Valentin, C. van Holsbeke, A. Testa, T. Bourne, S. van Huffel, and D. Timmerman (2010). Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. *BMC Medical Research Methodology 10*(1), 96.

Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. and J. Wellner (2000). *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

van Klaveren, D., E. Steyerberg, P. Perel, and Y. Vergouwe (2014). Assessing discriminative ability of risk models in clustered data. *BMC Medical Research Methodology 14*(5).

van Klaveren, D., E. Steyerberg, and Y. Vergouwe (2014). Interpretation of concordance measures for clustered data. *Statistics in Medicine 33*(4), 714–6.

van Oirbeek, R. and E. Lesaffre (2010). An application of Harrell's C-index to PH frailty models. *Statistics in Medicine 29*, 3160–71.

Varma, S. and R. Simon (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics 7*(1), 91.

Vickers, A. and E. Elkin (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making 26*(6), 565–74.

Vuilleumier, N., G. Le Gal, F. Verschuren, A. Perrier, H. Bounameaux, N. Turck, J.-C. Sanchez, N. Mensi, T. Perneger, D. Hochstrasser, and M. Righini (2008). Cardiac biomarkers for risk stratification in non-massive pulmonary embolism: a multicenter prospective study. *Journal of Thrombosis and Haemostasis 7*(3), 391–8.

Wang, Z. and Y.-C. I. Chang (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics 12*(2), 369–85.

Whittemore, A. and J. Halpern (2003). Logistic regression of family data from retrospective study designs. *Genetic Epidemiology 25*(3), 177–89.

Winter, B. (1979). Convergence rate of perturbed empirical distribution functions. *Journal of Applied Probability 16*, 163–73.

Wynants, L., W. Bouwmeester, K. Moons, M. Moerbeek, D. Timmerman, S. Van Huffel, B. Van Calster, and Y. Vergouwe (2015). A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *Journal of Clinical Epidemiology 68*(12), 1406–14.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association 93*(441), 120–31.

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

Zethelius, B., L. Berglund, J. Sundström, E. Ingelsson, S. Basu, A. Larsson, P. Venge, and J. Ärnlöv (2008). Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *New England Journal of Medicine 358*(20), 2107–16.

Appendix A

THEORETICAL RESULTS

A.1 Chapter 2

A.1.1 Optimal Combination for Conditionally Bivariate Normal Biomarkers with Proportional Covariance Matrices

Claim: If the biomarkers (X_1, X_2) are conditionally multivariate normal with proportional covariance matrices, that is,

$$(X_1, X_2 \mid D = 0) \sim N(\boldsymbol{\mu}_0, \Sigma), \;\; (X_1, X_2 \mid D = 1) \sim N(\boldsymbol{\mu}_1, \sigma^2\Sigma),$$

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. 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 a monotone increasing function thereof (McIntosh and Pepe, 2002). Let $\mathbf{M} = (X_1, X_2)$. Without loss of generality,

let $\boldsymbol{\mu}_0 = 0$ and let $\boldsymbol{\mu}_1 = \boldsymbol{\mu} = (\mu_{X_1}, \mu_{X_2})$. Then

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

Denote the entries of $(\Sigma)^{-1}$ by

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

Then

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

as claimed, where

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

□

A.1.2 Lemma 2.1

Lemma 2.1. *For a given function $f : \mathbb{R}^d \to \mathbb{R}$, define $\omega_n \in \Omega_n$ and $\omega_0 \in \Omega_0$ such that*

$$f(\omega_n) = \max_{\omega \in \Omega_n} f(\omega), \quad f(\omega_0) = \max_{\omega \in \Omega_0} f(\omega),$$

for two sets $\Omega_n \subseteq \mathbb{R}^d$ and $\Omega_0 \subseteq \mathbb{R}^d$. Further, define

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

where d is the Euclidean distance. If $d_n, e_n \to 0$ w.p. 1 and f is Lipschitz with constant $K > 0$, then $|f(\omega_0) - f(\omega_n)| \to 0$ w.p. 1.

Proof. We prove this by contradiction. Suppose $d_n, e_n \to 0$ w.p. 1 and f is Lipschitz with constant $K > 0$, yet for some $\epsilon > 0$, there exists a subsequence $\{n_i\} \subseteq \mathbb{N}$ such that

$$P\left(\liminf_{n_i \to \infty} |f(\omega_{n_i}) - f(\omega_0)| > \epsilon\right) > 0.$$

Then, there exists $n_\epsilon \in \mathbb{N}$, $\{\omega_{0,n}\} \subseteq \Omega_n$, and $\{\omega_{n,0}\} \subseteq \Omega_0$ such that for every $n \geq n_\epsilon$,

$$d(\omega_{0,n}, \omega_0) \leq \frac{\epsilon}{2K}, \quad d(\omega_{n,0}, \omega_n) \leq \frac{\epsilon}{2K} \text{ w.p. } 1.$$

Then for $n \geq n_\epsilon$,

$$|f(\omega_{0,n}) - f(\omega_0)| \leq \epsilon/2 \text{ w.p. } 1$$

and

$$|f(\omega_{n,0}) - f(\omega_n)| \leq \epsilon/2 \text{ w.p. } 1,$$

so

$$f(\omega_0) \leq f(\omega_{0,n}) + \epsilon/2, \quad f(\omega_n) \leq f(\omega_{n,0}) + \epsilon/2 \text{ w.p. } 1.$$

Since $\{\omega_{0,n}\} \subseteq \Omega_n$,

$$f(\omega_{0,n}) + \epsilon/2 \leq f(\omega_n) + \epsilon/2,$$

and since $\{\omega_{n,0}\} \subseteq \Omega_0$,

$$f(\omega_{n,0}) + \epsilon/2 \leq f(\omega_0) + \epsilon/2.$$

Thus, $|f(\omega_0) - f(\omega_n)| \leq \epsilon/2$ for all $n \geq n_\epsilon$ w.p. 1, giving a contradiction. □

A.1.3 Lemma 2.2

Lemma 2.2. *Under conditions (A1)-(A6) in Section 2.3.2, we have that*

$$\sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-FPR(\boldsymbol{\theta},\delta)\right|\to 0 \ \textit{w.p.}\ 1$$

$$\sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|T\tilde{P}R_{n_{D}}(\boldsymbol{\theta},\delta)-TPR(\boldsymbol{\theta},\delta)\right|\to 0 \ \textit{w.p.}\ 1$$

where $\Omega=\{\boldsymbol{\theta}\in\mathbb{R}^p,\delta\in\mathbb{R}:||\boldsymbol{\theta}||=1\}$.

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

$$\begin{aligned}\sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-FPR(\boldsymbol{\theta},\delta)\right| \leq & \sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\}\right| \\ & + \sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\}-FPR(\boldsymbol{\theta},\delta)\right|.\end{aligned}$$

First we consider $F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\}$. We can write this as

$$F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\}=\frac{1}{n_{\bar{D}}}\sum_{j=1}^{n_{\bar{D}}}\Phi\left(\frac{\boldsymbol{\theta}^\top\mathbf{X}_{\bar{D}j}-\delta}{h}\right)-\int\Phi\left(\frac{\boldsymbol{\theta}^\top\mathbf{x}-\delta}{h}\right)dF_{\bar{D}}(\mathbf{x}).$$

The class of functions $\mathcal{G}_1=\{(\boldsymbol{\theta},\delta)\mapsto\boldsymbol{\theta}^\top\mathbf{x}-\delta:\boldsymbol{\theta}\in\mathbb{R}^p,\delta\in\mathbb{R},\mathbf{x}\in\mathbb{R}^p\}$ is a Vapnik–Chervonenkis class, and, since $\Phi(\cdot/h),h>0$, is monotone, the class of functions $\mathcal{G}_2=\{(\boldsymbol{\theta},\delta)\mapsto\Phi\left\{(\boldsymbol{\theta}^\top\mathbf{x}-\delta)/h\right\}:\boldsymbol{\theta}\in\mathbb{R}^p,\delta\in\mathbb{R},\mathbf{x}\in\mathbb{R}^p,h>0\}$ is Vapnik–Chervonenkis (Kosorok, 2008; Van der Vaart, 2000; van der Vaart and Wellner, 2000). Since the constant 1 is an applicable envelope function for this class, $\mathcal{G}_2$ is $F_{\bar{D}}$-Glivenko–Cantelli, giving (Kosorok, 2008; van der Vaart and Wellner, 2000)

$$\sup_{(\boldsymbol{\theta},\delta)\in\Omega}\left|F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)-E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\}\right|\to 0 \text{ w.p. } 1.$$

Next we consider $E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\} - FPR(\boldsymbol{\theta},\delta)$. We can write this as

$$E\{F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta},\delta)\} - FPR(\boldsymbol{\theta},\delta) = \int \Phi\left(\frac{\boldsymbol{\theta}^\top \mathbf{x} - \delta}{h}\right) dF_{\bar{D}}(\mathbf{x}) - P(\boldsymbol{\theta}^\top \mathbf{X} > \delta \mid D = 0).$$

For a general random variable V with distribution function F where F is Lipschitz, consider

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

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

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

As in Winter (1979), let $M \in \mathbb{R}$ such that for all $v, t \in \mathbb{R}$, $|F(v-t) - F(v)| \leq M|t|$. Then

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

Since h is a function of n such that $h \to 0$ as $n \to \infty$, this gives

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

Returning now to $\boldsymbol{\theta}^\top \mathbf{X}$: consider the case where $p = 2$, so $\boldsymbol{\theta}^\top \mathbf{X} = \theta_1 X_1 + \theta_2 X_2$. Let $Y_1 = \theta_1 X_1 + \theta_2 X_2$, $Y_2 = \theta_2 X_2$. Then $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|\theta_1\theta_2|^{-1}$. Thus, for any $s \in \mathbb{R}$, we have

$$\int \Phi\left(\frac{s - \boldsymbol{\theta}^\top \mathbf{x}}{h}\right) f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \int \Phi\left(\frac{s-y_1}{h}\right) f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} = \int \Phi\left(\frac{s-y_1}{h}\right) f_{Y_1}(y_1)dy_1.$$

Since $P(\boldsymbol{\theta}^\top \mathbf{X} \leq \delta \mid D = 0) = P(Y_1 \leq \delta \mid D = 0)$, we can write

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| \int \Phi\left(\frac{\boldsymbol{\theta}^\top \mathbf{x} - \delta}{h}\right) dF_{\bar{D}}(\mathbf{x}) - P(\boldsymbol{\theta}^\top \mathbf{X} > \delta \mid D = 0) \right| \\
&= \sup_{\delta\in\mathbb{R}} \left| \int \Phi\left(\frac{y_1 - \delta}{h}\right) f_{Y_1|\bar{D}}(y_1) dy_1 - P(Y_1 > \delta \mid D = 0) \right| \\
&= \sup_{\delta\in\mathbb{R}} \left| \int \Phi\left(\frac{\delta - y_1}{h}\right) f_{Y_1|\bar{D}}(y_1) dy_1 - P(Y_1 \leq \delta \mid D = 0) \right|,
\end{aligned}$$

giving, by condition (A5) in Section 2.3.2 and the results above,

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

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

Combining these results, we have $\sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| \tilde{FPR}_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) - FPR(\boldsymbol{\theta}, \delta) \right| \to 0$ w.p. 1 as claimed. □

A.1.4 Proof of Theorem 2.1

Proof. First consider the claim $FPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \leq t + o_p(1)$. By Lemma 2.2, we have

$$\sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) - FPR(\boldsymbol{\theta}, \delta) \right| \to 0 \text{ w.p.1.}$$

In particular,

$$\begin{aligned}
FPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) &= F\tilde{P}R_{n_{\bar{D}}}(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) + \left[FPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) - F\tilde{P}R_{n_{\bar{D}}}(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \right] \\
&\leq F\tilde{P}R_{n_{\bar{D}}}(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) + \left| FPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) - F\tilde{P}R_{n_{\bar{D}}}(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \right| \\
&\leq F\tilde{P}R_{n_{\bar{D}}}(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) + \sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| FPR(\boldsymbol{\theta}, \delta) - F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) \right| \\
&\leq t + \sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| FPR(\boldsymbol{\theta}, \delta) - F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta) \right| \\
&= t + o_p(1),
\end{aligned}$$

giving the desired result.

Now consider the claim

$$\left| \max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta}, \delta) - TPR(\hat{\boldsymbol{\theta}}_t, \hat{\delta}_t) \right| \to 0 \text{ w.p. 1.}$$

Let $\omega = (\boldsymbol{\theta}, \delta)$. We first demonstrate that Lemma 2.1 holds for

$$\begin{aligned}
f(\omega) &:= TPR(\boldsymbol{\theta}, \delta) \\
\Omega_n &:= \Omega_{t,n_{\bar{D}}} \\
\Omega_0 &:= \Omega_{t,0}.
\end{aligned}$$

By condition (A4) in Section 2.3.2, $\omega_n \in \Omega_{t,n_{\bar{D}}}$ such that $f(\omega_n) = \max_{\omega\in\Omega_{t,n_{\bar{D}}}} f(\omega)$ and $\omega_0 \in \Omega_{t,0}$ such that $f(\omega_0) = \max_{\omega\in\Omega_{t,0}} f(\omega)$ exist. By condition (A6) in Section 2.3.2, $f(\omega)$

is Lipschitz. Thus, we must show $d_{n_{\bar{D}}}, e_{n_{\bar{D}}} \to 0$ w.p. 1, where, as in Lemma 2.1,

$$d_{n_{\bar{D}}} := \sup_{\omega \in \Omega_{t,n_{\bar{D}}}} \inf_{\tilde{\omega} \in \Omega_{t,0}} d(\omega, \tilde{\omega}), \quad e_{n_{\bar{D}}} := \sup_{\omega \in \Omega_{t,0}} \inf_{\tilde{\omega} \in \Omega_{t,n_{\bar{D}}}} d(\omega, \tilde{\omega}),$$

and d is the Euclidean distance.

Consider $d_{n_{\bar{D}}}$. Suppose for some $\kappa > 0$,

$$\sup_{\omega \in \Omega_{t,n_{\bar{D}}}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \leq \kappa,$$

where we abuse notation somewhat to allow $FPR(\omega) = FPR(\boldsymbol{\theta}, \delta)$ and $F\tilde{P}R_{n_{\bar{D}}}(\omega) = F\tilde{P}R_{n_{\bar{D}}}(\boldsymbol{\theta}, \delta)$. Then

$$\kappa \geq \sup_{\omega \in \Omega_{t,n_{\bar{D}}}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \geq \left| \sup_{\omega \in \Omega_{t,n_{\bar{D}}}} F\tilde{P}R_{n_{\bar{D}}}(\omega) - \sup_{\omega \in \Omega_{t,n_{\bar{D}}}} FPR(\omega) \right|,$$

so $\sup_{\omega \in \Omega_{t,n_{\bar{D}}}} FPR(\omega) \leq \kappa + t$.

For every $\omega \in \Omega_{t,n_{\bar{D}}}$, we have $F\tilde{P}R_{n_{\bar{D}}}(\omega) \leq t$ and $FPR(\omega) \leq \kappa + t$. Then for any given $\omega = (\boldsymbol{\theta}, \delta) \in \Omega_{t,n_{\bar{D}}}$, we can write

$$FPR(\omega) = P(\boldsymbol{\theta}^\top \mathbf{X} > \delta | D = 0) = t_2 \leq \kappa + t.$$

If $t_2 > t$ (so $\omega \notin \Omega_{t,0}$), there exists $\tilde{\delta}$ such that

$$P(\boldsymbol{\theta}^\top \mathbf{X} > \tilde{\delta} | D = 0) = t,$$

namely, $\tilde{\delta} = G^{-1}(1 - t)$, where $G(\cdot)$ is the distribution function of $(\boldsymbol{\theta}^\top \mathbf{X} | D = 0)$. This gives $\tilde{\omega} = (\boldsymbol{\theta}, \tilde{\delta}) \in \Omega_{t,0}$. Then since $G^{-1}(\cdot)$ is Lipschitz with constant C by condition (A5) in Section 2.3.2, we have

$$d(\omega, \tilde{\omega}) = |\delta - \tilde{\delta}| = |G^{-1}(1 - t_2) - G^{-1}(1 - t)| \leq C|t - t_2| \leq C\kappa.$$

Thus, for any given $\omega \in \Omega_{t,n_{\bar{D}}}$,

$$\inf_{\tilde{\omega} \in \Omega_{t,0}} d(\omega, \tilde{\omega}) \leq C\kappa,$$

so $d_{n_{\bar{D}}} \leq C\kappa$. Therefore, if $d_{n_{\bar{D}}} > \epsilon$ for some $\epsilon > 0$, then

$$\sup_{\omega \in \Omega_{t,n_{\bar{D}}}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| > \kappa_\epsilon = \epsilon/C.$$

This gives

$$P\left(\sup_{m \geq n_{\bar{D}}} d_m > \epsilon\right) \leq P\left(\sup_{m \geq n_{\bar{D}}} \sup_{\omega \in \Omega_{t,m}} |F\tilde{P}R_m(\omega) - FPR(\omega)| > \kappa_\epsilon\right) \to 0$$

since Lemma 2.2 and $\Omega_{t,n_{\bar{D}}} \subseteq \Omega$ give $\sup_{\omega \in \Omega_{t,n_{\bar{D}}}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \to 0$ w.p. 1.

Now consider $e_{n_{\bar{D}}}$. Suppose for some $\kappa \in (0, t)$

$$\sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \leq \kappa.$$

Then

$$\kappa \geq \sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \geq \left| \sup_{\omega \in \Omega_{t,0}} F\tilde{P}R_{n_{\bar{D}}}(\omega) - \sup_{\omega \in \Omega_{t,0}} FPR(\omega) \right|,$$

so $\sup_{\omega \in \Omega_{t,0}} F\tilde{P}R_{n_{\bar{D}}}(\omega) \leq \kappa + t$. Thus, for a given $\omega = (\boldsymbol{\theta}, \delta) \in \Omega_{t,0}$, $FPR(\omega) \leq t$ and $F\tilde{P}R_{n_{\bar{D}}}(\omega) \leq t + \kappa$. Suppose $\omega \notin \Omega_{t,n_{\bar{D}}}$, so $t < F\tilde{P}R_{n_{\bar{D}}}(\omega) \leq t + \kappa$. Since $\kappa \geq \sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)|$, $FPR(\omega) > t - \kappa$. Then

$$H^{-1}(1 + \kappa - t) > \delta \geq H^{-1}(1 - t),$$

where H is the distribution function of $(\boldsymbol{\theta}^\top \mathbf{X} | D = 0)$. Let $\tilde{\delta} = H^{-1}(1 + \kappa - t)$ and $\tilde{\omega} = (\boldsymbol{\theta}, \tilde{\delta})$. Since

$$\kappa \geq \sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)|,$$

and $FPR(\tilde{\omega}) = t - \kappa$ (so $\tilde{\omega} \in \Omega_{t,0}$), we have $F\tilde{P}R_{n_{\bar{D}}}(\tilde{\omega}) \leq t$, giving $\tilde{\omega} \in \Omega_{t,n_{\bar{D}}}$. Then since $H^{-1}(\cdot)$ is Lipschitz with constant B by condition (A5) in Section 2.3.2,

$$d(\omega, \tilde{\omega}) = |\delta - \tilde{\delta}| \leq |H^{-1}(1-t) - H^{-1}(1+\kappa-t)| \leq B\kappa.$$

Thus, for any given $\omega \in \Omega_{t,0}$,

$$\inf_{\tilde{\omega} \in \Omega_{t,n_{\bar{D}}}} d(\omega, \tilde{\omega}) \leq B\kappa,$$

and $e_{n_{\bar{D}}} \leq B\kappa$. Therefore, if $e_{n_{\bar{D}}} > \epsilon$ for some $\epsilon > 0$, then

$$\sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| > \kappa_\epsilon = \epsilon / B.$$

We then have

$$P\left(\sup_{m \geq n_{\bar{D}}} e_m > \epsilon\right) \leq P\left(\sup_{m \geq n_{\bar{D}}} \sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_m(\omega) - FPR(\omega)| > \kappa_\epsilon\right) \to 0$$

since Lemma 2.2 and $\Omega_{t,0} \subseteq \Omega$ give $\sup_{\omega \in \Omega_{t,0}} |F\tilde{P}R_{n_{\bar{D}}}(\omega) - FPR(\omega)| \to 0$ w.p. 1.

By Lemma 2.1, we now have

$$\left| \sup_{(\boldsymbol{\theta}, \delta) \in \Omega_{t,0}} TPR(\boldsymbol{\theta}, \delta) - \sup_{(\boldsymbol{\theta}, \delta) \in \Omega_{t,n_{\bar{D}}}} TPR(\boldsymbol{\theta}, \delta) \right| \to 0 \text{ w.p. } 1.$$

Then by Lemmas 2.1 and 2.2,

$$\begin{aligned}
&\left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \\
&\qquad \leq \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} TPR(\boldsymbol{\theta},\delta) \right| \\
&\qquad\quad + \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \\
&\qquad \leq \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} TPR(\boldsymbol{\theta},\delta) \right| \\
&\qquad\quad + \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \left| TPR(\boldsymbol{\theta},\delta) - \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \\
&\qquad \to 0 \text{ w.p. } 1.
\end{aligned}$$

By Lemma 2.2,

$$\left| TPR(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) - \tilde{TPR}_{n_D}(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) \right| \leq \sup_{(\boldsymbol{\theta},\delta)\in\Omega} \left| TPR(\boldsymbol{\theta},\delta) - \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \to 0 \text{ w.p. } 1,$$

and by condition (A4) in Section 2.3.2 and equation (2.1), $\max_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) = \tilde{TPR}_{n_D}(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t)$, giving

$$\begin{aligned}
\left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - TPR(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) \right| &\leq \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \\
&\quad + \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) - TPR(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) \right| \\
&= \left| \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,0}} TPR(\boldsymbol{\theta},\delta) - \sup_{(\boldsymbol{\theta},\delta)\in\Omega_{t,n_{\bar{D}}}} \tilde{TPR}_{n_D}(\boldsymbol{\theta},\delta) \right| \\
&\quad + \left| \tilde{TPR}_{n_D}(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) - TPR(\hat{\boldsymbol{\theta}}_t,\hat{\delta}_t) \right| \\
&\to 0 \text{ w.p. } 1,
\end{aligned}$$

completing the proof. □

A.2 Chapter 4

A.2.1 Risk Function for Conditionally Bivariate Normal Biomarkers

Claim: If the biomarkers X_1 and X_2 have the conditional distribution given by (4.7), the true risk function is given by

$$\text{logit } \{P(D=1|X_1,X_2,C=c)\} = \beta_0^c + \beta_1 X_1 + \beta_2 X_2,$$

where

$$\begin{aligned}
\beta_0^c &= \frac{-\mu_{X_1}^2 - \mu_{X_2}^2}{2(1-\rho^2)} + \frac{\rho\mu_{X_1}\mu_{X_2} + \rho\mu_{X_1}f_{X_2}(c) + \rho\mu_{X_2}f_{X_1}(c) - \mu_{X_1}f_{X_1}(c) - \mu_{X_2}f_{X_2}(c)}{1-\rho^2} \\
&\quad + \log\left(\frac{\gamma_c}{1-\gamma_c}\right) \\
\beta_1 &= \frac{\mu_{X_1} - \rho\mu_{X_2}}{1-\rho^2} \\
\beta_2 &= \frac{\mu_{X_2} - \rho\mu_{X_1}}{1-\rho^2}.
\end{aligned}$$

Proof. We can demonstrate this as follows:

$$\begin{aligned}
P(D=1|X_1,X_2,C=c) &= \frac{f(X_1,X_2|D=1,C=c)\gamma_c}{f(X_1,X_2|D=1,C=c)\gamma_c + f(X_1,X_2|D=0,C=c)(1-\gamma_c)} \\
&= \frac{1}{1+B/A},
\end{aligned}$$

where $A = f(X_1,X_2|D=1,C=c)\gamma_c$ and $B = f(X_1,X_2|D=0,C=c)(1-\gamma_c)$. We have

$$
\begin{aligned}
\frac{B}{A} &= \frac{f(X_1, X_2|D=0, C=c)(1-\gamma_c)}{f(X_1, X_2|D=1, C=c)\gamma_c} \\
&= \exp\Bigg(\frac{1}{2(1-\rho^2)} \big[\{X_1 - \mu_{X_1} - f_{X_1}(c)\}^2 - \{X_1 - f_{X_1}(c)\}^2 \quad \times \left(\frac{1-\gamma_c}{\gamma_c}\right) \\
&\qquad - 2\rho\{X_1 - \mu_{X_1} - f_{X_1}(c)\}\{X_2 - \mu_{X_2} - f_{X_2}(c)\} \\
&\qquad\qquad + 2\rho\{X_1 - f_{X_1}(c)\}\{X_2 - f_{X_2}(c)\} \\
&\qquad\qquad + \{X_2 - \mu_{X_2} - f_{X_2}(c)\}^2 - \{X_2 - f_{X_2}(c)\}^2 \big] \Bigg) \\
&= \exp\Bigg\{ \frac{1}{2(1-\rho^2)} \big(\mu_{X_1}^2 - 2\mu_{X_1}\{X_1 - f_{X_1}(c)\} + 2\rho[-\mu_{X_1}\mu_{X_2} + \mu_{X_1}\{X_2 - f_{X_2}(c)\} + \\
&\qquad\qquad \mu_{X_2}\{X_1 - f_{X_1}(c)\}] + \mu_{X_2}^2 - 2\mu_{X_2}\{X_2 - f_{X_2}(c)\} \big) + \log\left(\frac{1-\gamma_c}{\gamma_c}\right) \Bigg\}.
\end{aligned}
$$

If $P(D=1|X_1,\, X_2,\, C=c) = \dfrac{1}{1+\exp(\star)}$ then $P(D=1|X_1, X_2, C=c) = \text{expit}(-\star)$, so

$$
\begin{aligned}
&P(D=1|X_1, X_2, C=c) \\
&\qquad = \text{expit}\Bigg\{ \frac{-1}{2(1-\rho^2)} \big(\mu_{X_1}^2 - 2\mu_{X_1}\{X_1 - f_{X_1}(c)\} \\
&\qquad\qquad + 2\rho[-\mu_{X_1}\mu_{X_2} + \mu_{X_1}\{X_2 - f_{X_2}(c)\} + \mu_{X_2}\{X_1 - f_{X_1}(c)\}] \\
&\qquad\qquad\qquad + \mu_{X_2}^2 - 2\mu_{X_2}\{X_2 - f_{X_2}(c)\} \big) - \log\left(\frac{1-\gamma_c}{\gamma_c}\right) \Bigg\} \\
&\qquad = \text{expit}\Bigg\{ \frac{-\mu_{X_1}^2 - \mu_{X_2}^2}{2(1-\rho^2)} + \log\left(\frac{\gamma_c}{1-\gamma_c}\right) + \frac{\mu_{X_1} - \rho\mu_{X_2}}{1-\rho^2} X_1 + \frac{\mu_{X_2} - \rho\mu_{X_1}}{1-\rho^2} X_2 \\
&\qquad\qquad + \frac{\rho\mu_{X_1}\mu_{X_2} + \rho\mu_{X_1} f_{X_2}(c) + \rho\mu_{X_2} f_{X_1}(c) - \mu_{X_1} f_{X_1}(c) - \mu_{X_2} f_{X_2}(c)}{1-\rho^2} \Bigg\} \\
&\qquad = \text{expit}(\beta_0^c + \beta_1 X_1 + \beta_2 X_2).
\end{aligned}
$$

□

A.2.2 Center-Specific AUC for Conditionally Bivariate Normal Biomarkers

Claim: If the biomarkers X_1 and X_2 have the conditional distribution given by (4.7), $AUC_c(\boldsymbol{\theta})$ for a generic combinations $\boldsymbol{\theta}$ does not vary with c and the center-specific ROC curves for $\boldsymbol{\theta}^\top \mathbf{X}$ are concave.

Proof. Let X_{1D} denote X_1 for an arbitrary case and $X_{1\bar{D}}$ denote X_1 for an arbitrary control (and define X_{2D} and $X_{2\bar{D}}$ analogously). The $AUC_c(\boldsymbol{\theta})$ for a generic $\boldsymbol{\theta}$ can be written as:

$$\begin{aligned} AUC_c(\boldsymbol{\theta}) &= P(\theta_1 X_{1D} + \theta_2 X_{2D} > \theta_1 X_{1\bar{D}} + \theta_2 X_{2\bar{D}} C = c) \\ &= P\left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}^\top \begin{pmatrix} X_{1D} - X_{1\bar{D}} \\ X_{2D} - X_{2\bar{D}} \end{pmatrix} > 0 \,\middle|\, C = c \right), \end{aligned}$$

where the distribution of $\begin{pmatrix} X_{1D} - X_{1\bar{D}} \\ X_{2D} - X_{2\bar{D}} \end{pmatrix} \,\middle|\, C = c \Bigg)$ is constant across centers under (4.7). Since $\theta_1 X_{1D} + \theta_2 X_{2D}$ and $\theta_1 X_{1\bar{D}} + \theta_2 X_{2\bar{D}}$ are independently normally distributed (conditional on $C = c$) with

$$\begin{aligned} (\theta_1 X_{1D} + \theta_2 X_{2D} | C = c) &\sim N(\theta_1\{\mu_{X_1} + f_{X_1}(c)\} + \theta_2\{\mu_{X_2} + f_{X_2}(c)\}, \theta_1^2 + \theta_2^2 + 2\rho\theta_1\theta_2) \\ (\theta_1 X_{1\bar{D}} + \theta_2 X_{2\bar{D}} | C = c) &\sim N(\theta_1 f_{X_1}(c) + \theta_2 f_{X_2}(c), \theta_1^2 + \theta_2^2 + 2\rho\theta_1\theta_2), \end{aligned}$$

the center-specific AUC is a function of the variances of $(\theta_1 X_{1D} + \theta_2 X_{2D} | C = c)$ and $(\theta_1 X_{1\bar{D}} + \theta_2 X_{2\bar{D}} | C = c)$ and the difference in the means of $(\theta_1 X_{1D} + \theta_2 X_{2D} | C = c)$ and $(\theta_1 X_{1\bar{D}} + \theta_2 X_{2\bar{D}} | C = c)$ (Pepe, 2003). The difference in the means is $\theta_1 \mu_{X_1} + \theta_2 \mu_{X_2}$, and the variances also do not depend on center. Thus, $AUC_c(\boldsymbol{\theta})$ does not vary with c. Furthermore, since within each center center, the distribution of the combination given D is normally distributed with equal variance for cases and controls, the center-specific ROC curves for the combination are concave (Pepe, 2003). □

A.2.3 Proof of Lemma 4.1

Proof. Previous work by Han (1987) proved a similar claim for a related statistic. Consider a single center with n total observations, n_D cases, and $n_{\bar{D}}$ controls. Let $\mathbf{X}_i$ denote the biomarker vector for observation i. Han considered the statistic $h_{ij}(\boldsymbol{\theta})$:.for any pair of observations (i, j), let

$$h_{ij}(\boldsymbol{\theta}) = 1(D_i > D_j)1(\boldsymbol{\theta}\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j) + 1(D_i < D_j)1(\boldsymbol{\theta}^\top\mathbf{X}_i < \boldsymbol{\theta}^\top\mathbf{X}_j).$$

Then

$$\begin{aligned}
\sum_{\kappa} h_{ij}(\boldsymbol{\theta}) &= \sum_{i<j} h_{ij}(\boldsymbol{\theta}) \\
&= \frac{1}{2}\sum_{i\neq j} 1\{(D_i - D_j)(\boldsymbol{\theta}^\top\mathbf{X}_i - \boldsymbol{\theta}^\top\mathbf{X}_j) > 0\} \\
&= \sum_{i\neq j} 1(D_i > D_j)1(\boldsymbol{\theta}^\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j) \\
&= \sum_{i:D_i=1, j:D_j=0} 1(\boldsymbol{\theta}^\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j),
\end{aligned}$$

where κ denotes the collection of all possible pairs of distinct elements (regardless of D). Also, for any $i \neq j$,

$$\begin{aligned}
P(\boldsymbol{\theta}^\top\mathbf{X}_i < \boldsymbol{\theta}^\top\mathbf{X}_j, D_i < D_j) &+ P(\boldsymbol{\theta}^\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j, D_i > D_j) \\
&= P(\boldsymbol{\theta}^\top\mathbf{X}_i < \boldsymbol{\theta}^\top\mathbf{X}_j | D_i < D_j)P(D_i < D_j) \\
&\quad + P(\boldsymbol{\theta}^\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j | D_i > D_j)P(D_i > D_j) \\
&= P(\boldsymbol{\theta}^\top\mathbf{X}_i < \boldsymbol{\theta}^\top\mathbf{X}_j | D_i < D_j)\gamma(1-\gamma) \\
&\quad + P(\boldsymbol{\theta}^\top\mathbf{X}_i > \boldsymbol{\theta}^\top\mathbf{X}_j | D_i > D_j)\gamma(1-\gamma) \\
&= 2\gamma(1-\gamma)AUC(\boldsymbol{\theta}),
\end{aligned}$$

where γ is the prevalence.

We can consider

$$S_n(\boldsymbol{\theta}) = \frac{2}{n(n-1)} \sum_{\kappa} h_{ij}(\boldsymbol{\theta})$$
$$\hat{AUC}(\boldsymbol{\theta}) = \frac{1}{n_D n_{\bar{D}}} \sum_{i:D_i=1, j:D_j=0} 1(\boldsymbol{\theta}^\top \mathbf{X}_i > \boldsymbol{\theta}^\top \mathbf{X}_j),$$

where S_n is a one-sample U-statistic and $\hat{AUC}$ is a two-sample U-statistic. We would like to be able to study the asymptotic behavior of S_n; in particular, we would like to say that if

$$\sup_{\boldsymbol{\theta} \in B} |S_n(\boldsymbol{\theta}) - 2\gamma(1-\gamma) AUC(\boldsymbol{\theta})| = o_p(1)$$

then $\sup_{\boldsymbol{\theta} \in B} \left| \hat{AUC}(\boldsymbol{\theta}) - AUC(\boldsymbol{\theta}) \right| = o_p(1)$.

If $\sup_{\boldsymbol{\theta} \in B} |S_n(\boldsymbol{\theta}) - 2\gamma(1-\gamma) AUC(\boldsymbol{\theta})| = o_p(1)$, then

$$\sup_{\boldsymbol{\theta} \in B} \left| \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} - AUC(\boldsymbol{\theta}) \right| = o_p(1).$$

We would then have

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in B} \left| \hat{AUC}(\boldsymbol{\theta}) - AUC(\boldsymbol{\theta}) \right| &\leq \sup_{\boldsymbol{\theta} \in B} \left| \hat{AUC}(\boldsymbol{\theta}) - \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} \right| + \sup_{\boldsymbol{\theta} \in B} \left| \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} - AUC(\boldsymbol{\theta}) \right| \\
&= \sup_{\boldsymbol{\theta} \in B} \left| \hat{AUC}(\boldsymbol{\theta}) - \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} \right| + o_p(1) \\
&\leq \sup_{\boldsymbol{\theta} \in B} \left| \hat{AUC}(\boldsymbol{\theta}) - \frac{S_n(\boldsymbol{\theta}) n^2}{2 n_D n_{\bar{D}}} \right| + \sup_{\boldsymbol{\theta} \in B} \left| \frac{S_n(\boldsymbol{\theta}) n^2}{2 n_D n_{\bar{D}}} - \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} \right| + o_p(1) \\
&= \left| 1 - \frac{n^2}{n(n-1)} \right| \sup_{\boldsymbol{\theta} \in B} \hat{AUC}(\boldsymbol{\theta}) + \sup_{\boldsymbol{\theta} \in B} \left| \frac{S_n(\boldsymbol{\theta}) n^2}{2 n_D n_{\bar{D}}} - \frac{S_n(\boldsymbol{\theta})}{2\gamma(1-\gamma)} \right| + o_p(1) \\
&\leq o(1) + \left| \frac{n^2}{2 n_D n_{\bar{D}}} - \frac{1}{2\gamma(1-\gamma)} \right| + o_p(1) \\
&= o_p(1),
\end{aligned}$$

where the last inequality follows from the fact that $S_n(\boldsymbol{\theta}), A\hat{U}C(\boldsymbol{\theta}) \leq 1$ and the last equality follows from the Weak Law of Large Numbers, the continuous mapping theorem and Slutsky's theorem. Thus, to demonstrate uniform convergence of $A\hat{U}C(\boldsymbol{\theta})$, we can consider $S_n(\boldsymbol{\theta})$.

The following is taken nearly verbatim (with very minor variations) from part of the proof given in Han (1987). Let

$$\begin{aligned}
h(\boldsymbol{\theta}) &= E\{h_{ij}(\boldsymbol{\theta})\} \equiv 2\gamma(1-\gamma)AUC(\boldsymbol{\theta}) \\
\bar{g}_{ij}(\boldsymbol{\theta}, \delta) &= \sup_{b \in D_\delta(\boldsymbol{\theta})} \{h_{ij}(b) - h(b)\} \\
\underline{g}_{ij}(\boldsymbol{\theta}, \delta) &= \inf_{b \in D_\delta(\boldsymbol{\theta})} \{h_{ij}(b) - h(b)\} \\
\bar{g}(\boldsymbol{\theta}, \delta) &= E\{\bar{g}_{ij}(\boldsymbol{\theta}, \delta)\} \\
\underline{g}(\boldsymbol{\theta}, \delta) &= E\{\underline{g}_{ij}(\boldsymbol{\theta}, \delta)\}
\end{aligned}$$

where $D_\delta(\boldsymbol{\theta}) = \{b : b \in B, ||b - \boldsymbol{\theta}|| < \delta\}$.

It can be seen that $h_{ij}(\boldsymbol{\theta})$ is a step function uniformly bounded in (i, j) and $\boldsymbol{\theta}$. By condition (C3) in Section 4.3.3, $h_{ij}(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in B$ uniformly across (i, j) almost surely. Thus, $h(\boldsymbol{\theta})$ is uniformly bounded and continuous in $\boldsymbol{\theta} \in B$.

Note also that $\bar{g}_{ij}(\boldsymbol{\theta}, \delta)$ is measurable for all $\boldsymbol{\theta} \in B$ and $\delta > 0$ since B is separable and for any $\boldsymbol{\theta} \in B$ there exists a sequence $\{\boldsymbol{\theta}_t\}$ in a countable dense subset of B such that

$$\lim_{t \to \infty} h_{ij}(\boldsymbol{\theta}_t) = h_{ij}(\boldsymbol{\theta}), \quad \lim_{t \to \infty} h(\boldsymbol{\theta}_t) = h(\boldsymbol{\theta}).$$

Also, $\bar{g}_{ij}(\boldsymbol{\theta}, \delta)$ is uniformly bounded in $\boldsymbol{\theta}$ and

$$\lim_{\delta \to 0} \bar{g}_{ij}(\boldsymbol{\theta}, \delta) = h_{ij}(\boldsymbol{\theta}) - h(\boldsymbol{\theta}) \text{ almost surely.}$$

Thus, it follows that $\lim_{\delta \to 0} \bar{g}(\boldsymbol{\theta}, \delta) = 0$ for all $\boldsymbol{\theta} \in B$. A similar argument can be made for $\underline{g}(\boldsymbol{\theta}, \delta)$, giving $\lim_{\delta \to 0} \underline{g}(\boldsymbol{\theta}, \delta) = 0$ for all $\boldsymbol{\theta} \in B$.

To show convergence of $S_n(\boldsymbol{\theta})$ to $h(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}$, we have for a given $\epsilon > 0$,

$$P\left(\sup_{\boldsymbol{\theta}\in B}|S_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| > \epsilon\right) \leq P\left(\left|\frac{2}{n(n-1)}\sum_{\kappa}\sup_{\boldsymbol{\theta}\in B}\{h_{ij}(\boldsymbol{\theta}) - h(\boldsymbol{\theta})\}\right| > \epsilon\right) + P\left(\left|\frac{2}{n(n-1)}\sum_{\kappa}\inf_{\boldsymbol{\theta}\in B}\{h_{ij}(\boldsymbol{\theta}) - h(\boldsymbol{\theta})\}\right| > \epsilon\right).$$

Since B is compact, there exists a finite set of coverings $\{D_{\delta_l}(\boldsymbol{\theta}_l)\}$, $l = 1, ..., L$, such that

$$B \subset \bigcup_{l=1}^{L} D_{\delta_l}(\boldsymbol{\theta}_l) \text{ and } \bar{g}(\boldsymbol{\theta}_l, \delta_l), \underline{g}(\boldsymbol{\theta}_l, \delta_l) > \epsilon/2.$$

Thus,

$$P\left(\sup_{\boldsymbol{\theta}\in B}|S_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| > \epsilon\right) \leq \sum_{l=1}^{L} P\left(\left|\frac{2}{n(n-1)}\sum_{\kappa}\bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) - \bar{g}(\boldsymbol{\theta}_l, \delta_l)\right| \geq \epsilon/2\right) + \sum_{l=1}^{L} P\left(\left|\frac{2}{n(n-1)}\sum_{\kappa}\underline{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) - \underline{g}(\boldsymbol{\theta}_l, \delta_l)\right| \geq \epsilon/2\right).$$

However, note that for each l, $\frac{2}{n(n-1)}\sum_{\kappa}\bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l)$ is a one-sample U-statistic with kernel

$$\bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) = \sup_{\mathbf{b}\in D_{\delta_l}(\boldsymbol{\theta}_l)}\{h_{ij}(\mathbf{b}) - h(\mathbf{b})\}.$$

Since $E\{\bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l)\}^2 \leq E(1) < \infty$, we have

$$\frac{2}{n(n-1)}\sum_{\kappa}\bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) \xrightarrow{p} \bar{g}(\boldsymbol{\theta}_l, \delta_l).$$

These arguments also apply to $\underline{g}_{ij}(\boldsymbol{\theta}_l, \delta_l)$, giving

$$\frac{2}{n(n-1)} \sum_{\kappa} \bar{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) \xrightarrow{p} \bar{g}(\boldsymbol{\theta}_l, \delta_l), \quad l = 1, ..., L$$
$$\frac{2}{n(n-1)} \sum_{\kappa} \underline{g}_{ij}(\boldsymbol{\theta}_l, \delta_l) \xrightarrow{p} \underline{g}(\boldsymbol{\theta}_l, \delta_l), \quad l = 1, ..., L.$$

Thus, we have

$$P\left(\sup_{\boldsymbol{\theta} \in B} |S_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| > \epsilon\right) \to 0,$$

so $\sup_{\boldsymbol{\theta} \in B} |\hat{AUC}(\boldsymbol{\theta}) - AUC(\boldsymbol{\theta})| \xrightarrow{p} 0$. Since this holds for any center c, we have demonstrated $\sup_{\boldsymbol{\theta} \in B} |\hat{AUC}_c(\boldsymbol{\theta}) - AUC_c(\boldsymbol{\theta})| \xrightarrow{p} 0$ as $n_c \to \infty$. □

A.2.4 Proof of Lemma 4.2

Proof. We can write this claim as follows:

$$\sup_{\boldsymbol{\theta}\in B}|a\hat{AU}C(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})| = \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - \sum_{c=1}^{M} w_c AUC_c(\boldsymbol{\theta})\right| \xrightarrow{p} 0.$$

We can write

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - \sum_{c=1}^{M} w_c AUC_c(\boldsymbol{\theta})\right| \\
&\quad= \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - \sum_{c=1}^{m} w_c AUC_c(\boldsymbol{\theta}) - \sum_{c=m+1}^{M} w_c AUC_c(\boldsymbol{\theta})\right| \\
&\quad= \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\Big(\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - w_c AUC_c(\boldsymbol{\theta})\Big) - \sum_{c=m+1}^{M} w_c AUC_c(\boldsymbol{\theta})\right| \\
&\quad\leq \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\Big(\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - w_c AUC_c(\boldsymbol{\theta})\Big)\right| + \sum_{c=m+1}^{M}\sup_{\boldsymbol{\theta}\in B}|w_c AUC_c(\boldsymbol{\theta})| \\
&\quad= \sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\Big(\hat{w}_c A\hat{U}C_c(\boldsymbol{\theta}) - w_c AUC_c(\boldsymbol{\theta})\Big)\right| + o(1)
\end{aligned}$$

as $m \to M$.

Then

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\left(\hat{w}_c\hat{AUC}_c(\boldsymbol{\theta})-w_cAUC_c(\boldsymbol{\theta})\right)\right| \\
&\quad=\sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\left\{\hat{w}_c\left(\hat{AUC}_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right)+\hat{w}_cAUC_c(\boldsymbol{\theta})-w_cAUC_c(\boldsymbol{\theta})\right\}\right| \\
&\quad\leq\sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\hat{w}_c\left(\hat{AUC}_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right)\right|+\sup_{\boldsymbol{\theta}\in B}\left|\sum_{c=1}^{m}\left(\hat{w}_c-w_c\right)AUC_c(\boldsymbol{\theta})\right| \\
&\quad\leq\sum_{c=1}^{m}\sup_{\boldsymbol{\theta}\in B}\left|\hat{w}_c\left(\hat{AUC}_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right)\right|+\sum_{c=1}^{m}\sup_{\boldsymbol{\theta}\in B}\left|\left(\hat{w}_c-w_c\right)AUC_c(\boldsymbol{\theta})\right| \\
&\quad=\sum_{c=1}^{m}\hat{w}_c\sup_{\boldsymbol{\theta}\in B}\left|\hat{AUC}_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right|+\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|\sup_{\boldsymbol{\theta}\in B}AUC_c(\boldsymbol{\theta}) \\
&\quad=\sum_{c=1}^{m}\hat{w}_co_p(1)+\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|\sup_{\boldsymbol{\theta}\in B}AUC_c(\boldsymbol{\theta}) \\
&\quad=o_p(1)+\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|\sup_{\boldsymbol{\theta}\in B}AUC_c(\boldsymbol{\theta}) \\
&\quad\leq o_p(1)+\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|,
\end{aligned}
$$

where the second to last equality follows from the fact that $\sup_{\boldsymbol{\theta}\in B}\left|\hat{AUC}_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right|=o_p(1)$ by Lemma 4.1, the last equality follows from the fact that $\sum_{c=1}^{m}\hat{w}_c=1$ for every m, and the last inequality follows from $AUC_c(\boldsymbol{\theta})\leq 1$.

Now we must show that

$$\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|=o_p(1).$$

Equivalently, we must prove that for every $\epsilon>0$,

$$P\left(\sum_{c=1}^{m}\left|\hat{w}_c-w_c\right|>\epsilon\right)\to 0$$

as $n_c\to\infty$, $c=1,...,m$, and $m\to M$ such that $\sqrt{n_c}/m\to\infty$.

We have

$$P\left(\sum_{c=1}^{m}|\hat{w}_c - w_c| > \epsilon\right) \leq \frac{E(\sum_{c=1}^{m}|\hat{w}_c - w_c|)}{\epsilon} = \frac{\sum_{c=1}^{m}E|\hat{w}_c - w_c|}{\epsilon}$$
$$\leq \frac{\sum_{c=1}^{m}E\left|\hat{w}_c - E(\hat{w}_c)\right|}{\epsilon} + \frac{\sum_{c=1}^{m}|E(\hat{w}_c) - w_c|}{\epsilon}$$
$$\leq \frac{\sum_{c=1}^{m}\sqrt{Var(\hat{w}_c)}}{\epsilon} + o(1),$$

where the first inequality follows from Markov's inequality and the third inequality follows from Jensen's inequality and the fact that $\sum_{c=1}^{m}|E(\hat{w}_c) - w_c| = o(1)$ by condition (C2) in Section 4.3.3.

Let $f(R,S) = R/S$ and $\nu = (E(R), E(S))$. Then we can use a first-order Taylor approximation to write

$$f(R,S) \approx f(\nu) + f'_R(\nu)(R - E(R)) + f'_S(\nu)(S - E(S)),$$

where $f'_R(\nu) = \frac{\partial f(R,S)}{\partial R}|_\nu$ and $f'_S(\nu)$ is defined analogously. This gives $E[f(R,S)] \approx f(\nu)$. We can also write

$$Var[f(R,S)] \approx E\left[\{f(R,S) - f(\nu)\}^2\right] \approx \left[\frac{1}{\{E(S)\}^2}\right]Var(R) + \left[\frac{\{E(R)\}^2}{\{E(S)\}^4}\right]Var(S)$$
$$- 2\left[\frac{E(R)}{\{E(S)\}^3}\right]Cov(R,S).$$

In our case, we have $\hat{w}_c = n_D^c/n_D$, giving $R = n_D^c, S = n_D$, so $E(R) = n_c\gamma_c$, $Var(R) = n_c\gamma_c(1-\gamma_c)$, $E(S) = \sum_{c=1}^{m} n_c\gamma_c$, $Var(S) = \sum_{c=1}^{m} n_c\gamma_c(1-\gamma_c)$, and $Cov(R,S) = n_c\gamma_c(1-\gamma_c)$. Then

$$Var(\hat{w}_c) = Var\left(\frac{n_D^c}{n_D}\right) \approx \left(\frac{n_c\gamma_c}{\sum_{c=1}^{m} n_c\gamma_c}\right)^2 \left\{\frac{1-\gamma_c}{n_c\gamma_c} - 2\frac{1-\gamma_c}{\sum_{c=1}^{m} n_c\gamma_c} + \frac{\sum_{c=1}^{m} n_c\gamma_c(1-\gamma_c)}{(\sum_{c=1}^{m} n_c\gamma_c)^2}\right\}$$
$$\leq \frac{1-\gamma_c}{n_c\gamma_c} + \frac{\sum_{c=1}^{m} n_c\gamma_c(1-\gamma_c)}{(\sum_{c=1}^{m} n_c\gamma_c)^2}.$$

By Hölder's inequality

$$\sum_{c=1}^{m} \sqrt{Var(\hat{w}_c)} \leq \sqrt{mA},$$

where

$$A = \sum_{c=1}^{m} \left\{ \frac{1-\gamma_c}{n_c\gamma_c} + \frac{\sum_{c=1}^{m} n_c\gamma_c(1-\gamma_c)}{(\sum_{c=1}^{m} n_c\gamma_c)^2} \right\}.$$

We can write

$$A = \sum_{c=1}^{m} \frac{1-\gamma_c}{n_c\gamma_c} + m \left\{ \frac{\sum_{c=1}^{m} n_c\gamma_c(1-\gamma_c)}{(\sum_{c=1}^{m} n_c\gamma_c)^2} \right\} \leq \frac{1}{\min_c n_c} \sum_{c=1}^{m} (1/\gamma_c) + \frac{m}{\sum_{c=1}^{m} n_c\gamma_c}.$$

Furthermore, we have $1/\gamma_c \leq V < \infty$. Then

$$A \leq \frac{mV}{\min_c n_c} + \frac{m}{\min_c n_c \sum_{c=1}^{m} \gamma_c} \leq \frac{(m+1)V}{\min_c n_c} \approx \frac{2mV}{\min_c n_c}.$$

Then $\sum_{c=1}^{m} \sqrt{Var(\hat{w}_c)} \to 0$ if $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\frac{\sqrt{\min_c n_c}}{m} \to \infty$. This holds if $n_c \to \infty$, $c = 1, ..., m$, and $m \to M$ such that $\sqrt{n_c}/m \to \infty$. This then gives $P(\sum_{c=1}^{m} |\hat{w}_c - w_c| > \epsilon) \to 0$, completing the proof. □

A.2.5 Proof of Theorem 4.1

Proof. We can write this claim as

$$\left|a\hat{AUC}(\hat{\boldsymbol{\theta}}) - aAUC(\boldsymbol{\theta}_0)\right| = \left|a\hat{AUC}(\hat{\boldsymbol{\theta}}) - aAUC(\hat{\boldsymbol{\theta}}) + aAUC(\hat{\boldsymbol{\theta}}) - aAUC(\boldsymbol{\theta}_0)\right| = o_p(1).$$

Then

$$\left|a\hat{AUC}(\hat{\boldsymbol{\theta}}) - aAUC(\boldsymbol{\theta}_0)\right| \leq \left|a\hat{AUC}(\hat{\boldsymbol{\theta}}) - aAUC(\hat{\boldsymbol{\theta}})\right| + \left|aAUC(\hat{\boldsymbol{\theta}}) - aAUC(\boldsymbol{\theta}_0)\right|.$$

By the uniform convergence of $a\hat{AUC}(\boldsymbol{\theta})$ (Lemma 4.2),

$$\left|a\hat{AUC}(\hat{\boldsymbol{\theta}}) - aAUC(\hat{\boldsymbol{\theta}})\right| \leq \sup_{\boldsymbol{\theta}\in B}\left|a\hat{AUC}(\boldsymbol{\theta}) - aAUC(\boldsymbol{\theta})\right| = o_p(1).$$

Next, we can write

$$\begin{aligned}\left|aAUC(\hat{\boldsymbol{\theta}}) - aAUC(\boldsymbol{\theta}_0)\right| &= \left|\sum_{c=1}^{M} w_c(AUC_c(\hat{\boldsymbol{\theta}}) - AUC_c(\boldsymbol{\theta}_0))\right| \\ &\leq \sum_{c=1}^{M} w_c\left|AUC_c(\hat{\boldsymbol{\theta}}) - AUC_c(\boldsymbol{\theta}_0)\right|\end{aligned}$$

Then we can apply Taylor's theorem to $AUC_c(\hat{\boldsymbol{\theta}}) - AUC_c(\boldsymbol{\theta}_0)$ using condition (C5) in Section 4.3.3. This gives

$$AUC_c(\hat{\boldsymbol{\theta}}) - AUC_c(\boldsymbol{\theta}_0) \approx \left(\frac{\partial}{\partial \mathbf{t}} AUC_c(\mathbf{t})\bigg|_{\mathbf{t}=\boldsymbol{\theta}_0}\right)^{\top}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Then

$$\sum_{c=1}^{M} w_c \left| AUC_c(\hat{\boldsymbol{\theta}}) - AUC_c(\boldsymbol{\theta}_0) \right| \approx \sum_{c=1}^{M} w_c \left| AUC_c'(\boldsymbol{\theta}_0)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \right|$$
$$\leq \sqrt{p} \times T \times ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| \times \sum_{c=1}^{M} w_c = o_p(1),$$

by condition (C5) in Section 4.3.3, the Cauchy-Schwarz inequality, the convergence of $\hat{\boldsymbol{\theta}}$, the continuous mapping theorem, and the fact that $\sum_{c=1}^{M} w_c = 1$. □

A.3 Chapter 5

A.3.1 Proof of Theorem 5.1

Proof. 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

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

We can write

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

Under conditions (A1)-(A4) in Section 5.3.2, we have shown (Lemmas 4.1 and 4.2)

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in B}\left|a\hat{AU}C(\boldsymbol{\theta})-aAUC(\boldsymbol{\theta})\right|&=o_p(1)\\ \sup_{\boldsymbol{\theta}\in B}\left|A\hat{U}C_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta})\right|&=o_p(1),\quad c=1,...,M.\end{aligned}$$

We can write

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

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 Lemma 4.2,

$$
\sup_{\boldsymbol{\theta}\in B}|Q_n(\boldsymbol{\theta};\lambda)-Q(\boldsymbol{\theta};\lambda)|\le o_p(1)+o(1)+\lambda\sum_{c=1}^{m}\sup_{\boldsymbol{\theta}\in B}\left|w_cY_1^c(\boldsymbol{\theta})^2-\hat{w}_c(Y_2^c(\boldsymbol{\theta})+Y_1^c(\boldsymbol{\theta})+Y_3(\boldsymbol{\theta}))^2\right|,
$$

where

$$
\begin{aligned}
Y_1^c(\boldsymbol{\theta}) &= AUC_c(\boldsymbol{\theta})-aAUC(\boldsymbol{\theta}) \\
Y_2^c(\boldsymbol{\theta}) &= A\hat{U}C_c(\boldsymbol{\theta})-AUC_c(\boldsymbol{\theta}) \\
Y_3(\boldsymbol{\theta}) &= aAUC(\boldsymbol{\theta})-aA\hat{U}C(\boldsymbol{\theta});
\end{aligned}
$$

note that $|Y_1^c(\boldsymbol{\theta})| \leq 1$, $|Y_2^c(\boldsymbol{\theta})| \leq 1$ and $|Y_3(\boldsymbol{\theta})| \leq 1$. Then

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

We have (by Lemmas 4.1 and 4.2)

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in B} |Y_2^c(\boldsymbol{\theta})| = o_p(1), c = 1, ..., M \\
&\sup_{\boldsymbol{\theta}\in B} |Y_3(\boldsymbol{\theta})| = o_p(1).
\end{aligned}$$

This gives

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

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 (Lemma 4.2) 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

$$\sup_{\boldsymbol{\theta} \in B} \left| aR_n(\boldsymbol{\theta}) - a\hat{A}\hat{U}C(\boldsymbol{\theta}) \right| = o_p(1).$$

Ma and Huang (2007) demonstrated that $\sup_{\boldsymbol{\theta} \in B} \left| R^c_{n_c}(\boldsymbol{\theta}) - \hat{AU}C_c(\boldsymbol{\theta}) \right| = o_p(1)$ as $n_c \to \infty$.

We can write

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

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

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

where

$$Z_1^c(\boldsymbol{\theta}) = A\hat{U}C_c(\boldsymbol{\theta}) - aA\hat{U}C(\boldsymbol{\theta})$$
$$Z_2^c(\boldsymbol{\theta}) = R_{n_c}^c(\boldsymbol{\theta}) - A\hat{U}C_c(\boldsymbol{\theta})$$
$$Z_3(\boldsymbol{\theta}) = aA\hat{U}C(\boldsymbol{\theta}) - aR_n(\boldsymbol{\theta});$$

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

$$\sup_{\boldsymbol{\theta}\in B}\left|Z_2^c(\boldsymbol{\theta})\right| = o_p(1), c = 1, ..., M$$
$$\sup_{\boldsymbol{\theta}\in B}\left|Z_3(\boldsymbol{\theta})\right| = o_p(1).$$

This gives

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

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 $\sqrt{n}_c / m \to \infty$. □

Appendix B

ADDITIONAL SIMULATION RESULTS

B.1 Chapter 2

The tables below present additional results related to the simulations described in Section 2.4; 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 B.1: 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 recalculated 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 B.2: 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 recalculated 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 B.3: 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 recalculated 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 B.4: 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 $\beta_0 = 0.6$ 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 recalculated 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)

B.2 Chapter 3

B.2.1 Constructing Combinations

Cumulative Logit Model Does Not Hold: Results for $\Sigma_X = \Sigma_2$ and $\Sigma_X = \Sigma_4$

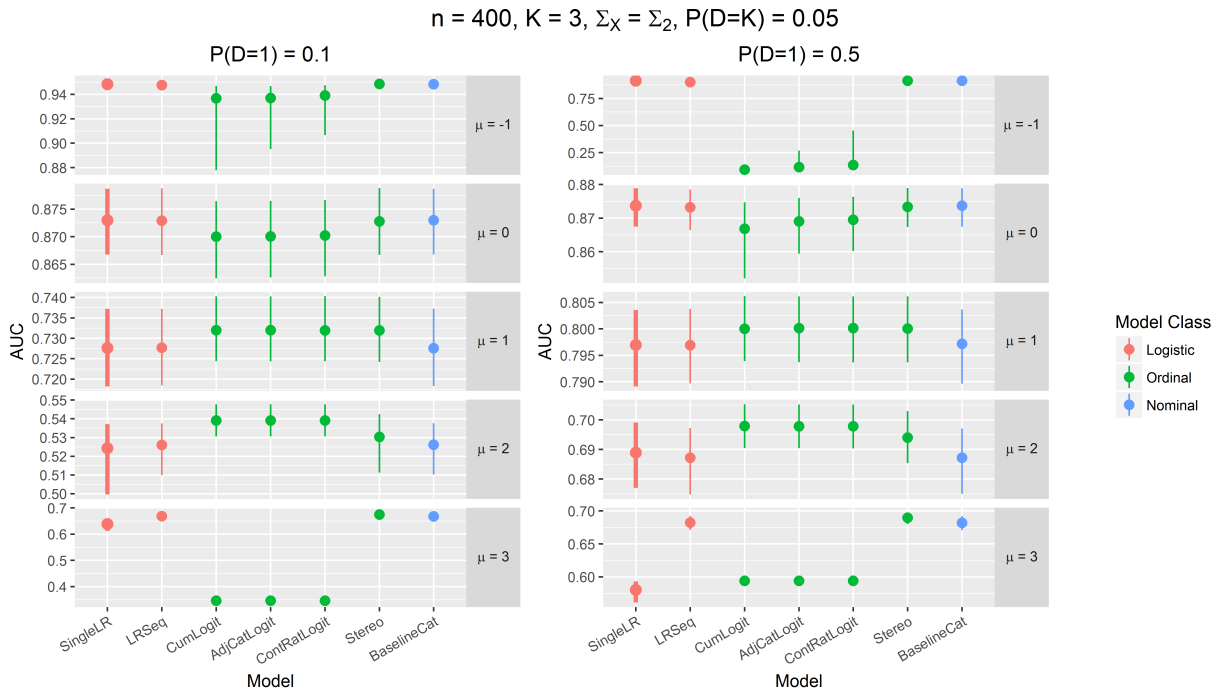

Figure B.1: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.05$ and $\Sigma_X = \Sigma_2$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

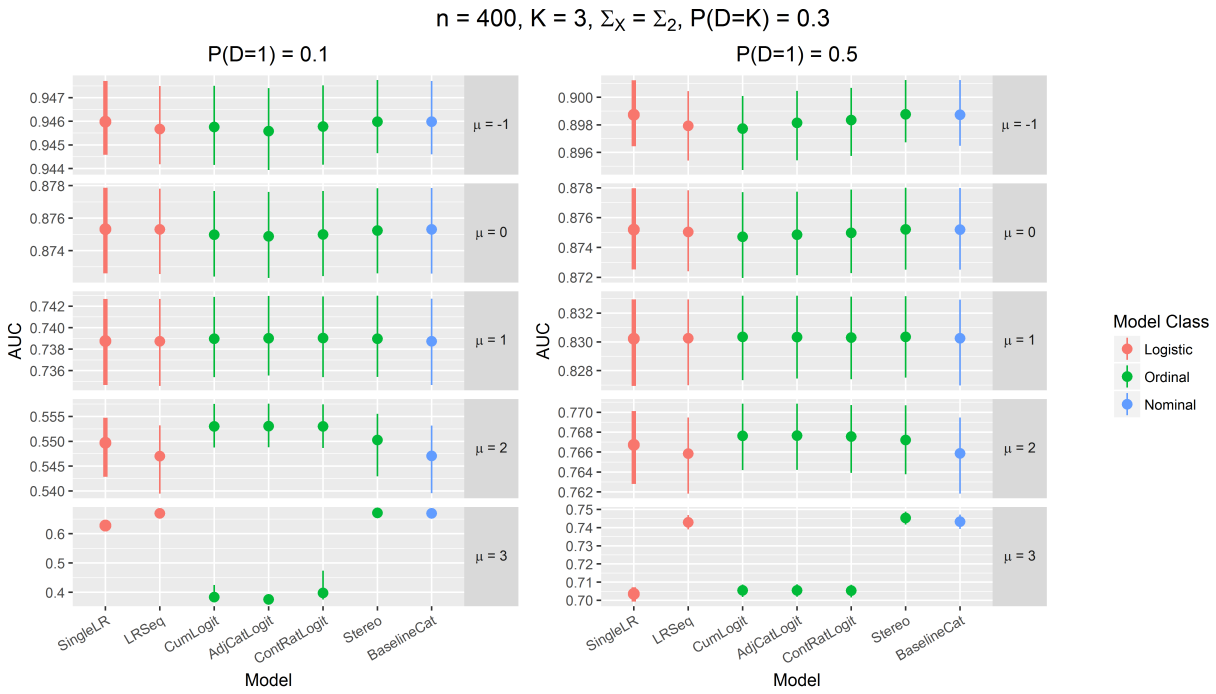

Figure B.2: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.3$ and $\Sigma_X = \Sigma_2$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

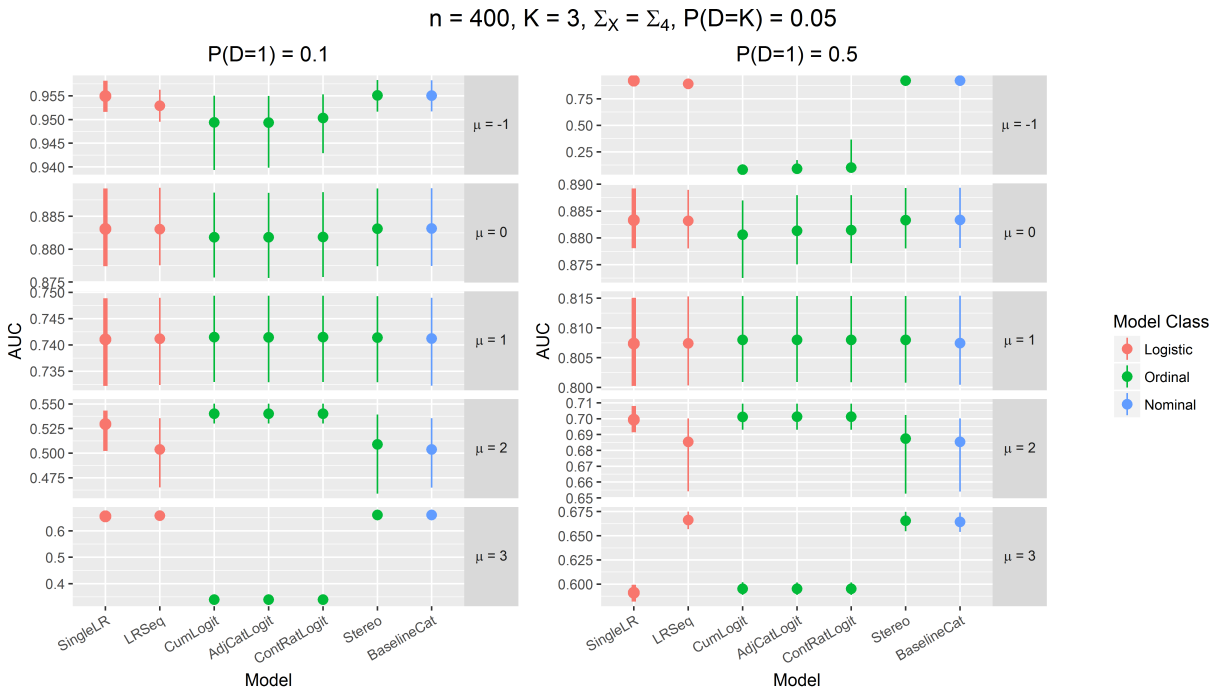

Figure B.3: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.05$ and $\Sigma_X = \Sigma_4$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

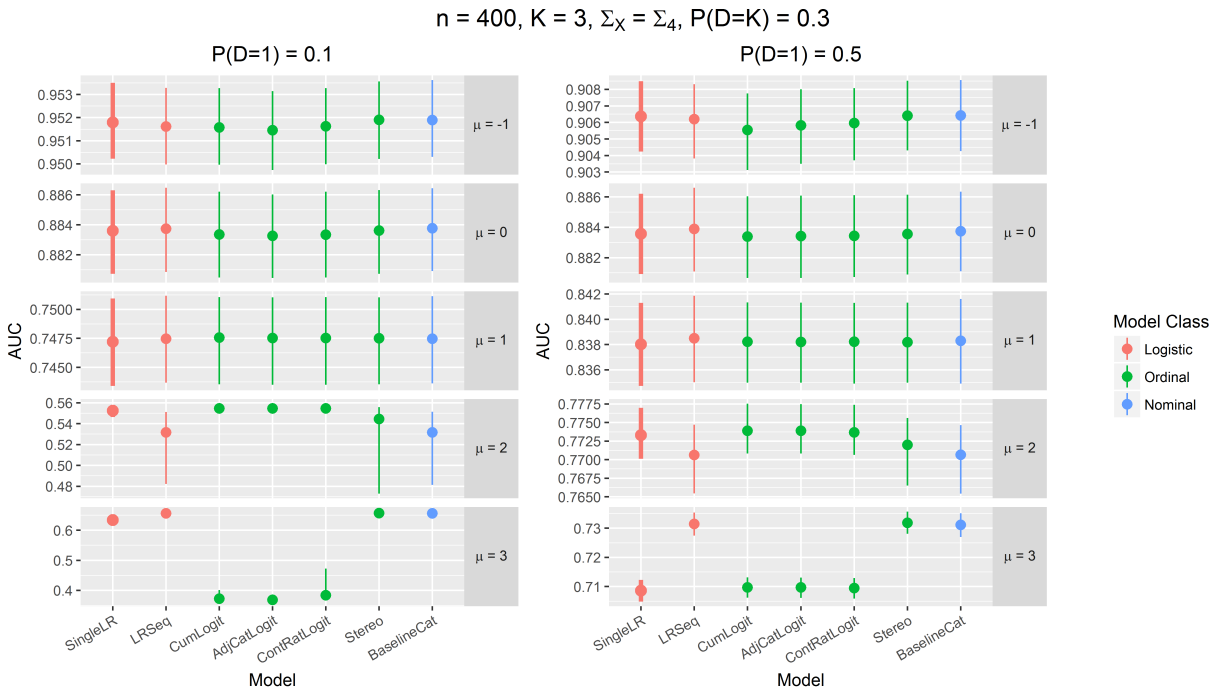

Figure B.4: Simulation results for $K = 3$ when the cumulative logit model does not hold, $P(D = 3) = 0.3$ and $\Sigma_X = \Sigma_4$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

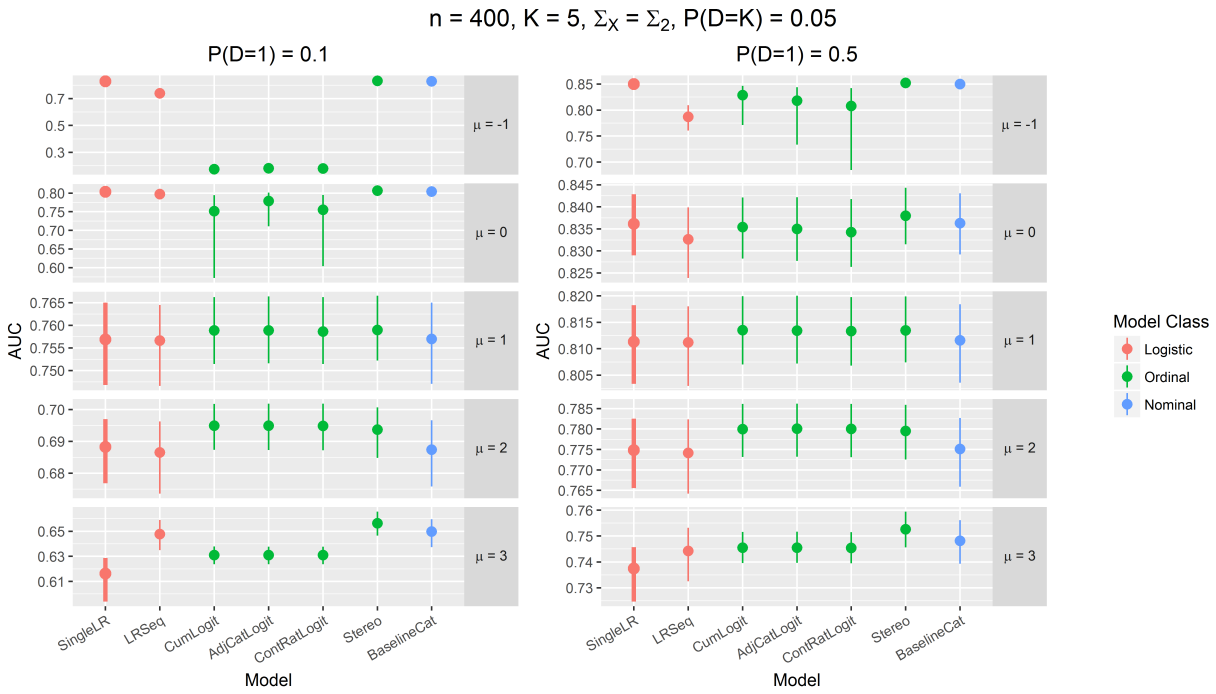

Figure B.5: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.05$ and $\Sigma_X = \Sigma_2$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

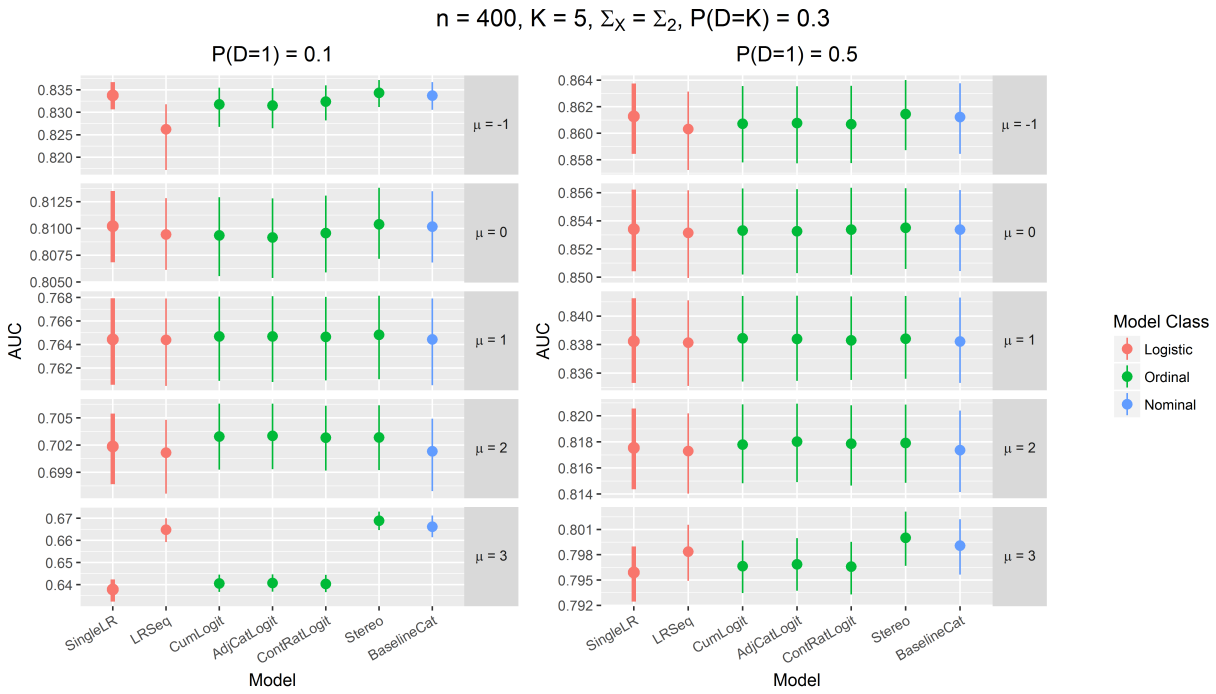

Figure B.6: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.3$ and $\Sigma_X = \Sigma_2$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

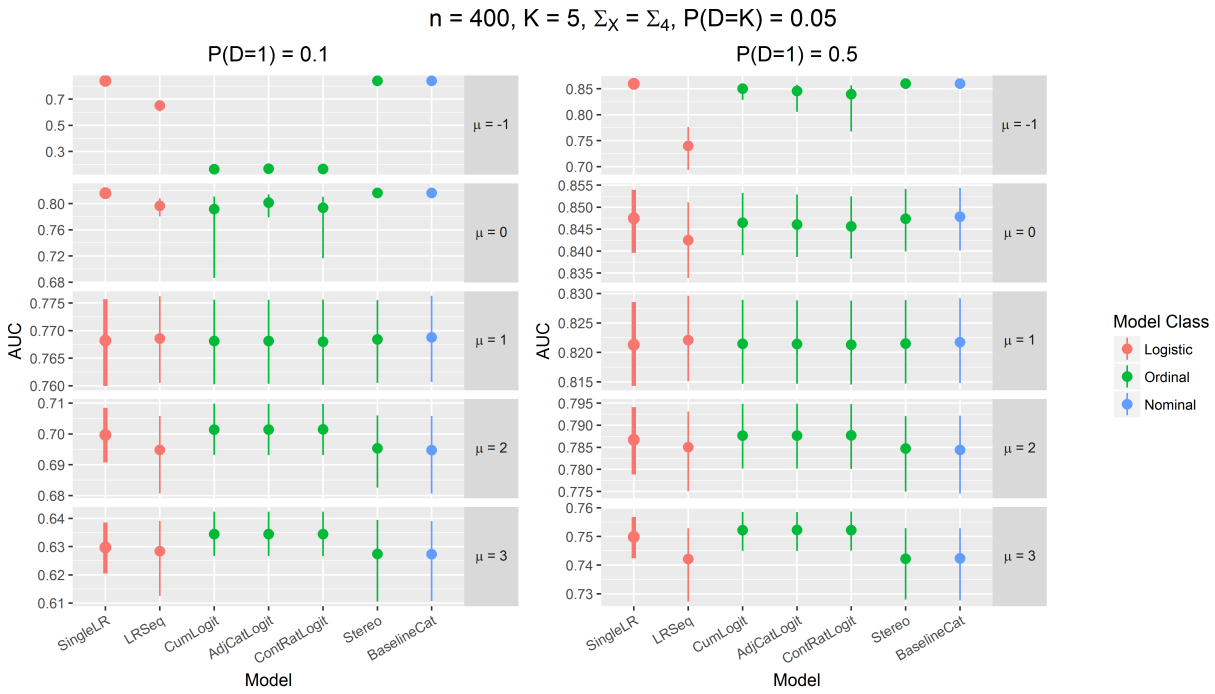

Figure B.7: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.05$ and $\Sigma_X = \Sigma_4$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

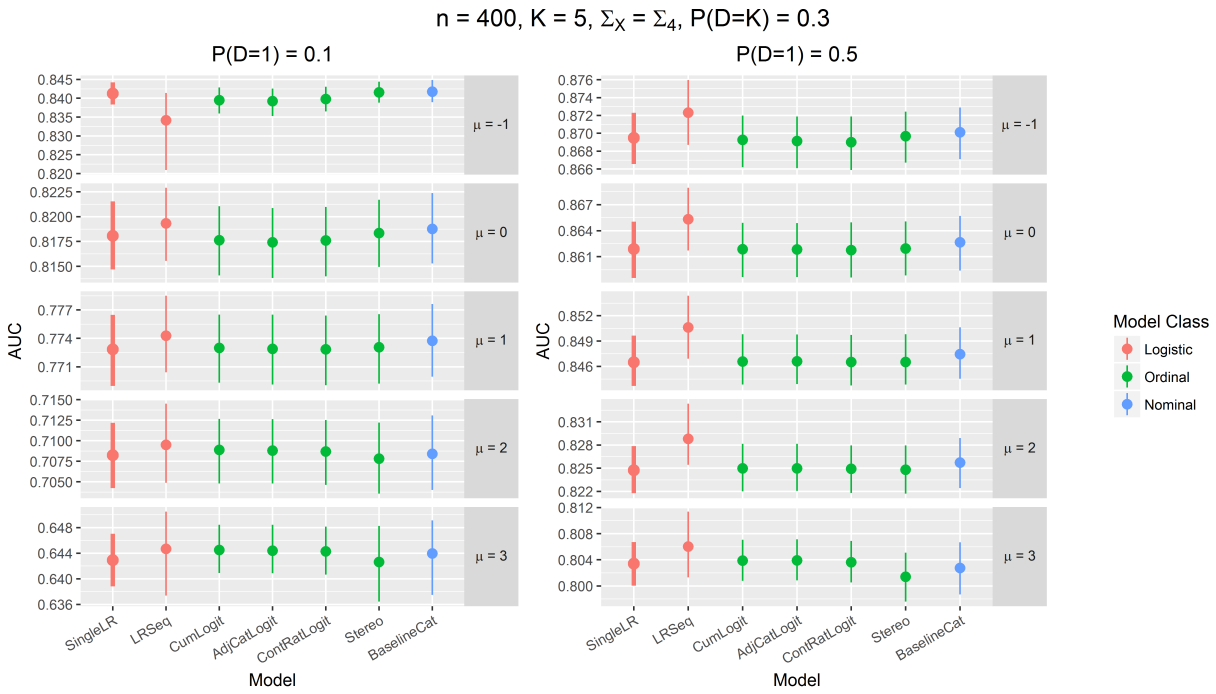

Figure B.8: Simulation results for $K = 5$ when the cumulative logit model does not hold, $P(D = 5) = 0.3$ and $\Sigma_X = \Sigma_4$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\mu}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

Cumulative Logit Model Holds: Results for $P(D = K) = 0.3$

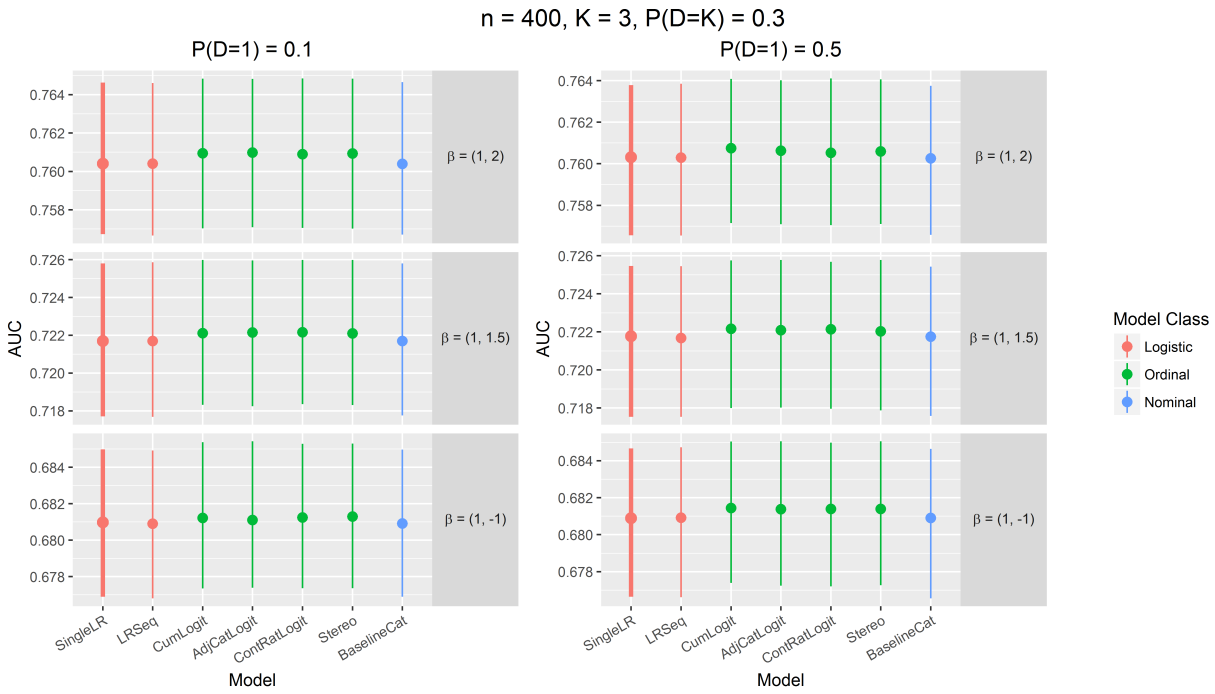

Figure B.9: Simulation results for $K = 3$ when the cumulative logit model with proportional odds holds and $P(D = 3) = 0.3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\beta}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

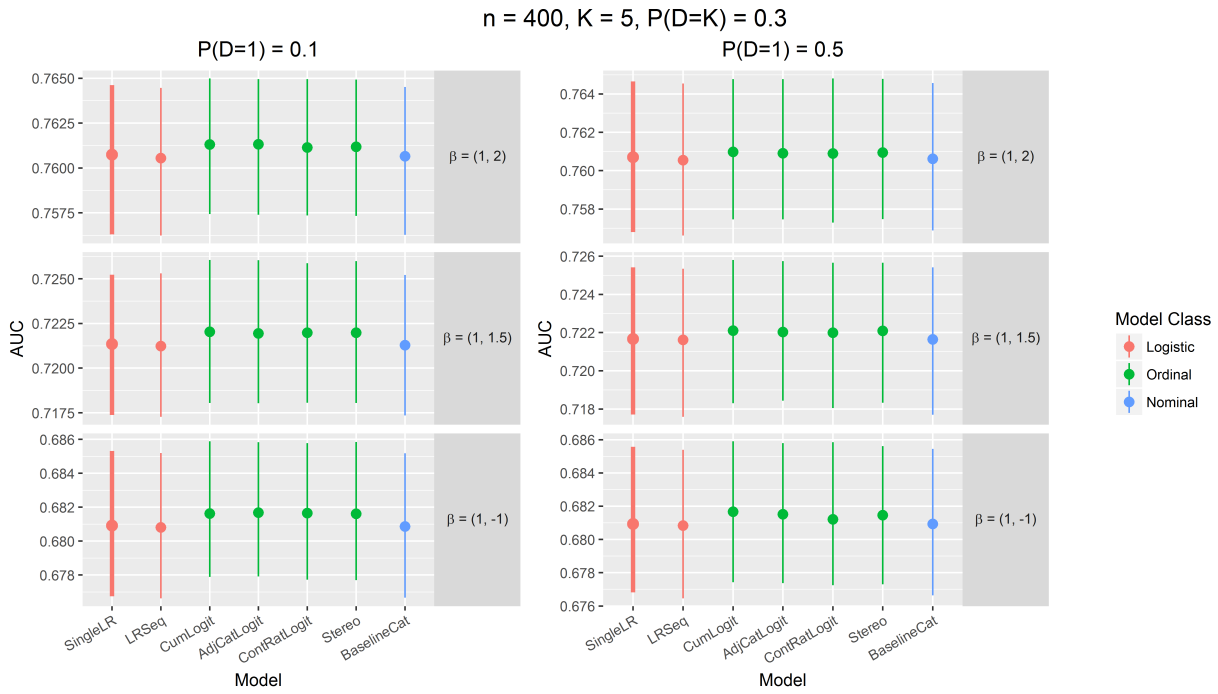

Figure B.10: Simulation results for $K = 5$ when the cumulative logit model with proportional odds holds and $P(D = 5) = 0.3$. Each plot presents the median and interquartile range of the AUCs for $D = K$ vs. $D < K$ in the test data for the combinations fitted by each strategy, which are given on the x-axis. The results are presented by $P(D = 1)$ (columns) and $\boldsymbol{\beta}$ (rows). The "standard" approach is the single binary logistic regression model ("SingleLR") and is indicated by a slightly thicker line and larger point.

B.2.2 Combination Selection

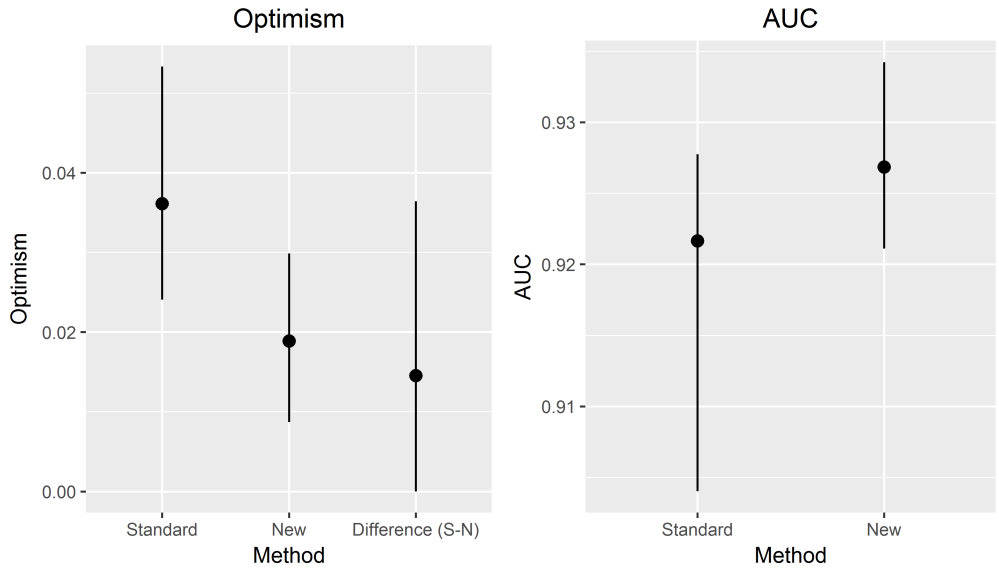

Figure B.11: Results for the proposed combination selection method for simulation Example 2, where the cumulative logit model with proportional odds holds. The plots give the results for the standard approach, that is, choosing the combination based on the estimated AUC (corrected for optimism due to resubstitution bias) for $D = 3$ vs. $D < 3$, and the results for the new approach, that is, choosing the combination based on the AUC for $D = 3$ vs. $D < 3$ and the AUC for $D = 2$ vs. $D = 1$. The plot on the left gives the median and interquartile range for the estimated optimism due to model selection bias (the difference between the estimated AUC, corrected for optimism due to resubstitution bias, and the AUC in test data) for the selected combinations and the difference in the estimated optimism between the two approaches. The plot on the right gives the $D = 3$ vs. $D < 3$ AUC in test data for the combinations selected by the two approaches.

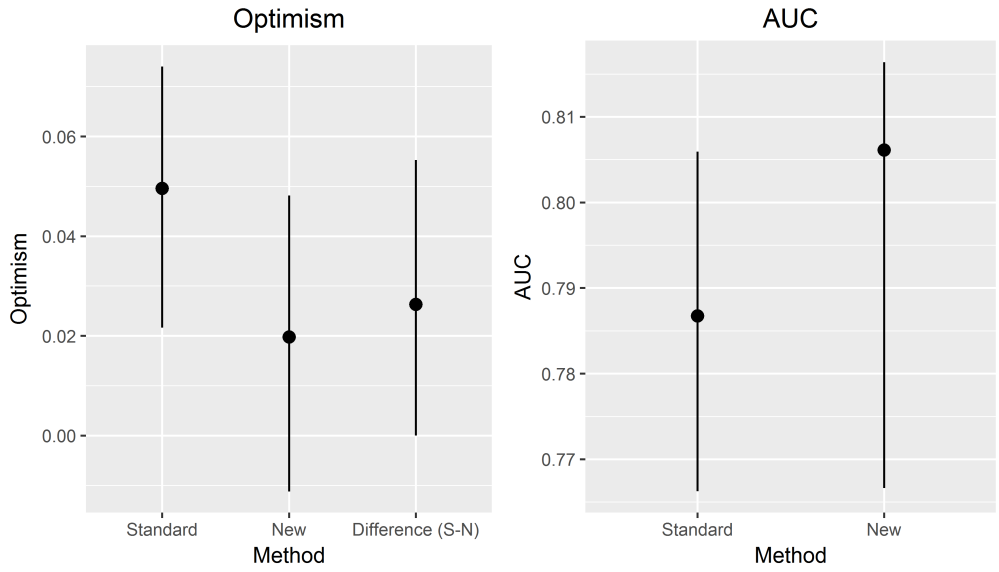

Figure B.12: Results for the proposed combination selection method for simulation Example 3, where the cumulative logit model with proportional odds does not hold. The plots give the results for the standard approach, that is, choosing the combination based on the estimated AUC (corrected for optimism due to resubstitution bias) for $D = 3$ vs. $D < 3$, and the results for the new approach, that is, choosing the combination based on the AUC for $D = 3$ vs. $D < 3$ and the AUC for $D = 2$ vs. $D = 1$. The plot on the left gives the median and interquartile range for the estimated optimism due to model selection bias (the difference between the estimated AUC, corrected for optimism due to resubstitution bias, and the AUC in test data) for the selected combinations and the difference in the estimated optimism between the two approaches. The plot on the right gives the $D = 3$ vs. $D < 3$ AUC in test data for the combinations selected by the two approaches.

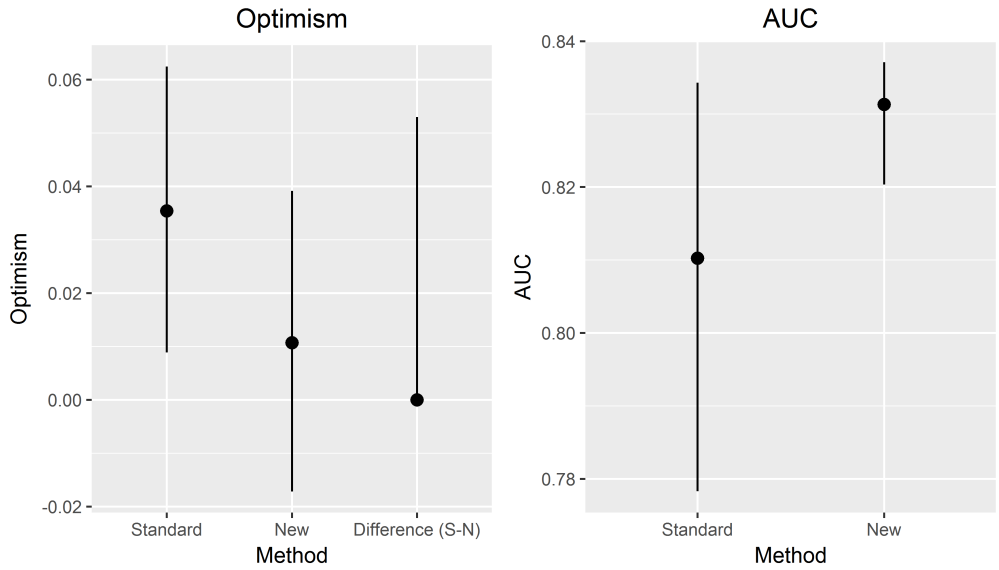

Figure B.13: Results for the proposed combination selection method for simulation Example 5, where the cumulative logit model with proportional odds does not hold. The plots give the results for the standard approach, that is, choosing the combination based on the estimated AUC (corrected for optimism due to resubstitution bias) for $D = 3$ vs. $D < 3$, and the results for the new approach, that is, choosing the combination based on the AUC for $D = 3$ vs. $D < 3$ and the AUC for $D = 2$ vs. $D = 1$. The plot on the left gives the median and interquartile range for the estimated optimism due to model selection bias (the difference between the estimated AUC, corrected for optimism due to resubstitution bias, and the AUC in test data) for the selected combinations and the difference in the estimated optimism between the two approaches. The plot on the right gives the $D = 3$ vs. $D < 3$ AUC in test data for the combinations selected by the two approaches.

B.3 Chapter 4

B.3.1 Ignoring Center

The tables below present the full results for the simulations reported in Section 4.4.1. In each table, we present the average across simulations of the coefficient estimates and AUCs, the percent bias, and the mean squared error (MSE). The percent bias and MSE for $\hat{\alpha}_1$ and $\hat{\beta}_1$ are relative to β_1, the percent bias and MSE for $\hat{\alpha}_2$ and $\hat{\beta}_2$ are relative to β_2, and the percent bias and MSE for $AUC(\hat{\boldsymbol{\alpha}})$, $AUC_c(\hat{\boldsymbol{\alpha}})$, $AUC(\hat{\boldsymbol{\beta}})$, and $AUC_c(\hat{\boldsymbol{\beta}})$ are relative to $AUC_c(\boldsymbol{\beta})$. We have multiplied the MSE by 10^4. The label "Confounder (+)" refers to the setting of positive correlation between logit(γ_c) and $f(c)$, while "Confounder (-)" refers to the setting of negative correlation between logit(γ_c) and $f(c)$. Although we refer to the differences between the parameter estimates ($\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$) and $\boldsymbol{\beta}$ and between the AUC values ($AUC(\hat{\boldsymbol{\alpha}})$, $AUC_c(\hat{\boldsymbol{\alpha}})$, $AUC(\hat{\boldsymbol{\beta}})$, and $AUC_c(\hat{\boldsymbol{\beta}})$) and $AUC_c(\boldsymbol{\beta})$ as "bias", $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ reflect different population parameters (in general), as do $AUC(\boldsymbol{\alpha})$, $AUC_c(\boldsymbol{\alpha})$, $AUC(\boldsymbol{\beta})$, and $AUC_c(\boldsymbol{\beta})$.

Table B.5: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates and the AUC across 500 simulations when there were 6 centers in the training data and center was either a case mix variable, a calibration variable, a confounder with positive correlation between logit(γ_c) and $f(c)$ ("Confounder (+)") or a confounder with negative correlation between logit(γ_c) and $f(c)$ ("Confounder (-)"). The coefficient estimates $(\hat{\alpha}_1, \hat{\alpha}_2)$ and marginal AUC ($AUC(\cdot)$) correspond to ignoring center during construction and evaluation, respectively. The coefficient estimates $(\hat{\beta}_1, \hat{\beta}_2)$ and conditional AUC ($AUC_c(\cdot)$) correspond to accounting for center during construction and evaluation, respectively. The MSE is multiplied by 10^4.

	Case Mix			Calibration			Confounder (+)			Confounder (-)		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$\hat{\alpha}_1$	0.116	1.468	50.07	-0.124	-207.952	613.40	0.002	-98.554	211.96	-0.251	-319.351	1429.40
$\hat{\alpha}_2$	0.491	0.625	59.57	0.249	-48.974	625.19	0.374	-23.223	205.83	0.125	-74.306	1408.40
$\hat{\beta}_1$	0.116	1.660	61.30	0.117	2.105	45.55	0.118	3.308	59.96	0.112	-2.252	62.56
$\hat{\beta}_2$	0.492	0.783	71.89	0.490	0.483	49.31	0.490	0.523	62.94	0.490	0.459	61.89
$AUC(\hat{\boldsymbol{\alpha}})$	0.651	-0.133	0.04	0.577	-11.618	59.05	0.707	8.360	66.59	0.577	-11.495	118.38
$AUC_c(\hat{\boldsymbol{\alpha}})$	0.651	-0.130	0.04	0.612	-6.246	22.42	0.646	-0.972	1.44	0.479	-26.582	349.41
$AUC(\hat{\boldsymbol{\beta}})$	0.651	-0.163	0.06	0.574	-12.036	64.59	0.708	8.544	68.43	0.432	-33.804	556.73
$AUC_c(\hat{\boldsymbol{\beta}})$	0.651	-0.162	0.06	0.652	-0.119	0.03	0.651	-0.162	0.05	0.651	-0.164	0.05

Table B.6: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates and the AUC across 500 simulations when there were 500 centers in the training data and center was either a case mix variable, a calibration variable, a confounder with positive correlation between logit(γ_c) and $f(c)$ ("Confounder (+)") or a confounder with negative correlation between logit(γ_c) and $f(c)$ ("Confounder (-)"). The coefficient estimates $(\hat{\alpha}_1, \hat{\alpha}_2)$ and marginal AUC ($AUC(\cdot)$) correspond to ignoring center during construction and evaluation, respectively. The coefficient estimates $(\hat{\beta}_1, \hat{\beta}_2)$ and conditional AUC ($AUC_c(\cdot)$) correspond to accounting for center during construction and evaluation, respectively. The MSE is multiplied by 10^4.

	Case Mix			Calibration			Confounder (+)			Confounder (-)		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$\hat{\alpha}_1$	0.114	-0.713	5.71	-0.147	-228.472	688.20	-0.014	-112.174	169.94	-0.275	-340.089	1519.49
$\hat{\alpha}_2$	0.488	-0.028	5.80	0.226	-53.703	690.71	0.363	-25.627	161.58	0.097	-80.021	1527.88
$\hat{\beta}_1$	0.114	-0.808	7.41	0.115	0.900	7.06	0.113	-1.420	6.99	0.114	-0.380	7.02
$\hat{\beta}_2$	0.487	-0.071	7.44	0.488	0.046	6.75	0.490	0.390	7.60	0.488	0.094	7.80
$AUC(\hat{\boldsymbol{\alpha}})$	0.652	-0.026	0.02	0.574	-12.016	61.47	0.719	10.293	45.54	0.624	-4.402	8.91
$AUC_c(\hat{\boldsymbol{\alpha}})$	0.652	-0.026	0.02	0.599	-8.170	28.81	0.649	-0.531	0.17	0.447	-31.473	422.54
$AUC(\hat{\boldsymbol{\beta}})$	0.652	-0.031	0.02	0.559	-14.290	86.94	0.719	10.169	44.47	0.395	-39.418	662.03
$AUC_c(\hat{\boldsymbol{\beta}})$	0.652	-0.031	0.02	0.652	-0.042	0.02	0.652	-0.015	0.02	0.652	-0.026	0.02

B.3.2 RILR vs. FILR

The tables below present the full results for the simulations reported in Section 4.4.2. The results are separated by the role of center (i.e., case mix variable, calibration variable, or confounder with negative, no, or positive correlation between $f(c)$ and $\text{logit}(\gamma_c)$) and the distribution of $\text{logit}(\gamma_c)$ and/or $f(c)$, F (i.e., normal, Gumbel, Laplace, or uniform). In the first table we report the mean, percent bias, and mean squared error (MSE) for the coefficient estimates based on RILR $(\hat{\tau}_1, \hat{\tau}_2)$ and FILR $(\hat{\beta}_1, \hat{\beta}_2)$ relative to (β_1, β_2), as well as the average, percent bias and MSE for the conditional AUCs based on these coefficients ($AUC_c(\hat{\boldsymbol{\tau}})$ and $AUC_c(\hat{\boldsymbol{\beta}})$, respectively) relative to $AUC_c(\boldsymbol{\beta})$. In the second table we report the average, percent bias and MSE for the overall (fixed) intercept estimate provided by RILR ($\hat{\tau}_0$). In the setting where center is a calibration variable, asterisks indicate the results for $\gamma_c = 0.1$; the other calibration variable results are for $\gamma_c = 0.5$. Again, we have multiplied the MSE by 10^4. As noted above, the "bias" is calculated as the difference between the parameter estimates and $\boldsymbol{\beta}$ and between the AUC values and $AUC_c(\boldsymbol{\beta})$; it is not exactly accurate to call these differences biases, as they arise because (in the case of $\hat{\boldsymbol{\tau}}$ and $AUC_c(\hat{\boldsymbol{\tau}})$) they correspond to different population parameters.

Case Mix

Normal

Table B.7: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a case mix variable and logit(γ_c) was normally distributed. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
1	0.114	-0.07	60.07	0.114	-0.08	60.71	0.494	1.19	58.32	0.494	1.23	59.01	0.651	-0.18	0.39	0.651	-0.19	0.39
0.5	0.115	0.54	53.28	0.115	0.53	53.95	0.493	1.15	55.62	0.493	1.18	56.25	0.651	-0.14	0.36	0.651	-0.15	0.37
0.5/1.5	0.116	1.09	60.76	0.116	1.18	61.39	0.490	0.47	61.84	0.490	0.51	62.67	0.651	-0.18	0.38	0.651	-0.18	0.38
3	0.117	2.23	75.12	0.117	2.32	76.06	0.492	0.79	90.99	0.492	0.82	92.23	0.651	-0.16	0.79	0.651	-0.16	0.79
1/5	0.112	-2.12	75.17	0.112	-2.08	75.93	0.485	-0.60	75.80	0.485	-0.60	76.69	0.651	-0.23	0.44	0.651	-0.23	0.44
								$m = 500$										
1	0.119	3.79	7.87	0.114	-0.53	7.75	0.510	4.59	13.01	0.489	0.23	8.08	0.652	-0.06	0.42	0.652	-0.06	0.42
0.5	0.119	3.57	7.04	0.115	0.07	7.07	0.505	3.61	9.94	0.488	-0.03	6.91	0.653	0.04	0.34	0.653	0.04	0.34
0.5/1.5	0.119	4.30	6.44	0.114	-0.02	6.18	0.510	4.56	11.85	0.489	0.23	6.88	0.652	0.02	0.33	0.652	0.01	0.33
3	0.120	5.24	9.72	0.115	0.26	9.56	0.511	4.82	14.82	0.487	-0.16	9.41	0.652	0.01	0.84	0.652	0.00	0.84
1/5	0.120	5.10	9.10	0.115	0.11	8.68	0.514	5.44	15.82	0.490	0.38	8.64	0.653	0.04	0.37	0.653	0.03	0.37

Table B.8: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a case mix variable and logit(γ_c) was normally distributed. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	Mean	Bias	MSE
	$m = 6$		
1	-0.149	-3.00	1777.65
0.5	-0.161	4.72	1021.73
0.5/1.5	-0.189	23.10	1701.55
3	-0.173	12.56	4994.47
1/5	-0.079	-48.24	5019.95
	$m = 500$		
1	-0.158	2.92	25.53
0.5	-0.162	5.71	15.04
0.5/1.5	-0.162	5.39	27.89
3	-0.163	6.27	65.01
1/5	-0.167	8.93	58.85

Gumbel

Table B.9: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a case mix variable and logit(γ_c) had a Gumbel distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
1	0.111	-2.69	59.00	0.111	-2.70	59.68	0.492	0.87	68.51	0.492	0.88	69.21	0.651	-0.22	0.47	0.651	-0.22	0.47
0.5	0.120	4.71	57.30	0.120	4.80	57.91	0.492	0.84	56.23	0.492	0.85	56.84	0.651	-0.15	0.37	0.651	-0.15	0.37
0.5/1.5	0.118	3.51	53.66	0.118	3.55	54.29	0.493	1.16	55.68	0.494	1.19	56.28	0.651	-0.16	0.42	0.651	-0.16	0.42
3	0.120	4.53	72.70	0.120	4.63	73.66	0.490	0.45	74.68	0.490	0.48	75.59	0.652	-0.04	0.86	0.652	-0.05	0.86
1/5	0.111	-3.25	66.01	0.111	-3.28	66.65	0.497	1.98	69.62	0.497	2.00	70.39	0.651	-0.20	0.59	0.651	-0.20	0.59
								$m = 500$										
1	0.121	5.94	7.34	0.116	1.62	6.89	0.509	4.38	12.25	0.488	0.07	7.64	0.652	-0.00	0.47	0.652	-0.00	0.47
0.5	0.120	4.99	7.29	0.116	1.45	7.21	0.504	3.35	9.73	0.487	-0.22	7.26	0.652	-0.03	0.34	0.652	-0.03	0.34
0.5/1.5	0.121	5.39	6.75	0.116	1.05	6.35	0.507	4.00	10.24	0.486	-0.26	6.48	0.653	0.04	0.34	0.653	0.04	0.33
3	0.120	4.70	8.30	0.114	-0.21	8.15	0.512	4.95	13.81	0.488	-0.02	8.11	0.652	-0.07	0.74	0.652	-0.08	0.74
1/5	0.118	3.49	8.06	0.113	-1.48	7.94	0.515	5.51	15.56	0.490	0.49	8.40	0.652	-0.06	0.39	0.652	-0.06	0.39

Table B.10: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a case mix variable and logit(γ_c) had a Gumbel distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	Mean	Bias	MSE
	$m = 6$		
1	-0.124	-19.47	1864.27
0.5	-0.172	12.02	822.96
0.5/1.5	-0.164	6.96	1763.39
3	-0.113	-26.10	5151.42
1/5	-0.104	-32.38	5129.53
	$m = 500$		
1	-0.195	26.84	38.91
0.5	-0.174	13.28	18.73
0.5/1.5	-0.192	25.32	35.47
3	-0.260	69.38	175.98
1/5	-0.251	63.91	146.37

Laplace

Table B.11: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a case mix variable and logit(γ_c) had a Laplace distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
1	0.113	-1.56	62.51	0.113	-1.62	63.17	0.491	0.65	58.82	0.491	0.68	59.48	0.651	-0.14	0.47	0.651	-0.14	0.48
0.5	0.117	2.25	55.74	0.117	2.29	56.41	0.491	0.65	56.96	0.491	0.67	57.47	0.651	-0.13	0.37	0.651	-0.13	0.37
0.5/1.5	0.115	0.69	62.09	0.115	0.71	62.71	0.494	1.33	57.52	0.494	1.33	58.00	0.651	-0.22	0.42	0.651	-0.22	0.42
3	0.112	-1.94	75.19	0.112	-1.97	76.07	0.493	1.03	68.33	0.493	1.04	69.13	0.652	-0.12	3.28	0.652	-0.12	3.28
1/5	0.117	2.32	64.94	0.117	2.36	65.50	0.493	1.07	72.65	0.493	1.07	73.23	0.651	-0.19	0.43	0.651	-0.19	0.43
								$m = 500$										
1	0.121	5.77	6.84	0.116	1.47	6.51	0.508	4.14	11.39	0.487	-0.14	7.38	0.652	-0.08	0.44	0.652	-0.08	0.44
0.5	0.119	4.04	6.30	0.115	0.26	6.10	0.505	3.51	9.37	0.488	-0.01	6.45	0.652	0.02	0.33	0.652	0.02	0.33
0.5/1.5	0.117	2.64	7.14	0.112	-1.70	7.13	0.508	4.14	11.18	0.487	-0.07	7.36	0.652	0.00	0.35	0.652	-0.00	0.35
3	0.122	6.71	9.88	0.116	1.54	9.42	0.511	4.78	15.05	0.486	-0.29	9.70	0.653	0.13	2.22	0.653	0.12	2.22
1/5	0.119	3.92	8.15	0.113	-1.06	8.03	0.513	5.27	14.60	0.489	0.20	8.02	0.652	-0.10	0.39	0.652	-0.10	0.39

Table B.12: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a case mix variable and logit(γ_c) had a Laplace distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	Mean	Bias	MSE
	$m = 6$		
1	-0.191	24.52	1881.44
0.5	-0.160	4.31	938.70
0.5/1.5	-0.155	0.82	1713.69
3	-0.184	20.28	4805.53
1/5	-0.234	52.49	5052.67
	$m = 500$		
1	-0.159	3.71	20.97
0.5	-0.158	2.75	14.74
0.5/1.5	-0.163	5.97	21.87
3	-0.155	1.34	49.87
1/5	-0.157	2.59	48.05

Uniform

Table B.13: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a case mix variable and logit(γ_c) had a uniform distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
1	0.117	2.55	63.17	0.117	2.62	63.93	0.491	0.75	63.06	0.492	0.79	63.74	0.651	-0.14	0.44	0.651	-0.14	0.44
0.5	0.117	2.17	55.26	0.117	2.23	55.68	0.488	0.15	58.61	0.489	0.17	59.00	0.652	-0.12	0.36	0.652	-0.12	0.36
0.5/1.5	0.116	1.33	59.55	0.116	1.35	60.25	0.496	1.70	62.01	0.496	1.74	62.71	0.652	-0.10	0.35	0.652	-0.10	0.35
3	0.112	-2.50	79.76	0.112	-2.45	81.06	0.498	2.13	77.55	0.499	2.25	78.84	0.651	-0.27	0.76	0.651	-0.28	0.76
1/5	0.124	8.61	78.28	0.124	8.70	79.40	0.490	0.52	80.29	0.490	0.56	81.50	0.651	-0.26	0.46	0.651	-0.26	0.46
								$m = 500$										
1	0.120	4.91	7.14	0.115	0.75	6.91	0.509	4.42	12.14	0.488	-0.00	7.55	0.652	-0.12	0.39	0.652	-0.12	0.39
0.5	0.119	3.94	6.69	0.115	0.27	6.68	0.506	3.70	10.66	0.487	-0.04	7.54	0.652	-0.01	0.30	0.652	-0.01	0.30
0.5/1.5	0.118	3.40	7.14	0.113	-1.09	7.04	0.510	4.50	12.78	0.488	0.13	7.81	0.653	0.03	0.33	0.653	0.03	0.33
3	0.117	2.10	10.38	0.111	-2.75	10.65	0.514	5.39	16.39	0.491	0.61	9.66	0.652	-0.09	0.69	0.652	-0.10	0.69
1/5	0.119	4.29	9.84	0.113	-0.93	9.72	0.513	5.16	15.87	0.489	0.19	9.63	0.652	-0.01	0.39	0.652	-0.01	0.39

Table B.14: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a case mix variable and logit(γ_c) had a uniform distribution. The number of centers in the training data, m, and the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{\gamma_c}$), denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	Mean	Bias	MSE
	$m = 6$		
1	-0.157	2.38	1665.04
0.5	-0.160	4.31	833.86
0.5/1.5	-0.141	-7.92	1734.34
3	-0.184	19.84	4863.75
1/5	-0.163	6.16	5291.33
	$m = 500$		
1	-0.165	7.83	29.53
0.5	-0.161	5.19	17.88
0.5/1.5	-0.154	0.56	28.13
3	-0.157	2.46	91.62
1/5	-0.164	6.62	73.88

Calibration

Normal

Table B.15: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a calibration variable and $f(c)$ was normally distributed. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
5	0.092	-19.56	51.35	0.118	3.29	44.79	0.463	-4.98	62.24	0.490	0.40	54.41	0.651	-0.19	0.29	0.651	-0.16	0.28
1	0.092	-19.63	55.56	0.117	2.13	48.57	0.467	-4.34	54.61	0.491	0.74	48.48	0.651	-0.19	0.32	0.651	-0.16	0.30
2/8	0.088	-23.30	58.45	0.114	-0.61	49.57	0.468	-4.10	60.69	0.494	1.27	55.32	0.651	-0.16	0.33	0.652	-0.12	0.33
*5	0.018	-84.31	281.73	0.118	2.83	131.50	0.392	-19.58	281.08	0.492	0.81	134.10	0.640	-1.82	6.59	0.650	-0.36	1.06
*2/8	0.029	-74.61	246.80	0.124	8.61	122.84	0.395	-18.95	282.48	0.491	0.77	137.85	0.642	-1.62	5.99	0.651	-0.25	0.85
								$m = 500$										
5	-0.142	-223.75	660.10	0.116	1.06	6.10	0.233	-52.13	651.20	0.488	0.14	6.63	0.605	-7.32	23.56	0.652	-0.07	0.29
1	-0.030	-126.22	214.42	0.115	0.78	6.46	0.348	-28.60	200.80	0.488	0.11	6.46	0.647	-0.78	0.63	0.652	-0.02	0.29
2/8	-0.142	-224.31	663.65	0.115	0.29	6.73	0.235	-51.88	644.59	0.489	0.31	6.12	0.605	-7.33	23.73	0.652	0.00	0.28
*5	-0.147	-228.13	693.12	0.114	-0.24	16.19	0.231	-52.60	669.62	0.490	0.41	15.82	0.602	-7.78	27.45	0.652	-0.06	0.84
*2/8	-0.145	-226.72	684.68	0.117	1.86	16.86	0.230	-52.85	675.66	0.490	0.45	16.25	0.602	-7.67	27.03	0.652	-0.01	0.76

Table B.16: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a calibration variable and $f(c)$ was normally distributed. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	Mean	Bias	MSE
5	-0.137	-10.52	2840.66
1	-0.141	-8.16	543.63
2/8	-0.112	-27.27	2778.08
*5	-2.313	-1.59	1816.01
*2/8	-2.336	-0.62	1740.03
	$m = 500$		
5	-0.038	-75.24	138.40
1	-0.089	-41.89	48.34
2/8	-0.039	-74.77	136.83
*5	-2.249	-4.31	116.82
*2/8	-2.251	-4.22	113.70

Gumbel

Table B.17: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a calibration variable and $f(c)$ had a Gumbel distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
5	0.082	-27.99	60.42	0.109	-5.06	48.57	0.469	-3.86	54.68	0.495	1.53	50.20	0.651	-0.18	0.34	0.651	-0.13	0.32
1	0.093	-18.44	57.45	0.118	2.86	50.76	0.466	-4.47	57.73	0.490	0.51	51.57	0.651	-0.17	0.31	0.651	-0.14	0.30
2/8	0.089	-22.09	52.15	0.115	0.58	43.91	0.468	-3.98	58.50	0.494	1.38	52.91	0.651	-0.17	0.31	0.651	-0.13	0.30
*5	0.020	-82.71	284.51	0.117	2.42	131.37	0.399	-18.09	253.74	0.496	1.77	120.85	0.639	-2.00	8.04	0.650	-0.39	0.99
*2/8	0.021	-81.87	249.45	0.112	-2.31	118.70	0.411	-15.66	237.42	0.503	3.12	135.69	0.642	-1.58	5.21	0.649	-0.45	1.05
								$m = 500$										
5	-0.141	-223.20	657.15	0.115	0.79	5.93	0.233	-52.16	652.07	0.486	-0.25	7.07	0.605	-7.28	23.30	0.652	-0.05	0.30
1	-0.031	-127.22	217.23	0.114	-0.38	5.50	0.350	-28.33	196.60	0.489	0.33	5.62	0.647	-0.83	0.64	0.652	-0.04	0.31
2/8	-0.142	-223.66	660.11	0.115	0.14	7.07	0.235	-51.81	643.94	0.489	0.24	6.68	0.605	-7.25	23.20	0.652	-0.03	0.26
*5	-0.149	-229.91	704.29	0.114	-0.40	16.52	0.229	-53.03	681.50	0.491	0.64	18.40	0.599	-8.12	29.95	0.652	-0.06	0.91
*2/8	-0.149	-229.97	706.01	0.113	-1.05	16.49	0.228	-53.30	689.28	0.488	-0.02	18.59	0.600	-8.10	29.72	0.652	-0.03	0.79

Table B.18: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a calibration variable and $f(c)$ had a Gumbel distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	Mean	Bias	MSE
	$m = 6$		
5	-0.154	0.20	2584.85
1	-0.137	-10.48	510.19
2/8	-0.155	0.90	2809.27
*5	-2.329	-0.94	1926.27
*2/8	-2.333	-0.76	2183.85
	$m = 500$		
5	-0.038	-74.98	137.05
1	-0.089	-42.21	47.75
2/8	-0.038	-75.25	137.85
*5	-2.249	-4.34	118.27
*2/8	-2.247	-4.42	122.61

Laplace

Table B.19: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a calibration variable and $f(c)$ had a Laplace distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
5	0.088	-23.50	63.25	0.113	-0.83	54.19	0.467	-4.27	56.90	0.493	1.05	51.56	0.651	-0.24	0.32	0.651	-0.20	0.31
1	0.100	-12.54	51.86	0.124	8.10	48.92	0.464	-4.84	59.25	0.488	0.02	52.40	0.651	-0.15	0.34	0.651	-0.13	0.32
2/8	0.088	-23.17	58.68	0.114	-0.51	49.62	0.462	-5.20	58.70	0.488	0.13	50.48	0.651	-0.16	0.34	0.652	-0.12	0.31
*5	0.035	-69.58	222.54	0.127	10.77	118.43	0.394	-19.15	256.50	0.486	-0.32	117.59	0.643	-1.46	5.05	0.650	-0.30	0.88
*2/8	0.015	-86.96	279.64	0.109	-4.85	127.07	0.400	-18.01	241.53	0.494	1.22	117.33	0.640	-1.83	6.99	0.650	-0.34	1.03
								$m = 500$										
5	-0.141	-223.25	657.19	0.115	0.43	5.90	0.235	-51.87	644.43	0.486	-0.26	6.00	0.605	-7.29	23.44	0.652	-0.07	0.31
1	-0.029	-125.69	212.47	0.115	0.48	6.01	0.349	-28.41	197.90	0.488	-0.03	6.11	0.647	-0.82	0.64	0.652	-0.08	0.31
2/8	-0.142	-224.06	661.70	0.114	0.04	5.91	0.236	-51.70	640.67	0.487	-0.05	6.14	0.605	-7.28	23.41	0.652	-0.06	0.29
*5	-0.145	-226.32	682.92	0.116	1.60	16.25	0.230	-52.93	678.92	0.489	0.32	17.64	0.602	-7.72	27.45	0.652	-0.13	0.92
*2/8	-0.147	-228.05	692.73	0.113	-1.10	15.90	0.230	-52.83	675.47	0.488	-0.01	16.08	0.601	-7.84	28.28	0.652	-0.01	0.82

Table B.20: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a calibration variable and $f(c)$ had a Laplace distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	Mean	Bias	MSE
	$m = 6$		
5	-0.125	-18.42	2795.42
1	-0.140	-8.70	533.83
2/8	-0.197	28.33	2448.45
*5	-2.337	-0.56	2059.07
*2/8	-2.347	-0.14	1702.48
	$m = 500$		
5	-0.040	-74.17	134.00
1	-0.089	-42.29	48.77
2/8	-0.039	-74.67	136.63
*5	-2.250	-4.26	115.70
*2/8	-2.252	-4.19	111.36

Uniform

Table B.21: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a calibration variable and $f(c)$ had a uniform distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$										
5	0.092	-19.99	56.46	0.118	2.83	49.16	0.464	-4.88	56.30	0.490	0.50	49.05	0.652	-0.08	0.35	0.652	-0.05	0.34
1	0.089	-22.35	57.74	0.114	-0.50	49.41	0.470	-3.72	55.11	0.494	1.38	50.84	0.651	-0.15	0.31	0.652	-0.12	0.30
2/8	0.092	-19.84	60.48	0.118	2.96	52.94	0.469	-3.83	58.21	0.495	1.56	53.25	0.651	-0.16	0.33	0.652	-0.12	0.31
*5	0.021	-81.58	289.01	0.123	7.31	134.23	0.388	-20.51	288.44	0.489	0.24	130.62	0.639	-2.00	7.42	0.650	-0.36	0.93
*2/8	0.018	-84.47	304.82	0.119	3.99	143.28	0.385	-21.15	303.48	0.486	-0.38	141.05	0.639	-2.12	7.83	0.650	-0.41	1.04
								$m = 500$										
5	-0.142	-224.17	662.29	0.114	-0.39	6.09	0.233	-52.32	655.66	0.486	-0.31	6.00	0.605	-7.32	23.48	0.652	0.00	0.31
1	-0.030	-126.31	215.05	0.116	1.23	6.79	0.348	-28.69	201.73	0.487	-0.13	6.83	0.647	-0.75	0.60	0.652	-0.01	0.31
2/8	-0.142	-223.81	660.16	0.116	0.94	6.16	0.233	-52.16	651.70	0.487	-0.08	6.56	0.605	-7.33	23.66	0.652	-0.04	0.30
*5	-0.144	-225.68	677.62	0.116	1.79	15.62	0.229	-52.98	678.44	0.489	0.18	14.81	0.602	-7.68	26.94	0.652	-0.00	0.85
*2/8	-0.143	-225.34	677.58	0.117	1.87	16.51	0.226	-53.57	694.59	0.485	-0.50	15.78	0.602	-7.71	27.30	0.653	0.04	0.81

Table B.22: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a calibration variable and $f(c)$ had a uniform distribution. The number of centers in the training data, m, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included heteroscedasticity (i.e., non-constant $\sigma^2_{f(c)}$), denoted by "a/b" in the $\sigma^2_{f(c)}$ column; in these scenarios, half of the centers had $\sigma^2_{f(c)} = a$ and half had $\sigma^2_{f(c)} = b$. The MSE is multiplied by 10^4.

$\sigma^2_{f(c)}$	Mean	Bias	MSE
	$m = 6$		
5	-0.133	-13.42	2975.43
1	-0.136	-11.59	600.93
2/8	-0.165	7.44	2669.35
*5	-2.311	-1.68	1848.61
*2/8	-2.363	0.53	1683.09
	$m = 500$		
5	-0.040	-74.02	134.03
1	-0.089	-42.29	48.48
2/8	-0.039	-74.41	136.09
*5	-2.249	-4.31	116.14
*2/8	-2.251	-4.22	115.37

Confounding (Negative Correlation)

Normal

Table B.23: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.104	-9.10	59.55	0.122	6.21	59.00	0.470	-3.55	64.46	0.488	0.04	61.27	0.650	-0.29	0.49	0.651	-0.27	0.48
3	5	0.100	-12.51	72.89	0.115	0.63	72.04	0.474	-2.82	81.19	0.489	0.29	80.14	0.651	-0.27	0.82	0.651	-0.25	0.80
0.5/1.5	2/8	0.099	-13.83	63.43	0.116	1.63	61.18	0.476	-2.50	61.04	0.493	1.12	59.41	0.651	-0.17	0.39	0.651	-0.15	0.38
1/5	2/8	0.099	-13.87	78.82	0.113	-1.64	75.69	0.478	-2.08	82.39	0.492	0.82	82.70	0.651	-0.21	0.46	0.651	-0.19	0.44
								$m = 500$											
1	5	-0.073	-163.61	358.57	0.116	1.03	7.76	0.316	-35.31	305.30	0.486	-0.31	8.10	0.641	-1.78	2.00	0.653	0.09	0.36
3	5	-0.027	-123.83	210.29	0.116	1.46	8.82	0.362	-25.68	165.28	0.488	-0.01	8.93	0.648	-0.67	1.15	0.653	0.08	0.97
0.5/1.5	2/8	-0.076	-166.22	370.17	0.114	-0.73	7.47	0.315	-35.46	307.58	0.487	-0.18	8.16	0.640	-1.96	2.28	0.652	-0.02	0.38
1/5	2/8	-0.031	-127.19	219.76	0.114	-0.26	7.57	0.362	-25.71	166.18	0.489	0.18	9.00	0.647	-0.80	0.76	0.652	-0.03	0.40

Table B.24: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.132	-13.80	6332.20
3	5	-0.154	0.44	11359.50
0.5/1.5	2/8	-0.118	-22.99	6772.60
1/5	2/8	-0.241	57.17	12645.23
		$m = 500$		
1	5	-0.074	-52.02	105.60
3	5	-0.087	-43.09	148.56
0.5/1.5	2/8	-0.073	-52.31	106.10
1/5	2/8	-0.095	-37.91	125.76

Gumbel

Table B.25: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.101	-11.38	55.21	0.119	3.94	53.84	0.465	-4.56	68.62	0.483	-0.95	63.20	0.651	-0.16	0.44	0.651	-0.15	0.44
3	5	0.096	-16.11	80.01	0.111	-2.75	77.82	0.477	-2.16	85.17	0.493	0.99	83.95	0.650	-0.33	1.06	0.650	-0.31	1.04
0.5/1.5	2/8	0.097	-15.38	58.28	0.114	-0.26	54.60	0.472	-3.15	57.94	0.490	0.42	55.29	0.651	-0.19	0.37	0.651	-0.17	0.36
1/5	2/8	0.101	-11.63	78.41	0.114	-0.05	77.27	0.475	-2.70	87.44	0.488	0.09	86.08	0.652	-0.10	0.46	0.652	-0.09	0.45
								$m = 500$											
1	5	-0.074	-165.03	364.51	0.116	1.10	7.28	0.314	-35.66	309.95	0.486	-0.36	6.97	0.640	-1.96	2.28	0.652	-0.03	0.40
3	5	-0.029	-124.92	212.76	0.117	2.61	8.05	0.361	-26.03	171.06	0.488	0.08	9.62	0.647	-0.76	1.41	0.652	-0.03	1.03
0.5/1.5	2/8	-0.079	-168.87	382.07	0.113	-1.46	7.27	0.315	-35.47	307.60	0.488	0.14	7.37	0.638	-2.14	2.66	0.652	-0.05	0.35
1/5	2/8	-0.033	-128.50	225.88	0.113	-0.92	8.46	0.361	-26.07	170.54	0.488	0.07	8.78	0.647	-0.80	0.86	0.652	0.02	0.48

Table B.26: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.165	7.86	6362.19
3	5	-0.127	-16.90	11817.45
0.5/1.5	2/8	-0.106	-31.06	6119.61
1/5	2/8	-0.197	28.42	11399.69
		$m = 500$		
1	5	-0.101	-33.93	65.83
3	5	-0.177	15.49	97.39
0.5/1.5	2/8	-0.099	-35.39	66.38
1/5	2/8	-0.165	7.45	86.90

Laplace

Table B.27: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.103	-9.61	59.75	0.120	5.22	57.72	0.475	-2.55	62.01	0.492	0.92	60.10	0.651	-0.22	0.42	0.651	-0.20	0.41
3	5	0.097	-15.11	75.58	0.111	-2.81	72.48	0.477	-2.13	68.59	0.492	0.79	69.47	0.651	-0.18	1.82	0.651	-0.16	1.83
0.5/1.5	2/8	0.103	-9.98	61.40	0.120	4.54	60.46	0.471	-3.49	66.32	0.487	-0.11	63.27	0.651	-0.18	0.35	0.651	-0.16	0.35
1/5	2/8	0.097	-15.19	73.53	0.111	-3.25	71.59	0.483	-1.01	64.44	0.497	1.81	65.46	0.651	-0.18	0.42	0.651	-0.16	0.41
$m = 500$																			
1	5	-0.080	-170.20	387.24	0.114	-0.74	7.27	0.311	-36.15	319.94	0.488	-0.02	7.26	0.638	-2.22	2.83	0.652	-0.10	0.41
3	5	-0.030	-126.32	217.87	0.116	1.62	7.79	0.360	-26.16	171.03	0.488	0.10	8.03	0.648	-0.66	1.23	0.653	0.10	0.90
0.5/1.5	2/8	-0.081	-170.79	389.89	0.115	0.45	6.38	0.310	-36.49	324.89	0.488	0.01	7.02	0.638	-2.15	2.64	0.653	0.03	0.35
1/5	2/8	-0.038	-133.37	242.00	0.113	-1.34	8.05	0.355	-27.22	185.18	0.487	-0.15	8.54	0.646	-0.94	0.92	0.652	-0.01	0.40

Table B.28: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.156	1.71	6782.50
3	5	-0.098	-36.12	10958.32
0.5/1.5	2/8	-0.128	-16.86	6388.38
1/5	2/8	-0.123	-19.86	10595.53
		$m = 500$		
1	5	-0.066	-56.74	113.16
3	5	-0.098	-36.37	123.96
0.5/1.5	2/8	-0.070	-54.29	107.62
1/5	2/8	-0.094	-38.67	119.46

Uniform

Table B.29: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.102	-10.98	57.99	0.120	4.79	56.31	0.467	-4.25	72.80	0.485	-0.49	68.84	0.651	-0.18	0.43	0.651	-0.16	0.43
3	5	0.106	-7.54	88.24	0.122	6.63	88.64	0.468	-4.10	92.29	0.484	-0.70	89.14	0.651	-0.24	0.74	0.651	-0.22	0.73
0.5/1.5	2/8	0.099	-13.52	68.47	0.116	1.67	66.50	0.475	-2.66	69.57	0.492	0.94	68.10	0.651	-0.21	0.40	0.651	-0.18	0.40
1/5	2/8	0.098	-14.16	74.28	0.113	-1.20	72.59	0.475	-2.52	86.25	0.491	0.58	86.29	0.650	-0.28	0.48	0.651	-0.26	0.47
								$m = 500$											
1	5	-0.069	-160.20	342.56	0.116	1.64	6.52	0.320	-34.43	289.25	0.487	-0.18	7.12	0.641	-1.67	1.74	0.653	0.04	0.38
3	5	-0.029	-125.60	217.41	0.113	-1.45	10.91	0.362	-25.68	166.58	0.487	-0.15	10.08	0.647	-0.77	0.95	0.652	-0.01	0.60
0.5/1.5	2/8	-0.070	-161.43	348.14	0.114	0.05	6.73	0.321	-34.18	285.15	0.488	0.04	7.17	0.641	-1.77	1.95	0.652	-0.05	0.36
1/5	2/8	-0.025	-121.96	204.09	0.115	0.61	9.10	0.364	-25.31	161.59	0.486	-0.35	9.48	0.648	-0.65	0.61	0.653	0.05	0.38

Table B.30: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with negative correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.189	23.32	5972.91
3	5	-0.123	-19.50	12011.31
0.5/1.5	2/8	-0.117	-23.95	6242.38
1/5	2/8	-0.168	9.60	13186.25
		$m = 500$		
1	5	-0.078	-48.97	104.34
3	5	-0.098	-36.07	147.12
0.5/1.5	2/8	-0.077	-50.00	104.96
1/5	2/8	-0.092	-39.73	157.28

Confounding (No Correlation)

Normal

Table B.31: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.094	-17.63	67.96	0.112	-1.92	63.76	0.475	-2.57	71.74	0.493	1.15	70.99	0.651	-0.16	0.42	0.651	-0.13	0.41
3	5	0.107	-6.70	86.53	0.119	3.91	87.89	0.475	-2.62	77.75	0.487	-0.09	77.87	0.651	-0.23	0.88	0.651	-0.22	0.87
0.5/1.5	2/8	0.094	-18.22	59.71	0.112	-1.85	55.70	0.479	-1.84	63.76	0.498	2.06	64.17	0.651	-0.15	0.40	0.652	-0.13	0.39
1/5	2/8	0.102	-10.61	75.08	0.115	0.73	73.98	0.482	-1.08	83.00	0.495	1.57	82.51	0.651	-0.21	0.47	0.651	-0.19	0.46
$m = 500$																			
1	5	-0.037	-132.35	236.82	0.115	0.17	7.93	0.356	-26.92	179.50	0.490	0.50	7.87	0.646	-0.91	0.80	0.652	-0.05	0.38
3	5	0.013	-88.86	112.09	0.112	-1.69	9.06	0.407	-16.60	74.16	0.488	0.12	9.12	0.651	-0.24	0.78	0.653	0.03	0.71
0.5/1.5	2/8	-0.040	-134.79	244.34	0.112	-2.37	7.12	0.354	-27.47	186.58	0.488	0.06	7.63	0.646	-0.97	0.80	0.652	-0.07	0.34
1/5	2/8	0.012	-89.52	114.15	0.115	0.58	9.31	0.402	-17.49	80.98	0.487	-0.16	8.24	0.651	-0.25	0.45	0.653	0.03	0.40

Table B.32: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.145	-5.77	4014.54
3	5	-0.085	-44.64	8173.21
0.5/1.5	2/8	-0.118	-22.96	4569.70
1/5	2/8	-0.135	-12.10	8494.52
		$m = 500$		
1	5	-0.089	-41.72	80.52
3	5	-0.115	-25.10	106.78
0.5/1.5	2/8	-0.086	-43.93	80.25
1/5	2/8	-0.108	-29.90	93.79

Gumbel

Table B.33: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.103	-9.97	64.81	0.121	5.68	63.48	0.476	-2.36	70.37	0.494	1.30	68.58	0.650	-0.32	0.49	0.650	-0.30	0.47
3	5	0.099	-13.60	80.38	0.112	-2.54	78.45	0.485	-0.48	80.35	0.498	2.18	82.65	0.651	-0.21	1.32	0.651	-0.19	1.32
0.5/1.5	2/8	0.095	-16.56	61.39	0.113	-1.01	55.83	0.476	-2.39	54.09	0.494	1.30	53.26	0.651	-0.19	0.37	0.651	-0.17	0.36
1/5	2/8	0.101	-11.88	79.96	0.113	-1.37	78.44	0.484	-0.72	77.77	0.496	1.78	80.06	0.651	-0.23	0.44	0.651	-0.21	0.43
$m = 500$																			
1	5	-0.040	-134.61	244.86	0.116	1.75	7.18	0.348	-28.58	200.87	0.487	-0.04	7.11	0.646	-1.00	0.89	0.652	-0.06	0.41
3	5	0.010	-91.31	117.52	0.115	0.93	9.07	0.402	-17.57	81.64	0.490	0.37	8.36	0.650	-0.34	2.04	0.652	-0.04	2.05
0.5/1.5	2/8	-0.043	-137.51	254.84	0.114	-0.64	7.15	0.350	-28.23	197.58	0.490	0.47	7.92	0.646	-1.01	0.94	0.652	-0.02	0.35
1/5	2/8	0.008	-93.29	121.66	0.116	1.63	7.42	0.396	-18.87	92.01	0.486	-0.39	8.45	0.650	-0.37	0.71	0.652	-0.07	0.62

Table B.34: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.169	10.38	4880.57
3	5	-0.176	14.78	8365.69
0.5/1.5	2/8	-0.094	-38.70	4443.63
1/5	2/8	-0.192	25.14	7828.69
		$m = 500$		
1	5	-0.115	-25.12	45.98
3	5	-0.201	30.94	105.51
0.5/1.5	2/8	-0.113	-26.53	46.49
1/5	2/8	-0.190	24.05	88.67

Laplace

Table B.35: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.099	-13.57	57.03	0.116	1.77	54.96	0.476	-2.45	71.64	0.493	1.17	70.85	0.651	-0.20	0.43	0.651	-0.19	0.43
3	5	0.103	-10.37	72.34	0.115	0.16	70.58	0.479	-1.69	73.54	0.492	0.80	72.02	0.650	-0.38	1.89	0.650	-0.36	1.87
0.5/1.5	2/8	0.095	-16.63	62.00	0.112	-2.16	57.03	0.476	-2.47	62.29	0.492	0.95	60.65	0.651	-0.20	0.40	0.651	-0.18	0.39
1/5	2/8	0.097	-14.82	78.76	0.109	-4.85	76.88	0.485	-0.62	65.74	0.496	1.70	66.42	0.651	-0.24	0.45	0.651	-0.23	0.44
$m = 500$																			
1	5	-0.042	-136.98	252.97	0.115	0.68	7.31	0.347	-28.75	203.83	0.488	0.05	7.68	0.646	-0.98	1.00	0.653	0.03	0.45
3	5	0.007	-93.68	123.05	0.116	0.97	8.52	0.399	-18.19	87.28	0.488	0.16	8.72	0.650	-0.40	0.71	0.652	-0.08	0.62
0.5/1.5	2/8	-0.046	-140.02	264.06	0.114	-0.78	7.00	0.347	-28.84	204.67	0.489	0.29	7.05	0.645	-1.12	1.01	0.652	-0.07	0.38
1/5	2/8	0.002	-98.19	134.61	0.116	1.39	8.26	0.391	-19.77	101.73	0.486	-0.28	8.69	0.649	-0.45	2.67	0.652	-0.10	4.16

Table B.36: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.149	-2.80	4468.99
3	5	-0.146	-5.05	8138.90
0.5/1.5	2/8	-0.171	11.38	3896.49
1/5	2/8	-0.184	20.28	8136.75
		$m = 500$		
1	5	-0.088	-42.38	76.42
3	5	-0.105	-31.57	96.69
0.5/1.5	2/8	-0.085	-44.29	78.80
1/5	2/8	-0.110	-28.03	83.38

Uniform

Table B.37: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.097	-15.16	61.43	0.116	1.24	57.48	0.480	-1.57	57.54	0.499	2.31	59.42	0.652	-0.11	0.44	0.652	-0.08	0.43
3	5	0.098	-13.96	84.27	0.111	-2.74	83.95	0.484	-0.70	86.00	0.498	2.05	88.39	0.650	-0.34	0.72	0.650	-0.32	0.71
0.5/1.5	2/8	0.094	-17.86	61.48	0.112	-1.69	58.08	0.477	-2.12	64.58	0.496	1.72	64.04	0.651	-0.20	0.37	0.651	-0.18	0.35
1/5	2/8	0.106	-7.23	73.93	0.119	3.82	74.54	0.473	-3.10	77.24	0.485	-0.51	75.34	0.651	-0.25	0.45	0.651	-0.24	0.45
$m = 500$																			
1	5	-0.034	-129.48	225.53	0.114	-0.04	6.60	0.358	-26.69	174.88	0.489	0.26	6.37	0.647	-0.81	0.71	0.652	-0.03	0.37
3	5	0.017	-84.73	103.26	0.114	-0.41	10.00	0.409	-16.16	71.60	0.488	0.04	10.74	0.651	-0.28	0.67	0.652	-0.03	0.63
0.5/1.5	2/8	-0.034	-129.88	227.15	0.114	-0.27	7.17	0.358	-26.69	176.02	0.489	0.20	6.99	0.647	-0.84	0.74	0.652	-0.03	0.37
1/5	2/8	0.019	-83.18	99.77	0.117	2.16	9.79	0.410	-15.95	69.06	0.489	0.30	9.01	0.651	-0.23	0.40	0.652	-0.00	0.37

Table B.38: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with no correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.135	-11.94	4635.71
3	5	-0.131	-14.80	8795.47
0.5/1.5	2/8	-0.210	37.05	4809.41
1/5	2/8	-0.139	-9.46	8070.33
		$m = 500$		
1	5	-0.093	-39.53	72.06
3	5	-0.121	-21.22	110.96
0.5/1.5	2/8	-0.098	-36.02	66.91
1/5	2/8	-0.108	-29.45	108.98

Confounding (Positive Correlation)

Normal

Table B.39: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.094	-17.81	68.49	0.116	1.46	63.20	0.473	-2.99	63.87	0.495	1.54	60.99	0.652	-0.12	0.44	0.652	-0.09	0.43
3	5	0.105	-7.92	72.88	0.114	-0.51	75.35	0.487	-0.12	74.90	0.496	1.68	75.67	0.651	-0.15	1.06	0.651	-0.15	1.06
0.5/1.5	2/8	0.094	-18.10	58.61	0.114	-0.16	54.42	0.467	-4.27	66.48	0.487	-0.05	61.71	0.651	-0.22	0.43	0.651	-0.19	0.42
1/5	2/8	0.113	-1.54	73.03	0.120	5.22	73.56	0.484	-0.81	81.52	0.491	0.76	82.16	0.651	-0.19	0.47	0.651	-0.18	0.47
								$m = 500$											
1	5	-0.002	-101.54	141.29	0.114	-0.25	7.16	0.388	-20.53	107.01	0.488	-0.02	7.67	0.650	-0.42	0.58	0.652	-0.04	0.46
3	5	0.067	-41.66	32.03	0.114	-0.71	10.47	0.458	-6.08	17.83	0.487	-0.08	10.50	0.652	0.00	0.84	0.653	0.05	0.85
0.5/1.5	2/8	-0.003	-102.86	144.86	0.114	-0.44	6.84	0.386	-20.76	109.18	0.488	0.04	8.11	0.650	-0.38	0.36	0.652	0.02	0.29
1/5	2/8	0.062	-45.88	35.17	0.116	1.53	8.35	0.452	-7.42	21.45	0.488	-0.03	9.01	0.651	-0.14	0.42	0.652	-0.07	0.41

Table B.40: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was normal with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.146	-4.95	2260.53
3	5	-0.151	-1.30	3965.97
0.5/1.5	2/8	-0.127	-17.01	1996.49
1/5	2/8	-0.118	-23.01	4230.00
		$m = 500$		
1	5	-0.106	-30.95	43.80
3	5	-0.131	-14.67	64.60
0.5/1.5	2/8	-0.100	-34.86	51.94
1/5	2/8	-0.140	-8.82	51.07

Gumbel

Table B.41: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.100	-12.72	68.96	0.121	5.86	66.10	0.464	-4.79	70.43	0.485	-0.46	63.72	0.651	-0.22	0.40	0.651	-0.19	0.39
3	5	0.107	-6.85	71.03	0.115	0.92	70.04	0.487	-0.23	81.76	0.496	1.61	82.96	0.651	-0.14	1.30	0.651	-0.13	1.29
0.5/1.5	2/8	0.097	-15.60	66.69	0.118	2.73	62.27	0.469	-3.86	65.67	0.490	0.47	59.93	0.651	-0.26	0.38	0.651	-0.22	0.36
1/5	2/8	0.113	-1.42	69.47	0.120	4.84	68.14	0.489	0.26	66.90	0.496	1.75	66.14	0.651	-0.17	0.50	0.651	-0.16	0.50
								$m = 500$											
1	5	-0.007	-105.85	152.93	0.114	-0.67	7.08	0.383	-21.41	115.65	0.488	0.03	7.83	0.649	-0.44	0.52	0.652	0.02	0.41
3	5	0.059	-48.03	38.18	0.114	-0.35	8.84	0.452	-7.28	21.40	0.489	0.35	9.71	0.652	-0.04	0.64	0.653	0.02	0.64
0.5/1.5	2/8	-0.005	-104.42	149.67	0.118	2.78	7.25	0.381	-21.83	120.33	0.488	0.08	8.02	0.649	-0.49	0.50	0.652	-0.06	0.39
1/5	2/8	0.054	-52.82	44.12	0.114	0.05	7.91	0.446	-8.62	25.28	0.488	0.09	8.15	0.652	-0.12	0.41	0.652	-0.04	0.41

Table B.42: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Gumbel with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.153	-0.53	2190.06
3	5	-0.153	-0.03	3729.06
0.5/1.5	2/8	-0.145	-5.66	2127.58
1/5	2/8	-0.105	-31.23	4403.84
		$m = 500$		
1	5	-0.126	-18.00	32.51
3	5	-0.215	40.42	87.65
0.5/1.5	2/8	-0.127	-16.90	29.13
1/5	2/8	-0.206	34.08	78.03

Laplace

Table B.43: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
$m = 6$																			
1	5	0.101	-11.97	57.62	0.122	6.30	54.82	0.470	-3.72	64.48	0.490	0.51	59.37	0.651	-0.21	0.53	0.651	-0.18	0.52
3	5	0.107	-6.77	78.73	0.117	2.00	78.58	0.485	-0.60	77.41	0.495	1.48	76.71	0.652	-0.12	3.56	0.652	-0.11	3.56
0.5/1.5	2/8	0.100	-12.35	64.72	0.120	4.45	59.64	0.468	-4.07	68.76	0.487	-0.15	64.83	0.651	-0.22	0.42	0.651	-0.19	0.41
1/5	2/8	0.113	-0.94	59.58	0.122	6.83	60.79	0.485	-0.59	64.34	0.494	1.20	62.39	0.652	-0.12	0.46	0.652	-0.11	0.46
$m = 500$																			
1	5	-0.007	-106.15	154.08	0.116	1.22	7.40	0.380	-22.10	122.75	0.487	-0.06	7.30	0.650	-0.40	0.63	0.653	0.04	0.52
3	5	0.054	-52.74	44.01	0.114	-0.75	8.41	0.446	-8.46	24.48	0.488	0.10	8.18	0.652	-0.07	1.52	0.652	0.01	1.51
0.5/1.5	2/8	-0.011	-109.75	164.43	0.112	-2.23	6.90	0.382	-21.73	119.69	0.490	0.44	8.11	0.649	-0.53	0.45	0.652	-0.04	0.29
1/5	2/8	0.051	-55.39	48.15	0.117	2.60	8.11	0.439	-9.90	30.38	0.488	0.05	7.92	0.652	-0.08	0.44	0.652	0.01	0.43

Table B.44: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was Laplace with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.161	4.93	2432.85
3	5	-0.173	13.03	4223.45
0.5/1.5	2/8	-0.157	2.14	2328.40
1/5	2/8	-0.142	-7.65	4262.70
		$m = 500$		
1	5	-0.102	-33.75	50.83
3	5	-0.133	-13.52	53.08
0.5/1.5	2/8	-0.101	-34.02	48.38
1/5	2/8	-0.128	-16.25	51.94

Uniform

Table B.45: Mean, percent bias, and mean squared error (MSE) of the coefficient estimates from random intercept logistic regression ($\hat{\tau}_1, \hat{\tau}_2$) and fixed intercept logistic regression ($\hat{\beta}_1, \hat{\beta}_2$) and center-specific AUC ($AUC_c(\cdot)$) across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	$\hat{\tau}_1$			$\hat{\beta}_1$			$\hat{\tau}_2$			$\hat{\beta}_2$			$AUC_c(\hat{\boldsymbol{\tau}})$			$AUC_c(\hat{\boldsymbol{\beta}})$		
		Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE	Mean	Bias	MSE
								$m = 6$											
1	5	0.094	-17.73	64.46	0.117	2.39	60.74	0.469	-3.74	62.06	0.493	1.03	59.05	0.651	-0.18	0.49	0.651	-0.14	0.49
3	5	0.110	-3.88	81.96	0.116	1.51	81.13	0.481	-1.28	75.02	0.488	0.11	80.90	0.651	-0.23	0.77	0.651	-0.23	0.78
0.5/1.5	2/8	0.091	-20.31	68.73	0.115	0.42	62.68	0.471	-3.34	59.35	0.495	1.52	58.37	0.651	-0.19	0.38	0.651	-0.16	0.36
1/5	2/8	0.113	-1.32	70.58	0.119	4.18	70.02	0.486	-0.38	72.73	0.492	0.96	76.58	0.652	-0.13	0.44	0.652	-0.12	0.44
								$m = 500$											
1	5	0.001	-99.48	135.59	0.115	0.87	7.32	0.389	-20.21	103.65	0.488	0.15	7.70	0.650	-0.41	0.51	0.652	-0.03	0.40
3	5	0.069	-40.02	29.85	0.111	-2.81	10.12	0.463	-5.00	14.30	0.489	0.23	10.07	0.652	-0.06	0.64	0.652	-0.03	0.63
0.5/1.5	2/8	0.000	-99.80	136.88	0.115	0.55	7.90	0.389	-20.33	104.40	0.488	0.04	7.62	0.650	-0.38	0.43	0.652	-0.03	0.35
1/5	2/8	0.070	-39.24	27.83	0.116	1.80	8.82	0.459	-5.88	15.84	0.488	0.03	8.65	0.652	-0.11	0.37	0.652	-0.07	0.37

Table B.46: Mean, percent bias, and mean squared error (MSE) of the fixed intercept estimate ($\hat{\tau}_0$) from random intercept logistic regression across 500 simulations when center was a confounder and the joint distribution of logit(γ_c) and $f(c)$ was uniform with positive correlation between logit(γ_c) and $f(c)$. The number of centers in the training data, m, the variance of logit(γ_c), $\sigma^2_{\gamma_c}$, and the variance of $f(c)$, $\sigma^2_{f(c)}$, were varied. Some settings included non-constant $\sigma^2_{\gamma_c}$, denoted by "a/b" in the $\sigma^2_{\gamma_c}$ column; in these scenarios, half of the centers had $\sigma^2_{\gamma_c} = a$ and half had $\sigma^2_{\gamma_c} = b$. Likewise, some settings included non-constant $\sigma^2_{f(c)}$. The MSE is multiplied by 10^4.

$\sigma^2_{\gamma_c}$	$\sigma^2_{f(c)}$	Mean	Bias	MSE
		$m = 6$		
1	5	-0.163	6.38	2209.13
3	5	-0.130	-15.12	3874.74
0.5/1.5	2/8	-0.162	5.32	2393.64
1/5	2/8	-0.131	-14.28	4849.74
		$m = 500$		
1	5	-0.107	-30.49	48.02
3	5	-0.140	-8.83	74.08
0.5/1.5	2/8	-0.103	-32.74	50.48
1/5	2/8	-0.133	-13.28	65.31

B.4 Chapter 5

B.4.1 Direct Maximization of aAUC

Robust Logistic Regression

The tables below present the full results for the simulations reported in Section 5.4.1, where robust logistic regression estimates were used as the starting values for the SaAUC method.

Table B.47: 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 ($\hat{\boldsymbol{\theta}}_{GLM}$), robust logistic regression ($\hat{\boldsymbol{\theta}}_{rGLM}$), and the SaAUC method ($\hat{\boldsymbol{\theta}}_{SaAUC}$) for $m = 6$. 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
				Scenario I					
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)
				Scenario II					
Yes	0.6240	0.6060	0.6420	0.6473	0.6294	0.6649	0.6851	0.6678	0.7021
	(0.012)	(0.013)	(0.012)	(0.030)	(0.030)	(0.029)	(0.009)	(0.009)	(0.009)
No	0.7033	0.6867	0.7199	0.7033	0.6867	0.7199	0.7031	0.6865	0.7198
	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)	(0.002)	(0.004)	(0.004)
				Scenario III					
Yes	0.5952	0.5426	0.6451	0.6212	0.5755	0.6643	0.6596	0.6270	0.6900
	(0.024)	(0.027)	(0.022)	(0.037)	(0.045)	(0.029)	(0.010)	(0.013)	(0.010)
No	0.6728	0.6381	0.7048	0.6727	0.6380	0.7048	0.6724	0.6377	0.7046
	(0.003)	(0.007)	(0.007)	(0.003)	(0.007)	(0.007)	(0.003)	(0.008)	(0.007)
				Scenario IV					
Yes	0.6384	0.6118	0.6643	0.6609	0.6358	0.6855	0.6954	0.6731	0.7173
	(0.011)	(0.012)	(0.012)	(0.027)	(0.029)	(0.026)	(0.009)	(0.010)	(0.009)
No	0.7158	0.6934	0.7373	0.7158	0.6934	0.7373	0.7156	0.6931	0.7372
	(0.002)	(0.005)	(0.004)	(0.002)	(0.005)	(0.004)	(0.002)	(0.005)	(0.004)

Table B.48: 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 ($\hat{\boldsymbol{\theta}}_{GLM}$), robust logistic regression ($\hat{\boldsymbol{\theta}}_{rGLM}$), and the SaAUC method ($\hat{\boldsymbol{\theta}}_{SaAUC}$) for $m = 50$. 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
				Scenario I					
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)
				Scenario II					
Yes	0.6236	0.5449	0.6995	0.6473	0.5691	0.7217	0.6842	0.6075	0.7562
	(0.008)	(0.013)	(0.012)	(0.026)	(0.029)	(0.026)	(0.004)	(0.011)	(0.009)
No	0.7036	0.6298	0.7727	0.7036	0.6298	0.7727	0.7034	0.6296	0.7725
	(0.001)	(0.010)	(0.008)	(0.001)	(0.010)	(0.008)	(0.001)	(0.010)	(0.008)
				Scenario III					
Yes	0.5909	0.4834	0.6941	0.6172	0.5150	0.7145	0.6603	0.5705	0.7438
	(0.015)	(0.019)	(0.017)	(0.032)	(0.040)	(0.026)	(0.003)	(0.013)	(0.010)
No	0.6734	0.5851	0.7553	0.6734	0.5851	0.7553	0.6732	0.5846	0.7554
	(0.002)	(0.011)	(0.010)	(0.002)	(0.011)	(0.010)	(0.002)	(0.011)	(0.010)
				Scenario IV					
Yes	0.6375	0.5540	0.7171	0.6593	0.5777	0.7366	0.6942	0.6157	0.7680
	(0.007)	(0.013)	(0.012)	(0.024)	(0.028)	(0.023)	(0.004)	(0.011)	(0.009)
No	0.7162	0.6397	0.7876	0.7162	0.6397	0.7876	0.7160	0.6394	0.7876
	(0.001)	(0.010)	(0.008)	(0.001)	(0.010)	(0.008)	(0.001)	(0.010)	(0.008)

Table B.49: 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 ($\hat{\boldsymbol{\theta}}_{GLM}$), robust logistic regression ($\hat{\boldsymbol{\theta}}_{rGLM}$), and the SaAUC method ($\hat{\boldsymbol{\theta}}_{SaAUC}$) for $m = 500$. 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
				Scenario I					
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)
				Scenario II					
Yes	0.6221	0.4680	0.7656	0.6338	0.4798	0.7760	0.6795	0.5287	0.8148
	(0.004)	(0.014)	(0.012)	(0.015)	(0.020)	(0.017)	(0.004)	(0.015)	(0.011)
No	0.7039	0.5580	0.8336	0.7039	0.5580	0.8336	0.7037	0.5578	0.8334
	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)
				Scenario III					
Yes	0.5884	0.4076	0.7560	0.6046	0.4252	0.7692	0.6595	0.4951	0.8067
	(0.008)	(0.017)	(0.014)	(0.019)	(0.026)	(0.020)	(0.002)	(0.016)	(0.012)
No	0.6737	0.5135	0.8168	0.6737	0.5135	0.8167	0.6736	0.5129	0.8171
	(0.001)	(0.014)	(0.012)	(0.001)	(0.014)	(0.012)	(0.001)	(0.014)	(0.012)
				Scenario IV					
Yes	0.6370	0.4792	0.7818	0.6480	0.4909	0.7911	0.6888	0.5358	0.8249
	(0.003)	(0.015)	(0.013)	(0.013)	(0.021)	(0.017)	(0.004)	(0.015)	(0.011)
No	0.7164	0.5690	0.8447	0.7164	0.5690	0.8447	0.7162	0.5686	0.8446
	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)	(0.001)	(0.014)	(0.010)

Conditional Logistic Regression

The tables below present the full results for the simulations reported in Section 5.4.1, where conditional logistic regression estimates were used both as the starting values for the SaAUC method and for comparison with the SaAUC method.

Table B.50: 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 ($\hat{\boldsymbol{\theta}}_{SaAUC}$) for $m = 50$. Conditional logistic regression estimates were used as the starting values for the SaAUC method.

Outliers	$aAUC(\hat{\boldsymbol{\theta}}_{GLM})$	$AUC_c(\hat{\boldsymbol{\theta}}_{GLM})$ Min	Max	$aAUC(\hat{\boldsymbol{\theta}}_{SaAUC})$	$AUC_c(\hat{\boldsymbol{\theta}}_{SaAUC})$ Min	Max
			Scenario I			
Yes	0.6233 (0.008)	0.5444 (0.014)	0.6992 (0.012)	0.6824 (0.004)	0.6062 (0.011)	0.7547 (0.009)
No	0.7036 (0.001)	0.6301 (0.009)	0.7731 (0.008)	0.7035 (0.001)	0.6299 (0.010)	0.7730 (0.008)
			Scenario II			
Yes	0.6236 (0.008)	0.5450 (0.013)	0.6995 (0.012)	0.6823 (0.004)	0.6054 (0.011)	0.7544 (0.009)
No	0.7036 (0.001)	0.6298 (0.010)	0.7727 (0.008)	0.7034 (0.001)	0.6296 (0.010)	0.7725 (0.008)
			Scenario III			
Yes	0.5909 (0.015)	0.4834 (0.019)	0.6940 (0.017)	0.6596 (0.003)	0.5688 (0.013)	0.7440 (0.010)
No	0.6734 (0.002)	0.5851 (0.011)	0.7553 (0.010)	0.6732 (0.002)	0.5846 (0.011)	0.7554 (0.010)
			Scenario IV			
Yes	0.6376 (0.007)	0.5540 (0.013)	0.7171 (0.012)	0.6920 (0.004)	0.6130 (0.011)	0.7663 (0.009)
No	0.7162 (0.001)	0.6397 (0.010)	0.7876 (0.008)	0.7160 (0.001)	0.6394 (0.010)	0.7876 (0.008)

Table B.51: 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 ($\hat{\boldsymbol{\theta}}_{SaAUC}$) for $m = 500$. Conditional logistic regression estimates were used as the starting values for the SaAUC method.

Outliers	$aAUC(\hat{\boldsymbol{\theta}}_{GLM})$	$AUC_c(\hat{\boldsymbol{\theta}}_{GLM})$ Min	$AUC_c(\hat{\boldsymbol{\theta}}_{GLM})$ Max	$aAUC(\hat{\boldsymbol{\theta}}_{SaAUC})$	$AUC_c(\hat{\boldsymbol{\theta}}_{SaAUC})$ Min	$AUC_c(\hat{\boldsymbol{\theta}}_{SaAUC})$ Max
			Scenario I			
Yes	0.6221 (0.004)	0.4684 (0.015)	0.7659 (0.013)	0.6764 (0.003)	0.5253 (0.015)	0.8128 (0.012)
No	0.7038 (0.001)	0.5574 (0.014)	0.8330 (0.010)	0.7037 (0.001)	0.5573 (0.014)	0.8329 (0.010)
			Scenario II			
Yes	0.6221 (0.004)	0.4680 (0.014)	0.7656 (0.012)	0.6763 (0.003)	0.5248 (0.015)	0.8122 (0.011)
No	0.7039 (0.001)	0.5580 (0.014)	0.8336 (0.010)	0.7037 (0.001)	0.5578 (0.014)	0.8334 (0.010)
			Scenario III			
Yes	0.5884 (0.008)	0.4076 (0.017)	0.7560 (0.014)	0.6580 (0.003)	0.4921 (0.016)	0.8065 (0.012)
No	0.6737 (0.001)	0.5135 (0.014)	0.8168 (0.012)	0.6736 (0.001)	0.5129 (0.014)	0.8171 (0.012)
			Scenario IV			
Yes	0.6370 (0.003)	0.4792 (0.015)	0.7818 (0.013)	0.6853 (0.003)	0.5318 (0.015)	0.8222 (0.011)
No	0.7164 (0.001)	0.5690 (0.014)	0.8447 (0.010)	0.7162 (0.001)	0.5686 (0.014)	0.8446 (0.010)

B.4.2 Penalized Estimation Examples

The figures below include several additional examples of the penalized estimation method described in Section 5.4.2. The layout of these figures is the same as described in Section 5.4.2.

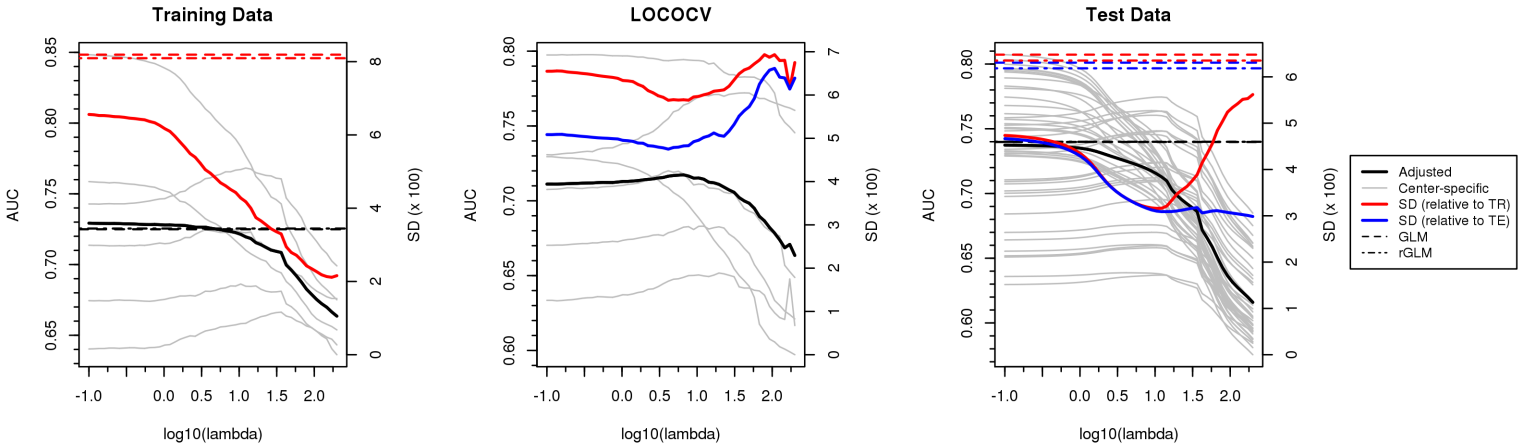

Figure B.14: Penalized estimation example 8. In this example, the LOCOCV procedure does a nice job capturing the trends seen in the test data.

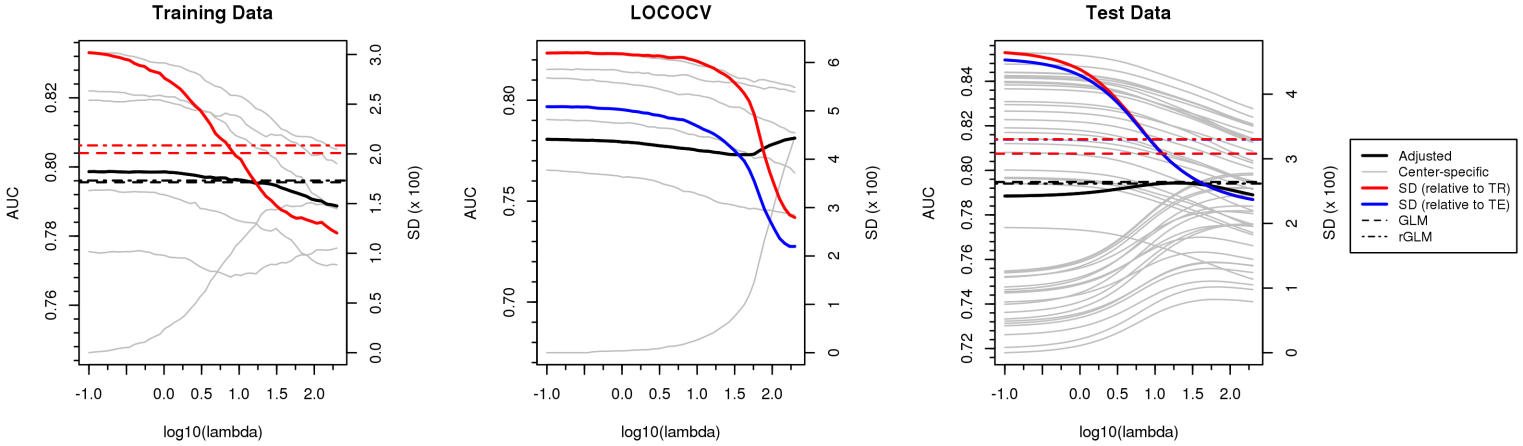

Figure B.15: Penalized estimation example 9. In this example, there is a clear benefit to penalization.

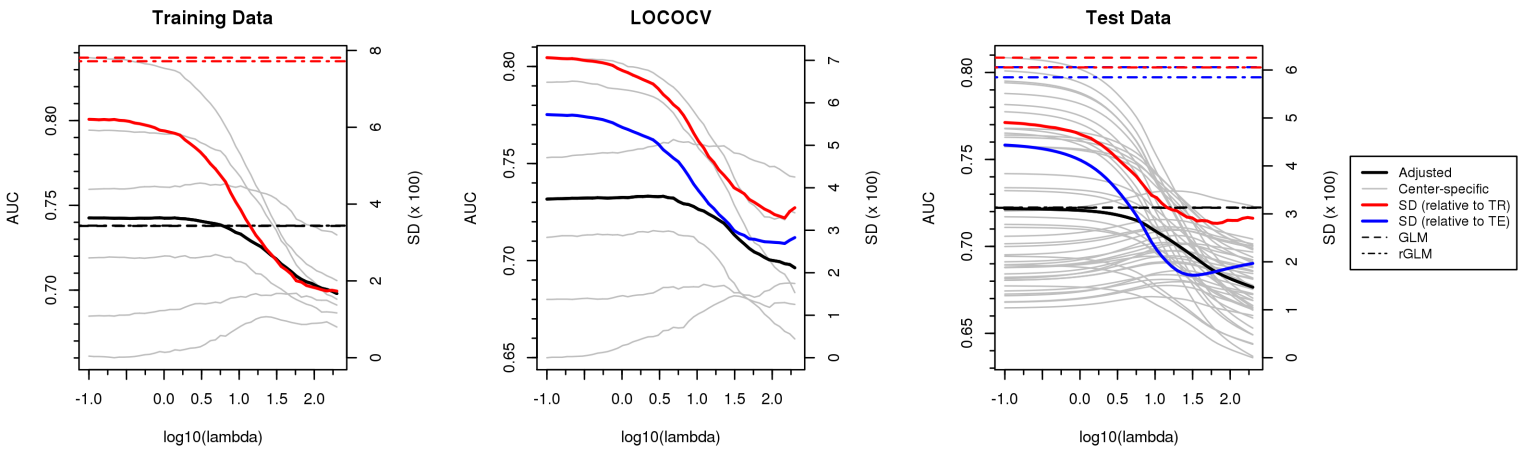

Figure B.16: Penalized estimation example 10. In this example, there is a clear benefit to penalization for a range of λ values.

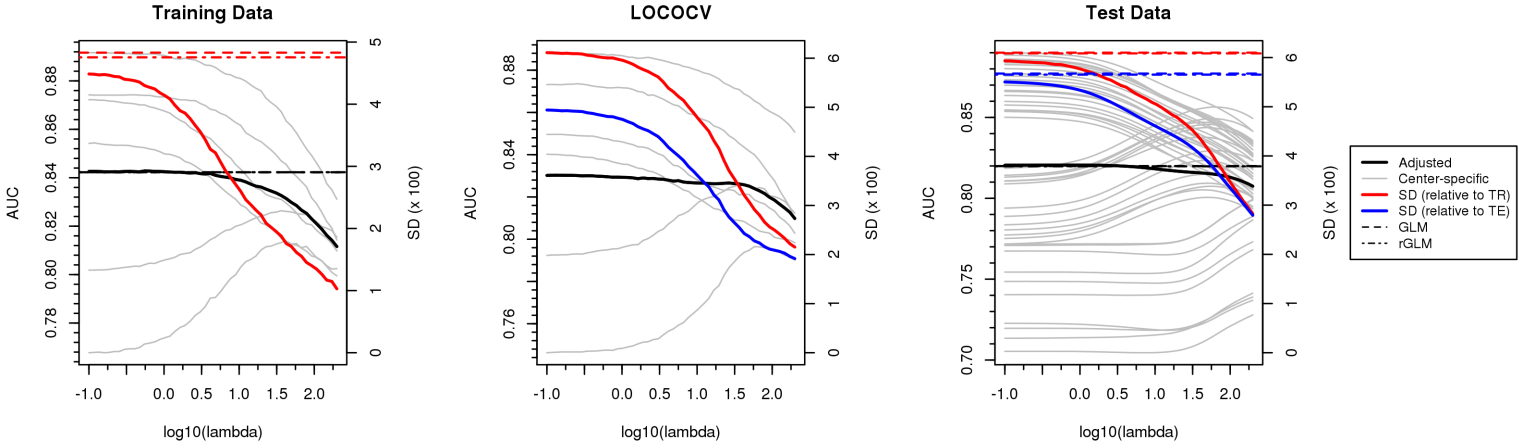

Figure B.17: Penalized estimation example 11. In this example, there is a definite benefit to penalization.

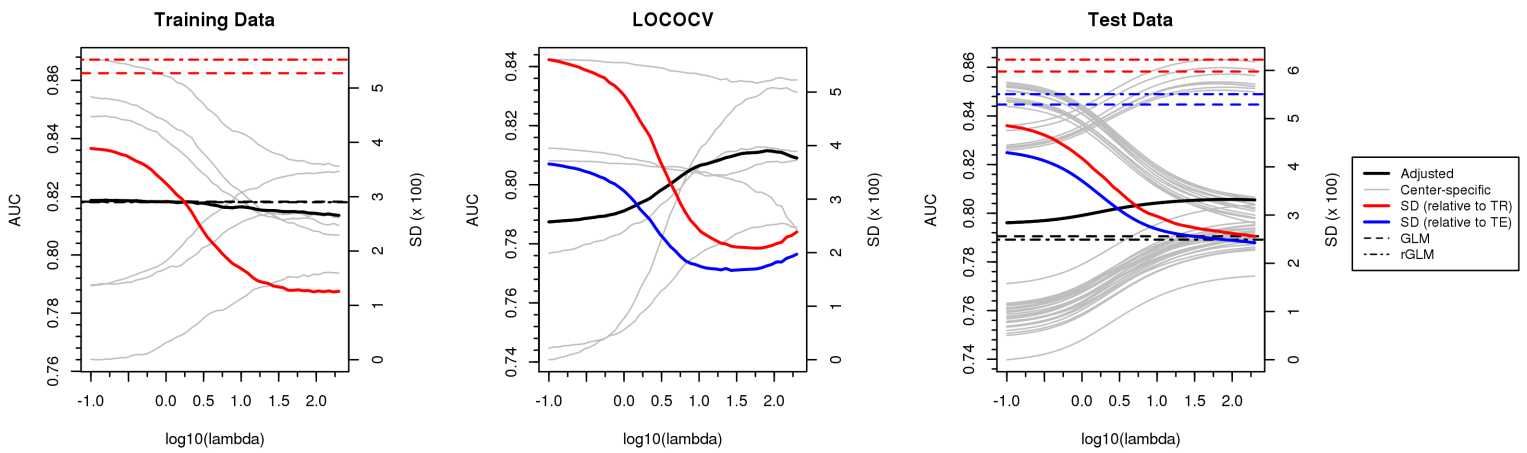

Figure B.18: Penalized estimation example 12. Here, there is a clear benefit to penalization, and the overall performance even increases slightly as λ increases.

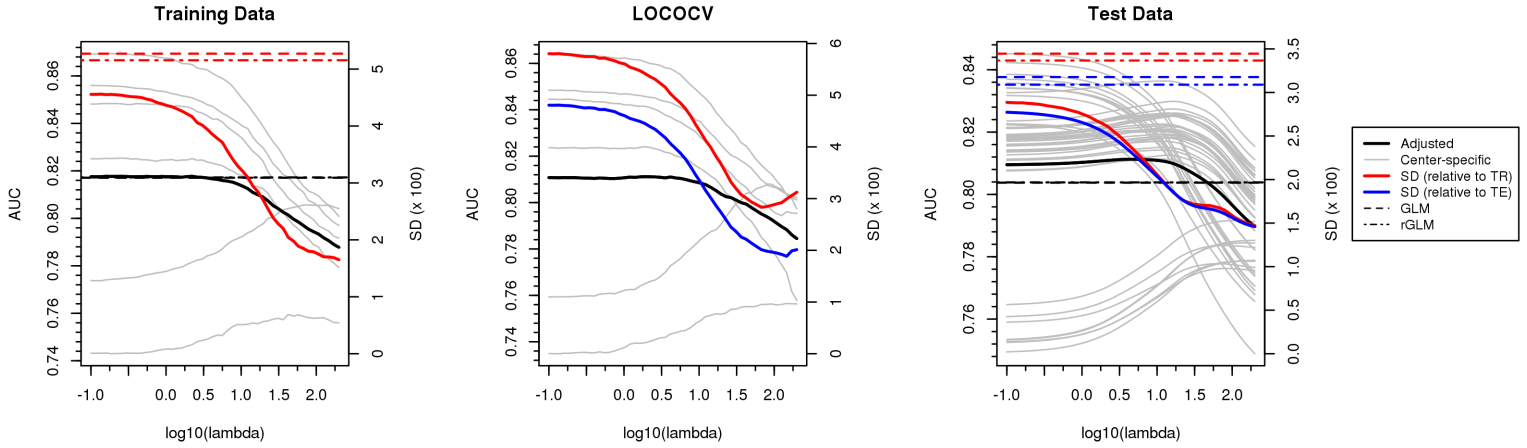

Figure B.19: Penalized estimation example 13. In this example, there is a definite benefit to penalization for a range of λ values.

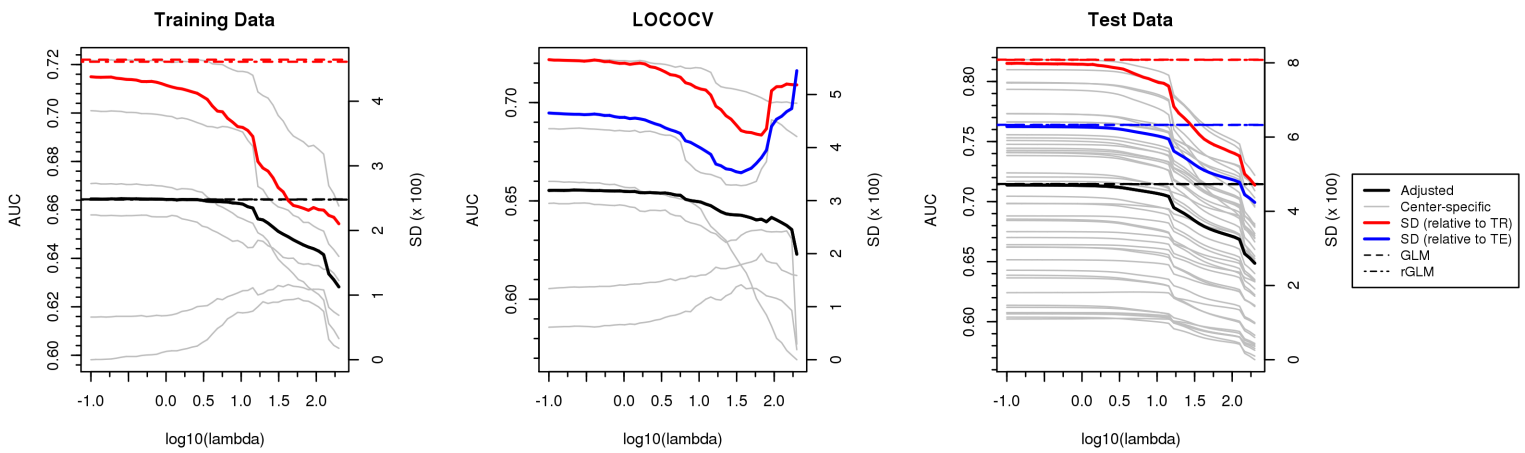

Figure B.20: Penalized estimation example 14. In this example, the LOCOCV procedure returns somewhat inconclusive results, making the choice of λ less clear.

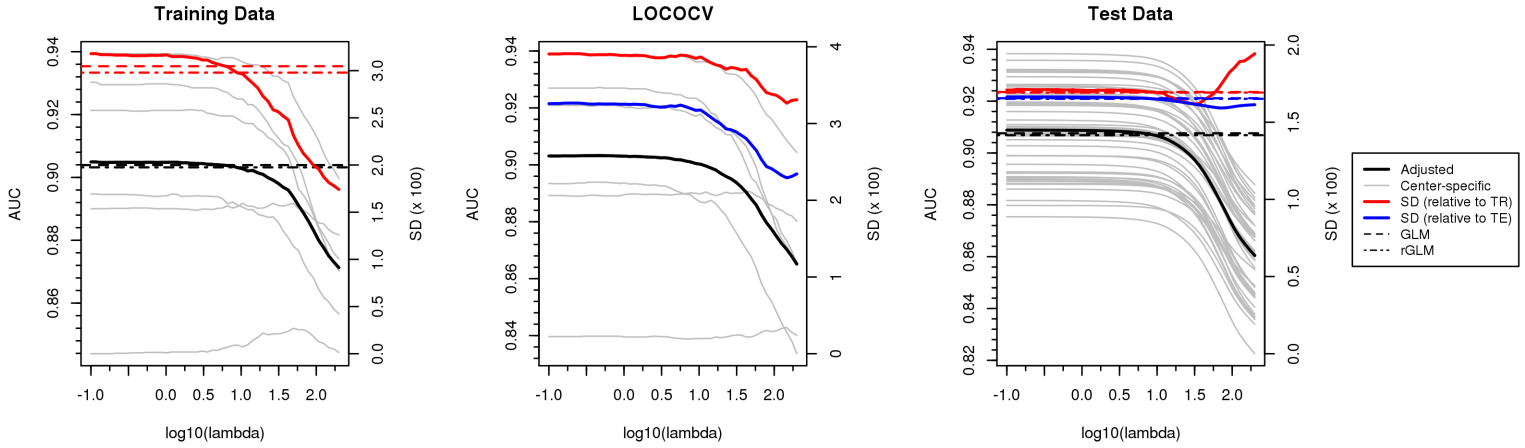

Figure B.21: Penalized estimation example 15. In this example, the LOCOCV procedure is a bit misleading, when compared to the results in test data.

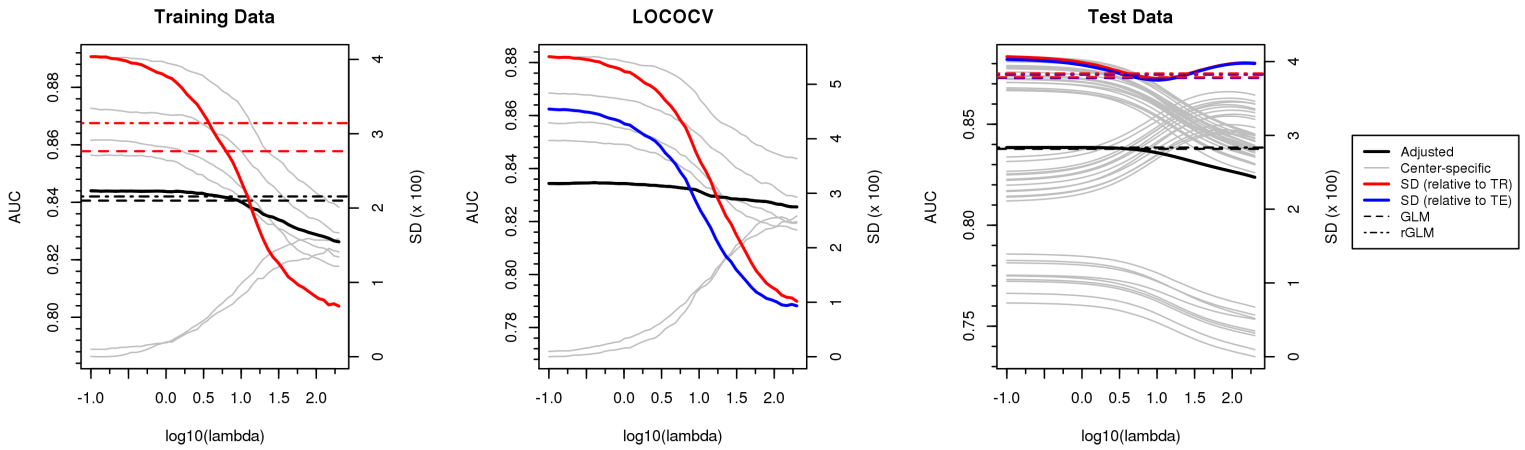

Figure B.22: Penalized estimation example 16. In this example, the LOCOCV procedure is a bit misleading, when compared to the results in test data.

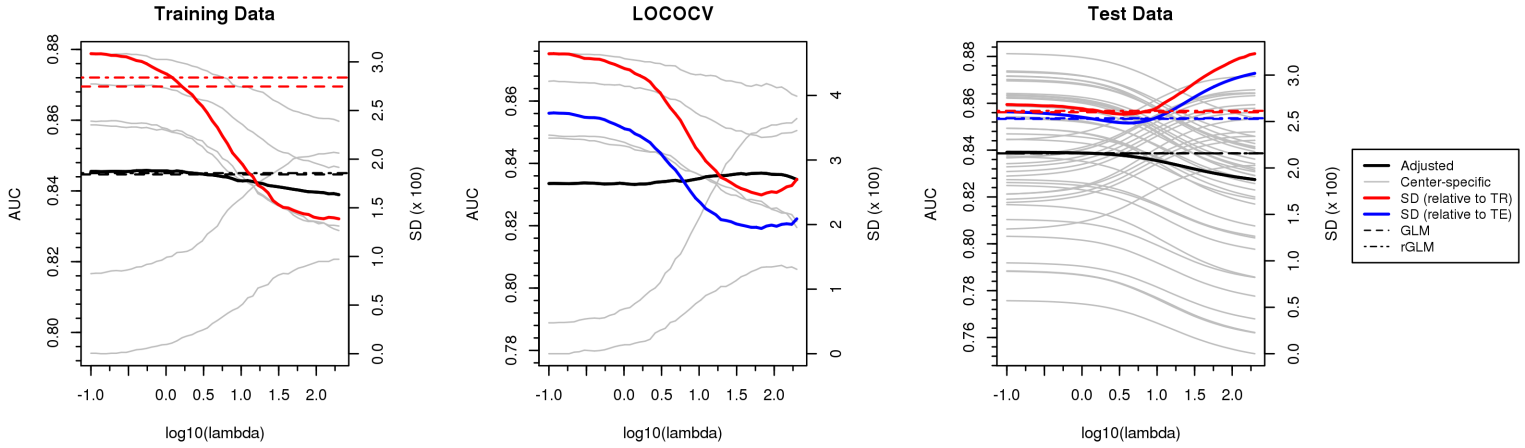

Figure B.23: Penalized estimation example 17. In this example, the variability in performance in test data increases slightly with increasing λ.

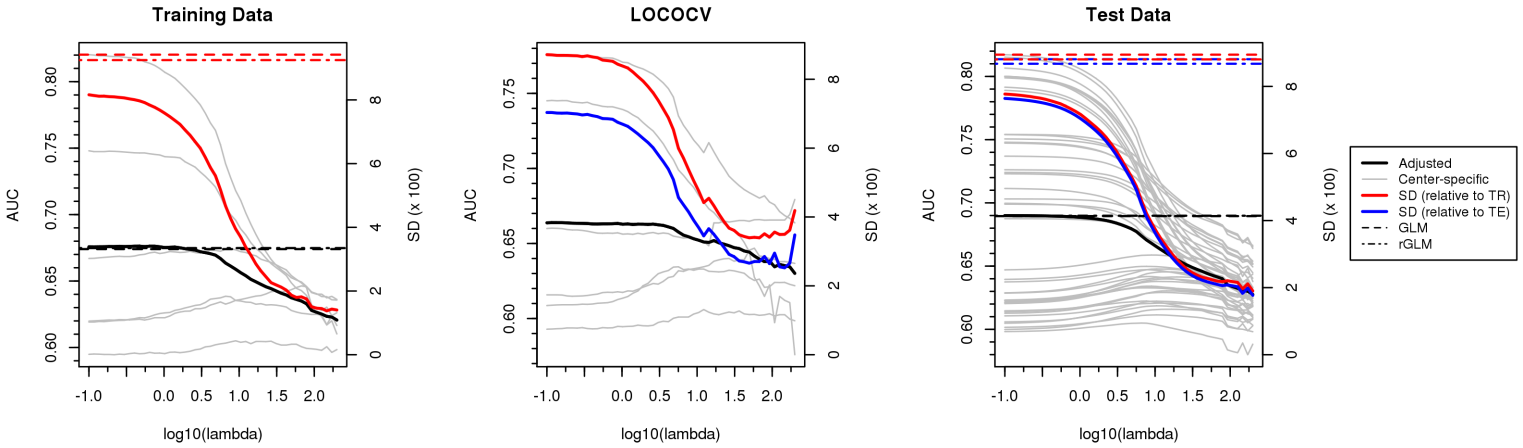

Figure B.24: Penalized estimation example 18. In this example, the overall performance in test data decreases more rapidly with increasing λ than is suggested by the LOCOCV results.

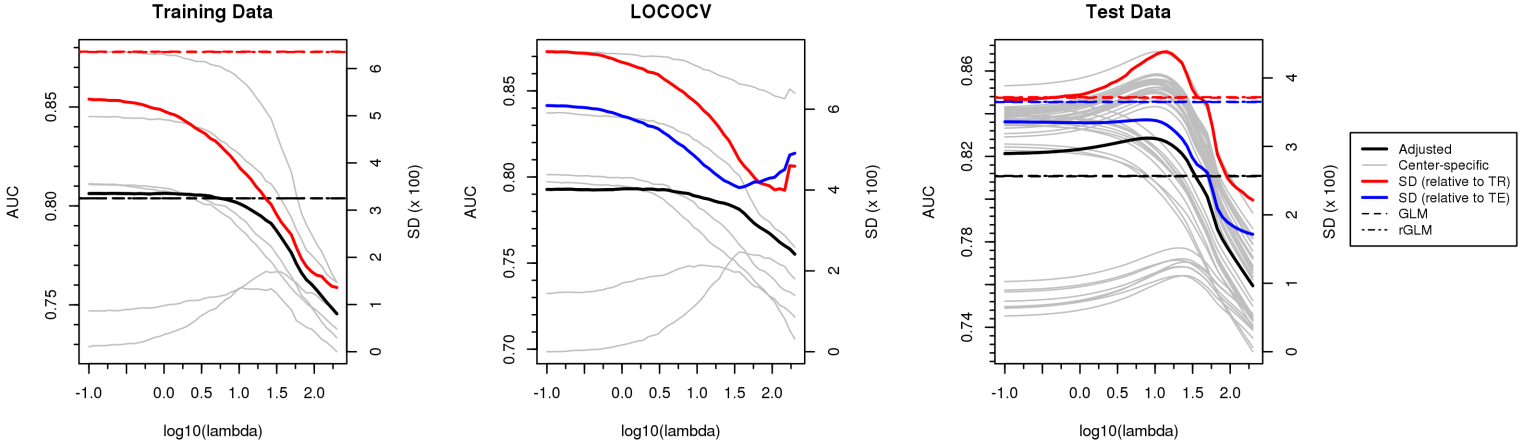

Figure B.25: Penalized estimation example 19. In this example, there is a strange relationship between λ and variability in performance in the test data that is not well-captured by the LOCOCV procedure.

VITA

Allison Meisner was raised in Glastonbury, Connecticut and received a Bachelors of Science in Biomedical Engineering from the University of Connecticut in 2008. She subsequently received a Master of Arts in Biostatistics from Boston University in 2011, and worked at the Harvard School of Public Health prior to moving to Seattle. She was awarded a Doctor of Philosophy in Biostatistics from the University of Washington in 2017.