6-29-2017

Biomarker Combinations for Diagnosis and Prognosis in Multicenter Studies: Principles and Methods

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

Chirag R. Parikh
Yale University, chirag.parikh@yale.edu

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

Biomarker Combinations for Diagnosis and Prognosis in Multicenter Studies: Principles and Methods

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

[1]Department of Biostatistics, University of Washington, Seattle, Washington

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

[3]Department of Internal Medicine, Veterans Affairs Medical Center, West Haven, Connecticut

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, the 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 a biomarker combination should be constructed and evaluated. We show that

Short Title: Biomarker Combinations in Multicenter Studies
Corresponding Author: Allison Meisner
Box 357232, University of Washington, Seattle, WA 98195-7232, USA
meisnera@uw.edu
Supplementary Material available at allisonmeisner.com

ignoring center, which is frequently done by clinical researchers, is often not appropriate. The limited statistical literature proposes using random intercept logistic regression models, an approach that we demonstrate is generally inadequate and may be 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.
Keywords: biomarkers, combinations, diagnosis, multicenter, prognosis

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 effect of differences across centers.[1] The literature on multicenter studies is especially extensive for randomized trials, where the need for careful design and analysis of such studies is widely acknowledged.[1]

Multicenter biomarker studies are increasingly common as investigators seek to increase power and generalizability (e.g., Feldstein et al.[2], Degos et al.[3], Nickolas et al.[4]). However, in contrast to randomized trials, the literature on multicenter biomarker studies 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 combinations of biomarkers. A multicenter study design can have implications for both the construction and evaluation of biomarker combinations.

Center plays a unique role in biomarker studies, where the goal is generally prediction. 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. 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. However, not all methods for accounting for center are suitable for biomarker studies, and we will illustrate shortcomings with some existing methods.

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 may play in their studies. We assess the impact of ignoring center and evaluate existing approaches for accounting for center in biomarker studies. Finally, we propose suitable methods for constructing and evaluating biomarker combinations using data from multiple centers. We restrict attention to biomarkers that will be used to identify individuals likely to have (in the diagnostic setting) or develop (in the 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.

This work was motivated by 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.[5] 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.[5] One goal of the study is to identify a combination of post-operative biomarkers that can provide an earlier diagnosis of AKI.

2 Notation and Terminology

We discuss existing methods for (i) modeling clustered data and (ii) adjusting for covariates in evaluating performance. We then apply these ideas to the multicenter setting, where center can be thought of as both a clustering variable and a covariate. Below, we primarily use the term "cluster," though this is (for our purposes) interchangeable with "covariate."

Let C indicate cluster and suppose the population consists of M clusters where cluster c has N_c observations, $c = 1, ..., M$. Further suppose that we observe data from m of these clusters with n_c observations from cluster c, giving n total observations. We consider a p-dimensional vector of predictors $\mathbf{X}$ and a binary outcome D. Cases (individuals who have or will develop the outcome) are denoted by either $D = 1$ or the subscript D, while controls (individuals who do not have or will not develop the outcome) are denoted by either $D = 0$ or the subscript $\bar{D}$. Let $(\mathbf{X}, D)$ be the predictors and outcome for an arbitrary observation. We use the subscript i on $\mathbf{X}$ and D to denote the predictors and outcome, respectively, for the i^{th} observation. We use the superscript c on $\mathbf{X}$ and D to denote the predictors and outcome, respectively, for an observation from cluster c. We denote the collection of predictors and outcomes for observations in cluster c as $(\mathcal{X}^c, \mathcal{D}^c)$.

In general, in the clustered data setting, predictors may be constant for all observations in a cluster

(often called cluster-level, cluster-constant, or between-cluster predictors), may vary across observations in a cluster (called cluster-varying or within-cluster predictors), or may vary both within and between clusters. We focus on predictors that have at least some variation within clusters. Throughout, we will assume a non-trivial cluster-specific prevalence of D; that is, $P(D = 1|C = c) := \gamma_c \in [1/V, 1 - 1/V]$, $c = 1, ..., M$, for some $V \in (2, \infty)$.

3 Background

3.1 Models for Clustered Data: Random Intercept Logistic Regression

The random intercept logistic regression (RILR) model can be written as:

$$\text{logit}\left\{P(D = 1|\mathbf{X}, C = c, b_c)\right\} = b_c + \tau_0 + \boldsymbol{\tau}^\top \mathbf{X}, \quad b_c \overset{iid}{\sim} F. \tag{1}$$

Typically, it is assumed that $b_c \sim N(0, \sigma^2)$ so σ^2 is an additional parameter in this model. If we view the random intercept b_c as σz_c, where $z_c \overset{iid}{\sim}$, σ is the regression coefficient for this standardized omitted (cluster-level) predictor.[6] In that sense, b_c is generally "interpreted as the combined effects of omitted cluster-level predictors."[7]

3.1.1 Assumptions

The key assumptions typically made by the RILR model given in equation (1) are:[8,9]

(A1) b_c and $\mathcal{X}^c$ are independent

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

(A3) Conditional on $(\mathcal{X}^c, b_c)$, $D_1^c, ...D_{n_c}^c$ are independent and $P(D_i^c = 1|\mathcal{X}^c, b_c) = P(D_i^c = 1|\mathbf{X}_i^c, b_c)$, $i = 1, ..., n_c$

Assumption (A1) can be written as $f(b_c|\mathcal{X}^c) = f(b_c)$, which is a fairly strong assumption in the non-randomized setting.[6] In particular, this assumption is often implausible when the distribution of the predictors varies by cluster.

3.1.2 Estimates

It is important to distinguish between marginal and conditional modeling approaches: the conditional (or cluster-specific) approach, for example, RILR, involves modeling the probability distribution of D as a function of predictors and cluster-specific parameters (e.g., cluster-specific intercepts), while the marginal (or population-averaged) approach involves modeling the marginal expectation of D as a function of predictors.[10] Due to the inclusion of cluster-specific parameters, parameter interpretation under the conditional approach is with respect to cluster.[10] For predictors that vary within clusters, conditional methods are often more appropriate than marginal methods, such as generalized estimating equations.[10]

Predictors frequently have both a between- and within-cluster component; that is, they vary both within and between clusters.[11] Estimates obtained via conditional methods 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 methods with predictors that vary within clusters.[1,12,13] 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

variations.[1,9–11,13–16] Importantly, between-cluster differences are likely to include the effects of cluster-constant confounders.[13]

3.1.3 Violations of Assumptions

First we consider (A1); that is, independence of b_c and $\mathcal{X}^c$. In the context of a randomized multicenter clinical trial, the assumption holds if randomization is stratified by center, since in this situation the distribution of the predictor, treatment, is the same across centers.[1,11] However, as noted above, it is generally the case predictors are not purely within-cluster and have both a between-cluster component and a within-cluster component.[11,17,18] When such predictors are included in a RILR model, the assumption b_c and $\mathcal{X}^c$ are independent may not hold, leading to distortions of the association of interest.[11,17]

As a concrete example, suppose the following model holds for the predictor X:

$$\begin{aligned}\text{logit}\left\{P(D=1|X,C=c,b_c')\right\} &= b_c' + \tau_0 + \tau_B h(\mathcal{X}^c) + \tau_W(X - h(\mathcal{X}^c)) \\ &= b_c' + \tau_0 + (\tau_B - \tau_W)h(\mathcal{X}^c) + \tau_W X, \end{aligned} \tag{2}$$

where $b_c' \sim N(0, \sigma^2)$ and $h(\cdot)$ is some cluster-level summary of $\mathcal{X}^c$ such that $X - h(\mathcal{X}^c)$ has the same distribution across clusters. Here, $X - h(\mathcal{X}^c)$ corresponds to the within-cluster component of X and $h(\mathcal{X}^c)$ corresponds to the between-cluster component of X. If the distribution of the predictor X is the same across clusters, then $(\tau_B - \tau_W)h(\mathcal{X}^c)$ will be constant in large samples, and can be combined with the fixed intercept τ_0. However, if the distribution of the predictor varies across clusters such that $h(\mathcal{X}^c)$ varies and the RILR model given in equation (1) is fit to the data, $b_c = b_c' + (\tau_B - \tau_W)h(\mathcal{X}^c)$ which is not independent of $\mathcal{X}^c$ if $\tau_B \neq \tau_W$, violating assumption (A1).

Results from research on omitted variable bias indicate that when (2) holds, and (1) is fit to the data, the estimate of τ_W will be a combination of the within- and between-effects τ_B and τ_W.[11] Importantly, the combination of within- and between-effects, if these effects differ, is not of substantive interest, lacking clinical relevance.[1,11] 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 problematic.[13,19] If cluster-level factors are associated with predictors, as is often true 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 the predictors.[7,16]

This issue is often called "confounding by cluster" since the within-cluster association, τ_W, is distorted by the between-cluster association, τ_B;[1,9,17,20,21] in the econometrics literature, it is called the "endogenous covariates problem."[7] In our example, omitting $h(\mathcal{X}^c)$ leads to correlation between b_c and $\mathcal{X}^c$, which, as described by Greenland et al., has the effect of confounding τ_W.[22] Thus, confounders are "now covariates that 'explain' the correlation between" b_c and $\mathcal{X}^c$;[22] that is, the cluster-level variable $h(\mathcal{X}^c)$.

Assumption (A2) requires that the random cluster-specific intercepts be independently and identically distributed according to a normal distribution with mean zero and variance σ^2. Broadly speaking, misspecifications of the random intercept distribution may lead to bias in the estimate of the fixed intercept and the coefficients for cluster-level variables but typically do not have a large effect on the estimates for cluster-varying predictors.[6,18,23,24]

3.1.4 Decomposing Predictors

One solution that has been proposed to address violations of assumption (A1) is to decompose predictors into a between-cluster component and a within-cluster component.[6,9,11,14,16,18,19,25,26] In the context of the model at (2), this means fitting a model with $h(\mathcal{X}^c)$ and X as predictors. When $h(\mathcal{X}^c) = \bar{\mathcal{X}}^c$, the cluster mean, this approach is called the "poor man's" method.[11] Using the cluster mean may be overly simplistic[17] and more flexible methods have been proposed based on modeling b_c as a function of $\mathcal{X}^c$.[27] Of course, these methods require that the model for b_c is correctly specified.[16,26,27]

3.1.5 Efficiency

RILR is often touted as being more efficient than alternative methods due in part to the assumption that b_c has some (parametric) distribution.[12,28] In addition, RILR can use both between- and within-cluster comparisons to estimate coefficients, which allows it to use more information in estimating these parameters.[8,9,12,29] Some studies have found reduced efficiency when the distribution of the random intercept is not normal and normality is assumed.[30]

3.2 Models for Clustered Data: Fixed Intercept Logistic Regression

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

$$\text{logit}\left\{P(D=1|\mathbf{X}, C=c, \beta_0^c)\right\} = \beta_0^c + \boldsymbol{\beta}^\top \mathbf{X}, \tag{3}$$

where β_0^c represents a cluster-specific intercept. The conditioning on β_0^c in (3) is only necessary if β_0^c is random. 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.[31]

3.2.1 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.[8] Thus, the key assumption for FILR is:[8,9]

(B1) Conditional on $\mathcal{X}^c$, $D_1^c, ...D_{n_c}^c$ are independent

If the β_0^c are random, then they must be independent across clusters and assumption (B1) must additionally condition on β_0^c.[8,9]

3.2.2 Estimates

FILR consistently estimates the within-cluster effect of predictors that vary within clusters, provided (B1) is satisfied and model (3) holds.[11,12,15,17] Thus, this method avoids the issue of confounding by cluster; in fact, the resulting estimates are not subject to confounding by any unmeasured cluster-constant variable.[13,17]

For both uFILR and cFILR, only within-cluster comparisons are used to estimate the coefficients, and, since clusters for which all observations have $D = 1$ or all observations have $D = 0$ (we call these "concordant clusters") do not contribute any information to the estimation of the within-cluster effect, they are not used in estimation.[8,12] This is also true of clusters that are

concordant on the predictors, though this situation is unlikely when there are multiple and/or continuous predictors.

3.2.3 Efficiency

Many investigators are hesitant to use FILR since the exclusion of concordant clusters could reduce efficiency.[11] However, previous research has shown that cFILR provides estimates that are efficient relative to RILR for predictors that vary predominantly within-clusters.[19,32] Indeed, as pointed out by Neuhaus and Kalbfleisch, for predictors with between- and within-cluster components, the increased efficiency of estimates from RILR that is sometimes observed is often largely due to the assumption of common within- and between-cluster effects.[11] If these effects are indeed equal, there will be some efficiency gain from using RILR since this approach uses both within- and between-cluster variations to estimate the coefficients.[19] However, as noted above, using both types of variation in estimation is generally not recommended since the between and within effects may not be equal and the potential exists for differential confounding. Furthermore, concordant clusters contribute to between-cluster variation and often exhibit strong between-cluster effects, which have the potential to heavily distort the estimated coefficients for predictors that vary within clusters if RILR is used.[14,15]

3.3 Evaluating Performance

Suppose we have a predictor Z and are interested in evaluating its performance. Without loss of generality, we will assume that higher values of Z are more indicative of D. We 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 and controls, and is 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.[33] The ROC curve for a useless predictor is the 45-degree line, and the corresponding AUC is 0.5.[33] 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.[33] The AUC has a probabilistic interpretation: it is the probability that, for a randomly selected case and control, the value of Z for the case is higher than the value of Z for the control.[33]

Covariate effects could influence the evaluation of the predictor Z; in particular, associations between Z and the covariate could allow the covariate to contribute to or attenuate the discriminatory accuracy of Z.[34] In order to prevent the covariate from affecting the assessment of the discriminatory accuracy of Z, the covariate-adjusted AUC should be evaluated. The covariate-adjusted ROC (aROC) and corresponding covariate-adjusted AUC (aAUC) for a discrete covariate C, proposed by Janes and Pepe, can be written as $aROC_Z$ and $aAUC_Z$, respectively, where[35]

$$\begin{aligned} aAUC_Z &= \int_0^1 aROC_Z(t)dt = \int_0^1 \sum_c ROC_{Z|C=c}(t)P(C=c|D=1)dt \\ &= \sum_c w_c AUC_{Z|C=c}, \end{aligned} \tag{4}$$

where t denotes the false positive rate, $ROC_{Z|C=c}$ and $AUC_{Z|C=c}$ denote the covariate-specific ROC and AUC, respectively, and $w_c = P(C=c|D=1)$ denotes the distribution of the covariate among cases. Thus, the aAUC is a weighted average of the covariate-specific AUCs, where the weights correspond to the proportion of cases with each covariate value.[35,36]

When the same data are used to construct a combination and evaluate its performance, the resulting estimate of performance is optimistic.[37] This can be addressed by using a bootstrapping procedure

to estimate the degree of optimism.[37,38] Bootstrapping assumes observations are exchangeable, which may not be reasonable when the data are clustered; thus, bootstrap resampling of clusters has been suggested.[1,36,39,40] However, Bouwmeester et al. found similar results for the average cluster-specific AUC whether resampling was done on clusters or individual observations.[39]

4 Methods

Our predictors consist of a collection of biomarkers, and both the covariate and the cluster variable are center.

When the data come from a single center, common practice is to first construct a 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, we propose using FILR to construct biomarker combinations and the center-adjusted AUC to evaluate them.

Throughout, we focus on constructing a single biomarker combination; 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.[29] Analogously, in the diagnostic and prognostic settings, it is preferable to give a single combination that is not center-specific, as this would make combination development highly localized. We focus on constructing linear combinations via the logistic regression framework. While this may seem restrictive, Pepe et al. noted that the class of linear combinations is actually quite large (taking into

consideration possible biomarker transformations and interactions) and the logistic form is fairly robust.[41]

4.1 The Role of Center

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 practices in each center, could lead to variations in biomarker measurements unrelated to D.

We focus on three possibilities for the role of center. We call center a *confounder* when it affects both the prevalence of D and biomarker measurements, a *case mix variable* when it affects only the prevalence of D, and a *calibration variable* when it affects only the biomarker measurements.

In the TRIBE-AKI study, where the goal is to use biomarkers to diagnose AKI, certain centers may serve particularly unhealthy communities and that this results in differences in biomarker levels across center; however, these differences may 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.

In Figure 1, we present graphical and probabilistic depictions of center as a case mix variable, a calibration variable, and a confounder for diagnostic or prognostic biomarkers $\mathbf{X}$. Diagnostic biomarkers represent some underlying disease or disease process, that is, $D \rightarrow \mathbf{X}$, while prognostic biomarkers that cause some future outcome, that is, $\mathbf{X} \rightarrow D$.

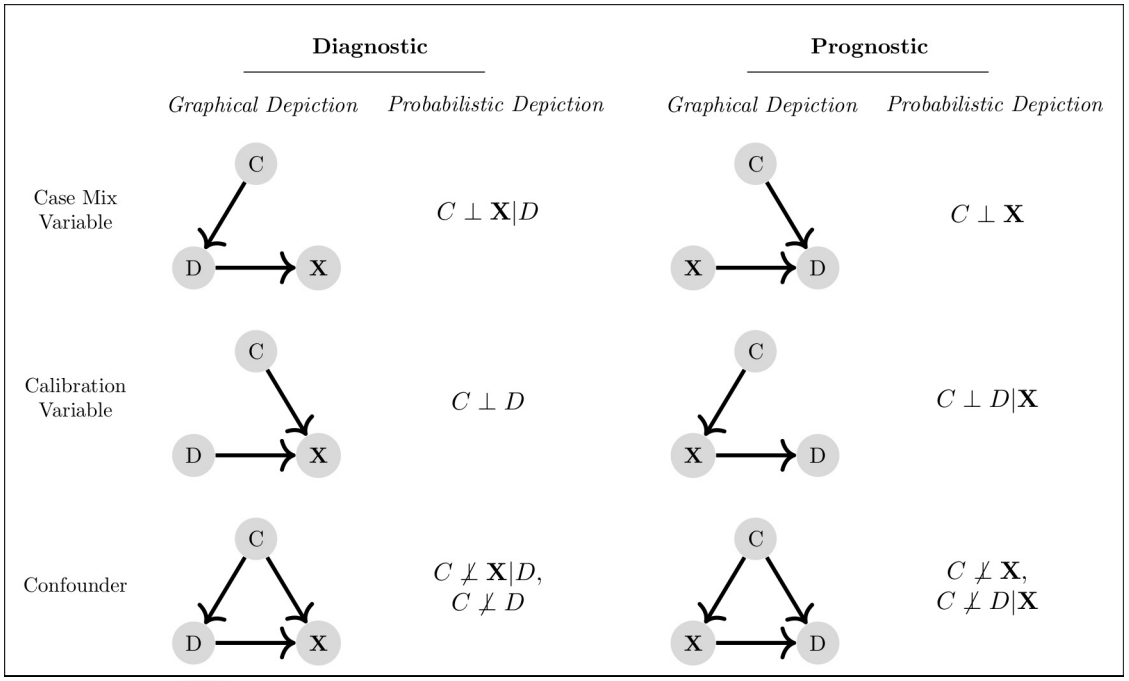

Figure 1: Select potential roles of center in biomarker studies.

It is important to distinguish center as a confounder, as defined in Figure 1, from "confounding by cluster" in the context of a RILR model. Our use of "confounding" in Figure 1 is in line with standard epidemiological notions of confounding, where a variable C distorts the effect of interest, the causal association between $\mathbf{X}$ and D. 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 biomarkers are not independent, leading to distortion of the effect of interest, the within-cluster association. As we will see, there are situations where center is not a confounder by the definitions in Figure 1, but 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."

4.2 Ignoring Center

Clinical researchers frequently ignore center in the construction and/or evaluation of combinations of diagnostic or prognostic biomarkers (e.g., Shapiro et al.[42] and Vuilleumier at al.[43]). This is likely due to the fact that investigators acknowledge that center should not naïvely be included as a predictor, but are not familiar with methods for accommodating center or the repercussions of ignoring it.

4.2.1 Construction

Suppose the linear-logistic model holds:

$$\text{logit}\left\{P(D=1|\mathbf{X}, C=c, \beta_0^c)\right\} = \beta_0^c + \boldsymbol{\beta}^\top \mathbf{X}. \tag{5}$$

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

$$\left(\begin{array}{c} X_1 \\ X_2 \end{array} \middle| D=d, C=c \right) \sim N\left(\left(\begin{array}{c} f_{X_1}(c) + \mu_{X_1} d \\ f_{X_2}(c) + \mu_{X_2} d \end{array} \right), \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right) \right) \tag{6}$$

where μ_{X_1} and μ_{X_2} are related to the center-specific AUC for each marker: $\mu_{X_1} = \sqrt{2}\Phi^{-1}(\lambda_1)$ and $\mu_{X_2} = \sqrt{2}\Phi^{-1}(\lambda_2)$, where Φ is the standard normal distribution function and λ_1 and λ_2 are the center-specific AUCs for X_1 and X_2, respectively. Thus, we consider constant center-specific AUCs for X_1 and X_2, and allowing for center effects on biomarker levels via conditional mean shifts ($f(c)$). Equation (6) gives

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

where β_0^c is a center-specific offset and, as shown in the Supplementary Materials (S1),

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

and

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

Returning to the general linear-logistic model given in (5), suppose that the model holds, but β_0^c is not allowed to vary across centers. That is, suppose we fit the following model to the data pooled across centers:

$$\text{logit}\left\{P(D=1|\mathbf{X})\right\} = \alpha_0 + \boldsymbol{\alpha}^\top \mathbf{X}. \tag{7}$$

When C and D are independent conditional on $\mathbf{X}$ or C and $\mathbf{X}$ are independent conditional on D, and model (5) holds, we have collapsibility,[44] so 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 or not we condition on center. Furthermore, under model (5), when C and D are independent conditional on $\mathbf{X}$, β_0^c will not vary across centers, so $\alpha_0 = \beta_0^c$.

However, when model (5) holds yet C and D are not independent conditional on $\mathbf{X}$ and C and $\mathbf{X}$ are not independent conditional on D, we may no longer have $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Furthermore, the linear-logistic model (5) may not hold, in which case the results on collapsibility 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.

4.2.2 Evaluation

Suppose $p = 2$ and we have a linear combination: $L_{\boldsymbol{\theta}}(\mathbf{X}) = \boldsymbol{\theta}^\top \mathbf{X} = \theta_1 X_1 + \theta_2 X_2$. When center is ignored in the evaluation of $L_{\boldsymbol{\theta}}(\mathbf{X})$, the data are pooled across centers, giving the marginal AUC, $AUC(\boldsymbol{\theta}) = P(L_{\boldsymbol{\theta}}(\mathbf{X}_D) > L_{\boldsymbol{\theta}}(\mathbf{X}_{\bar{D}}))$. In practice, $AUC(\boldsymbol{\theta})$ is estimated empirically:

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

where $1(a > b)$ is 1 if $a > b$ and 0 otherwise. If the $L_{\boldsymbol{\theta}}(\mathbf{X})$ is associated with center, the marginal AUC may not reflect the center-specific AUC.[34]

4.3 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.[29] This idea can be extended to the construction and evaluation of biomarker combinations for diagnosis and prognosis. In particular, we focus on methods that stratify (condition) on center in both the

construction and evaluation of biomarker combinations.

4.3.1 Construction

We will consider two methods for constructing combinations that involve conditioning on center, namely, RILR and FILR; for FILR, we will consider both cFILR and uFILR. For concreteness, we consider $p = 2$ in the discussion below.

4.3.2 Construction: RILR

To the extent that the literature has acknowledged the potential role of center in the prediction setting, RILR is often the approach used in constructing combinations.[45] This model can be written as

$$\begin{aligned} \text{logit}\left\{P(D=1|\mathbf{X}, C=c, b_c)\right\} &= b_c + \tau_0 + \tau_1 X_1 + \tau_2 X_2, \\ b_c &\overset{iid}{\sim} F(0, \sigma^2), \end{aligned} \tag{8}$$

where the distribution of the random center-specific intercepts. The model makes three key assumptions, (A1)–(A3). 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.

The "poor man's" method may be useful in addressing violations of (A1), and can be written

$$\begin{aligned} \text{logit}\left\{P(D=1|X_1, X_2, b_c^*)\right\} &= b_c^* + \tau_0^* + \tau_1^W (X_1 - \bar{X}_1^c) + \tau_2^W (X_2 - \bar{X}_2^c) + \tau_1^B \bar{X}_1^c + \tau_2^B \bar{X}_2^c, \\ b_c^* &\sim F(0, \sigma^{*2}), \end{aligned}$$

where $\bar{X}_1^c$ and $\bar{X}_2^c$ are the means of X_1 and X_2 in center c, respectively, b_c^* represents the random intercept in center c, τ_0^* represents the overall (fixed) intercept, 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.

4.3.3 Construction: FILR

An option that has been discussed at length in the literature on multicenter randomized trials,[1,11,17] but has been largely (if not entirely) neglected in the prediction literature is FILR. We propose using uFILR when the number of centers is modest, and cFILR when the number of centers is large in order to avoid the incidental parameters problem.[31] The FILR model can be written as

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

If the β_0^c are not random, this model relies on assumption (B1).

4.3.4 RILR vs. FILR in Diagnostic and Prognostic Research

Random intercept models are, at first glance, appealing in the context of prediction when the data arise from multiple centers: 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, 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.[8] Thus, while the notion of center-specific intercepts as random

quantities may have intuitive appeal, this is outweighed by the statistical reality that random intercept models rely on potentially untenable assumptions.

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 $P(D = 1|\mathbf{X})$.[21] Even if a valid estimate of b_c is available, the estimate of τ_0 from RILR can be badly biased if the random intercept distribution is misspecified.

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 makes external validation and clinical application (if predicted probabilities are sought) more challenging. In addition, since the "poor man's" method still relies on a RILR model, the estimate of the fixed intercept may face the same challenges as with the standard RILR model.

The goal of the poor man's method is to transform the biomarkers into predictors that are independent of b_c. 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 the within-center effect with no further assumptions or transformations of the data. This is compelling as estimates of biomarker associations (and thus, fitted biomarker combinations) that are unaffected by center differences are most useful in identifying promising combinations for further development.

Conversely, an obvious criticism of FILR is that it does not allow predicted probabilities to be calculated either in new centers (for uFILR) or at all (for cFILR). However, as discussed above, RILR does not necessarily solve this problem. Furthermore, the biomarker combination can still be useful, for example, to stratify patients within each center according to likelihood of having or developing the outcome.

4.3.5 Evaluation: Center-Adjusted ROC and AUC

When the data come from multiple centers, it is important to avoid allowing center to be predictive, so conditional approaches to constructing combinations are appropriate. Likewise, in order to prevent center differences from affecting the assessment of the discriminatory accuracy of a fitted combination, a conditional measure should be used to evaluate performance. In particular, the marginal AUC would be appropriate if between-center heterogeneity were able to be used in making decisions, but this is not typically true.[46] Thus, some summary of the conditional, or center-specific, AUCs should be used to avoid allowing center differences to influence the evaluation of performance. The summary measure defined in equation (4), that is, using the distribution of center among cases to weight the center-specific AUCs, is compelling because it is the area under 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.[35] That is, for a predictor Z, we can write $aROC_Z(t) = P\Big(Z > g_c(t)\Big|D = 1\Big)$, where $g_c(t)$ is the center-specific threshold giving a false positive rate of t in center c.

For a given combination $L_{\boldsymbol{\theta}}(\mathbf{X}) = \boldsymbol{\theta}^\top \mathbf{X}$, the center-adjusted AUC can be written as $aAUC(\boldsymbol{\theta}) = \sum_{c=1}^{M} w_c AUC_c(\boldsymbol{\theta})$ where the center-specific AUC is $AUC_c(\boldsymbol{\theta}) = P(L_{\boldsymbol{\theta}}(\mathbf{X}_D^c) > L_{\boldsymbol{\theta}}(\mathbf{X}_{\bar{D}}^c))$.

In practice, $AUC_c(\boldsymbol{\theta})$ is estimated empirically:

$$\hat{AUC}_c(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{n_D^c} \sum_{j=1}^{n_{\bar{D}}^c} 1(L_{\boldsymbol{\theta}}(\mathbf{X}_{Di}^c) > L_{\boldsymbol{\theta}}(\mathbf{X}_{\bar{D}j}^c)}{n_D^c n_{\bar{D}}^c}.$$

The empirical aAUC estimate is then

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

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, $\hat{w}_c = \frac{n_D^c}{n_D}$. The AUC_c can only be estimated in discordant centers.

When the ROC curve varies by a covariate, it is generally recommended that a separate ROC curve be estimated for each value of the covariate.[36] In the case of center, where only a fraction of the 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.[45]

Finally, as a consequence of focusing on the center-specific AUC, summarized via the aAUC, we do not need an estimate of the center-specific intercept to evaluate a combination, as the center-specific AUC is a rank-based measure and so would be unaffected by such offsets. This allows for identification of promising combinations of biomarkers for further development without the need for center-specific intercept estimates.

4.3.6 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, $a\hat{A}UC(\hat{\boldsymbol{\theta}})$ converges in probability to $aAUC(\boldsymbol{\theta}_0)$ if $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$. This is shown by Lemma 1 and Theorems 1 and 2, which are stated and proved in the Supplementary Material (S2).

4.4 Combining Construction and Evaluation

When constructing and evaluating biomarker combinations, there are two binary decisions to make regarding center, giving four possibilities (using the notation of models (5) and (7)):

1. Pool the data across centers for both construction and evaluation, giving $AUC(\boldsymbol{\alpha})$
2. Pool the data across centers for construction, but stratify by center for evaluation, giving $aAUC(\boldsymbol{\alpha})$
3. Stratify by center for construction, but pool across centers for evaluation, giving $AUC(\boldsymbol{\beta})$
4. Stratify by center for both construction and evaluation, giving $aAUC(\boldsymbol{\beta})$

Proposition 1, given in the Supplementary Material (S3), follows directly from Pepe[33] and shows that the marginal and center-adjusted AUCs of a combination based on some $\boldsymbol{\theta}$ are equivalent if C and $L_{\boldsymbol{\theta}}(\mathbf{X})$ are independent among controls. If C and $\mathbf{X}$ are independent conditional on D, then C and $L_{\boldsymbol{\theta}}(\mathbf{X})$ will be independent among controls. Thus, if model (5) holds and C and $\mathbf{X}$ are independent conditional on D, then $AUC(\boldsymbol{\beta}) = aAUC(\boldsymbol{\beta}) = aAUC(\boldsymbol{\alpha}) = AUC(\boldsymbol{\alpha})$, since $\boldsymbol{\alpha} = \boldsymbol{\beta}$ by collapsibility. Proposition 2, given in the Supplementary Material (S3), also follows directly from Pepe[33] and shows that when the prevalence and center-specific AUC do not vary with center and the center-specific ROC curves are concave, the aAUC for a given biomarker combination will be at least as large as the marginal AUC. In general, the center-specific ROC curves will be concave if for a given $\boldsymbol{\theta}$, in each center, increasing $L_{\boldsymbol{\theta}}(\mathbf{X})$ increases the likelihood that $D = 1$.[37]

When model (5) holds, optimality of the risk score $P(D = 1|\mathbf{X}, C = c, \beta_0^c)$ implies that the combination based on $\boldsymbol{\beta}$ is optimal within each center, in terms of maximizing center-specific AUC.[33,47] Thus, under this model,

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

for any $\boldsymbol{\theta}$. Furthermore, by the collapsibility results discussed above, when model (5) holds and C and D are independent conditional on $\mathbf{X}$, $\boldsymbol{\alpha} = \boldsymbol{\beta}$, so

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

5 Simulations

5.1 Ignoring Center

We studied 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 (as summarized in Figure 1). The two biomarkers X_1 and X_2 were distributed as described in equation (6) with $\rho = 0.5$, and $\lambda_1 = 0.6$ and $\lambda_2 = 0.65$ in all centers.

Throughout, $f_{X_1}(c) = f_{X_2}(c) = f(c)$. 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 $\delta \in \{-0.75, 0.75\}$.

We constructed combinations in a training dataset consisting of either 6 centers with 200 observations each or 500 centers with 20 observations each. The combinations were constructed via logistic regression, where center was either ignored or incorporated using FILR (uFILR for $m = 6$ and cFILR for $m = 500$). These estimates correspond to estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (defined in equations (7) and (5)), respectively.

We evaluated 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. As shown in the Supplementary Material (S4), the conditional AUC is constant across centers under our data-generating model so $AUC_c(\cdot) = aAUC(\cdot)$. 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 true coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $AUC_c(\boldsymbol{\beta})$ were determined analytically for comparison. The simulations were repeated 500 times.

Figure 2 presents the results of the simulations with 500 centers. These simulations support the conclusions given above: that is, when center is a case mix variable, the combination and its performance (in terms of the AUC) are not affected by ignoring center in construction and/or

evaluation. Likewise, the simulation results when center is a calibration variable are consistent with the relationships described 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 actual performance of the combination in a new center.

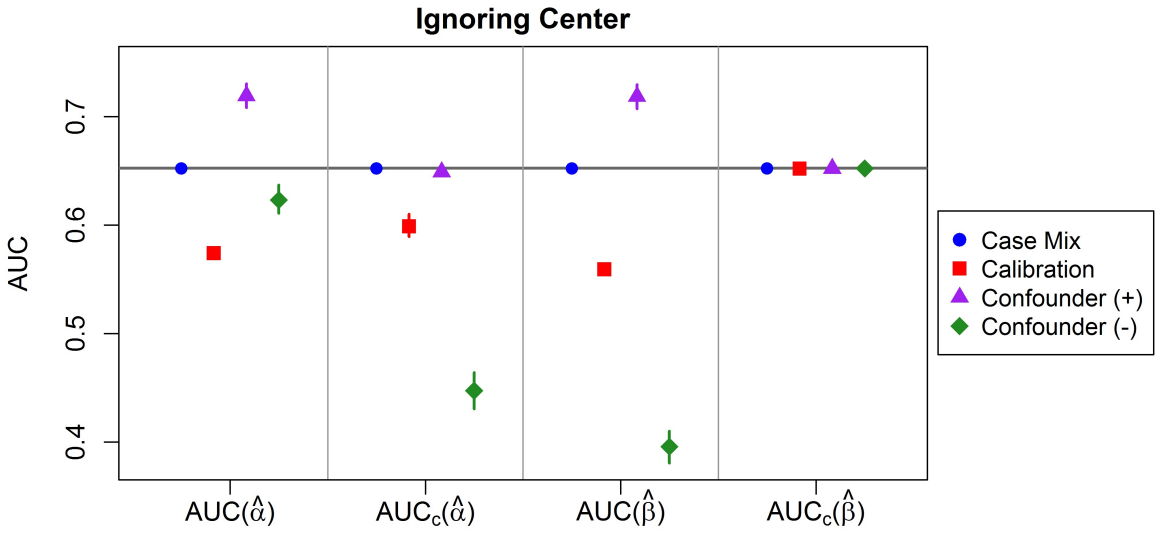

Figure 2: Simulation results for training data with 500 centers. The first column, $AUC(\hat{\boldsymbol{\alpha}})$, is the marginal AUC based on the combination constructed by ignoring center and the second column, $AUC_c(\hat{\boldsymbol{\alpha}})$, is the conditional AUC based on the combination constructed by ignoring center. The third column, $AUC(\hat{\boldsymbol{\beta}})$, is the marginal AUC based on the combination constructed by stratifying by center and the fourth column, $AUC_c(\hat{\boldsymbol{\beta}})$, is the conditional AUC based on the combination constructed by stratifying by center. 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 center is ignored during construction (yielding $\hat{\boldsymbol{\alpha}}$), further ignoring center during evaluation tends to give a measure of performance that is higher than the actual performance in a new center (i.e., $AUC(\hat{\boldsymbol{\alpha}})$ tends to be 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 can give a measure of performance that may be higher or lower than the performance of the combination in a new center; that is, $AUC(\hat{\boldsymbol{\beta}})$ may be larger or smaller than than $AUC_c(\hat{\boldsymbol{\beta}})$, depending on the correlation δ. As expected, ignoring center during construction generally results in a combination with worse performance in new centers ($AUC_c(\hat{\boldsymbol{\alpha}})$ vs. $AUC_c(\hat{\boldsymbol{\beta}})$).

Supplementary Material (S5.1) gives the full results.

5.2 Including Center

We conducted simulations to compare combinations constructed by RILR to those constructed by FILR. The set-up of these simulations is similar to those described in Section 5.1. When center was a case mix variable, $\text{logit}(\gamma_c) \sim F$ 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, $\text{logit}(\gamma_c) \sim F$ with mean 0 and variance $\sigma^2_{\gamma_c}$, $f(c) \sim F$ with mean 0 and variance $\sigma^2_{f(c)}$, and $\text{Corr}(\text{logit}(\gamma_c), f(c)) = \delta$. We varied F (Normal, Gumbel, Laplace, or Uniform), $\sigma^2_{\gamma_c}$ (0.5, 1, 3, (0.5, 1.5), or (1, 5)), $\sigma^2_{f(c)}$ (1, 5, or (2,8)) and δ (-0.5, 0, 0.5). We considered some settings where the variances of $\text{logit}(\gamma_c)$ (or $f(c)$) were not constant (those pairs of values in parentheses); 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. When center was a calibration variable, $\gamma_c = 0.5$ in most simulations; however, to study the impact of concordance, we also considered simulations where $\gamma_c = 0.1$. This was also the motivation for including large values of $\sigma^2_{\gamma_c}$.

Linear combinations were constructed in training data (which had either 6 or 200 centers) via logistic regression, where center was either (i) incorporated using RILR assuming $b_c \overset{iid}{\sim} N(0, \sigma^2)$ 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 RILR and FILR were evaluated in test data, which consisted of a single new center with 10 000 observations. These simulations were repeated 500 times.

In Figure 3, we present the results for $m = 500$ centers in the training data with $F =$ 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. In all scenarios, the results from FILR are close to the true values. The differences in the coefficient estimates when RILR is used are clear, particularly when center is a calibration variable or a confounder. This leads to substantially different conditional AUCs for RILR compared to FILR, 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. Additionally, the differences in the coefficient estimates and the AUC are much larger when center is a calibration variable than when it is a confounder. The full results are given in the Supplementary Materials (S5.2). In general, we see that the differences between RILR and FILR tend to be smaller when there are fewer centers ($m = 6$ vs. $m = 500$), $\sigma^2_{f(c)}$ is small, or $\sigma^2_{\gamma_c}$ is large.

The superior performance of FILR persisted even when we considered situations where there were 500 centers and, on average, 7−12% were concordant (Supplementary Material S5.2). In simulations not designed specifically to have high concordance, up to 2% of centers were concordant, on average.

Finally, we evaluated the estimate of the overall fixed intercept provided by RILR and found absolute biases of more than 20% in many scenarios (Supplementary Material S5.2).

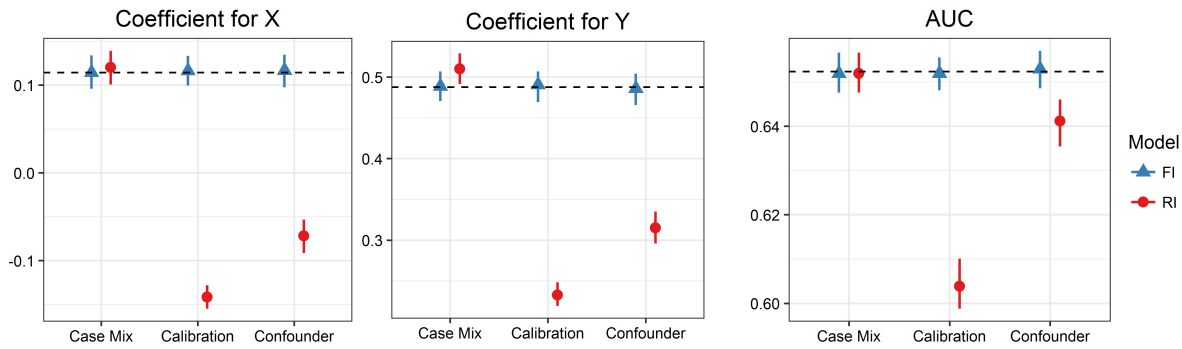

Figure 3: Simulation results comparing random and fixed intercept logistic regression for $m = 500$ in the training data, where F = Normal, $\sigma^2_{\gamma_c} = 1$, $\sigma^2_{f(c)} = 5$, $\gamma_c = 0.5$ when center was a calibration variable, and $\delta = -0.5$ when center was a confounder. The median and interquartile ranges across the 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 AUC in a single new test center are shown in the third plot. In each plot, the dashed horizontal line indicates the true value.

We have provided the `R` functions used to conduct these simulations in the Supplementary Material (S6).

6 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 at six medical centers, and there is interest in using biomarkers to provide an earlier diagnosis of post-operative AKI. All participants provided written informed consent and details regarding subject recruitment and sample collection and storage have been previously reported.[5] These data are used as illustration and not to report new findings of the TRIBE-AKI study. We consider three biomarkers, urine NGAL, h-FABP, and plasma TNI, and use the measurements taken immediately after surgery. After removing observations with missing values

for any of these biomarkers, 962 observations remained. The three biomarkers were log-transformed.

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. There is variation in the distribution of the biomarker measurements across centers among controls (Figure 4). Additionally, the center-specific AKI prevalences were between 7.8% and 22.9%. These results strongly suggest that center is a confounder in this study.

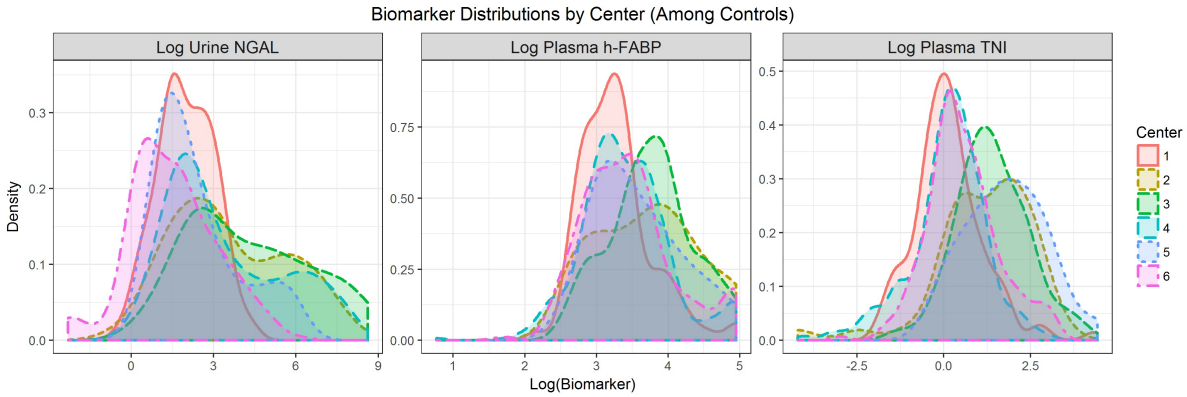

Figure 4: Distribution of log urine NGAL, log plasma h-FABP, and log plasma TNI in the TRIBE-AKI study among controls. The biomarker distributions are stratified by center.

We constructed linear biomarker combinations and evaluated their performance by estimating the center-adjusted AUC. We corrected this estimate for optimism due to resubstitution bias by bootstrapping the individual observations.

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. 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.

7 Discussion

We have created a unified framework for constructing and evaluating biomarker combinations in multicenter studies, including a taxonomy to differentiate the role center can play, tools for identifying the role of center, and methods for constructing a biomarker combination and evaluating its 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 such center differences are often not scientifically relevant and are expected to vary in magnitude from center to center, using conditional approaches for construction and evaluation of biomarker combinations is advised in order to avoid allowing center differences to influence either the combination itself or the assessment of its performance. The concepts and

methods we describe apply to biomarker combinations, and also to combinations of biomarkers and other clinical or demographic variables.

The center-specific AUC may not be the same across centers; in this situation, it is generally informative to evaluate the variability in the center-specific AUCs across center. This offers 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 assessing the center-specific AUCs, it is important to keep in mind that AUC estimates from centers with fewer observations are less reliable.

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[41] and the center-adjusted AUC is unaffected by center-dependent sampling among controls,[35] though the asymptotic results we have provided may not hold under certain sampling schemes. If a multicenter study also involves matching, care must be taken to adjust the AUC for the matching in addition to center.[34]

Future research will consider approaches that do not rely on empirical estimates of the AUC, perhaps by modeling the fitted combination parametrically (e.g., using a model to relate the combination to center among controls);[36] such an approach may be useful 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 biomarker data are frequently not appropriate. Biomarkers hold great potential for use as diagnostic and prognostic tools, but have for the most part been relatively disappointing thus far. 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.[48] 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. Carefully addressing these issues can increase the likelihood of identifying clinically useful combinations, ultimately leading to improvements in patient care.

8 Acknowledgments

The authors wish to acknowledge the TRIBE-AKI study investigators: Steven G. Coca (Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, New York), Amit X. Garg (Institute for Clinical Evaluative Sciences Western, London, Ontario, Canada; Division of Nephrology, Department of Medicine, and Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada), Jay Koyner (Section of Nephrology, Department of Medicine, University of Chicago Pritzker School of Medicine, Chicago, Illinois), and Michael Shlipak (Kidney Health Research Collaborative, San Francisco Veterans Affairs Medical Center, University of California, San Francisco, San Francisco, California).

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

9 Declaration of Conflicting Interests

The authors declare that there is no conflict of interest.

10 Funding

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

References

1. Localio AR, Berlin JA, Ten Have TR et al. Adjustments for center in multicenter studies: an overview. *Ann Intern Med* 2001; 135: 112–123.

2. Feldstein AE, Wieckowska A, Lopez AR et al. Cytokeratin-18 fragment levels as noninvasive biomarkers for nonalcoholic steatohepatitis: a multicenter validation study. *Hepatology* 2009; 50: 1072–1078.

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

4. Nickolas TL, Schmidt-Ott KM, Canetta P et al. Diagnostic and prognostic stratification in the emergency department using urinary biomarkers of nephron damage. *J Am Coll Cardiol* 2012; 59: 246–255.

5. Parikh CR, Coca SG, Thiessen-Philbrook H et al. Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. *J Am Soc Nephrol* 2011; 22: 1748–1757.

6. Heagerty PJ and Zeger SL. Marginalized multilevel models and likelihood inference. *Stat Sci* 2000; 15: 1–26.

7. Skrondal A and Rabe-Hesketh S. Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *J R Stat Soc Ser C Appl Stat* 2014; 63: 211–237.

8. Gardiner JC, Luo Z and Roman L. Fixed effects, random effects and GEE: what are the differences? *Stat Med* 2009; 28: 221–239.

9. Seaman S, Pavlou M and Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Stat Med* 2014; 33: 5371–5387.

10. Neuhaus JM, Kalbfleisch JD and Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev* 1991; 59: 25–35.

11. Neuhaus JM and Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; 54: 638–645.

12. Ten Have TR, Landis JR and Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Stat Med* 1995; 14: 413–429.

13. Gunasekara FI, Richardson K, Carter K et al. Fixed effects analysis of repeated measures data. *Int J Epidemiol* 2014; 43: 264–269.

14. Ten Have TR, Landis JR and Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Stat Med* 1996; 15: 1227–1229.

15. Hu FB, Goldberg J, Hedeker D et al. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *Am J Epidemiol* 1998; 147: 694–703.

16. Neuhaus JM and McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J R Stat Soc Series B Stat Methodol* 2006; 68: 859–872.

17. Berlin JA, Kimmel SE, Ten Have TR et al. An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics* 1999; 55: 470–476.

18. McCulloch CE and Neuhaus JM. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Stat Sci* 2011; 26: 388–402.

19. Schildcrout JS and Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics* 2008; 9: 735–749.

20. Localio AR, Berlin JA and Ten Have TR. Confounding due to cluster in multicenter studies - causes and cures. *Health Serv Outcomes Res Methodol* 2002; 3: 195–210.

21. Pavlou M, Ambler G, Seaman S et al. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol* 2015; 15.

22. Greenland S, Robins JM and Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999; 14: 29–46.

23. Neuhaus JM, Hauck WW and Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* 1992; 79: 755–762.

24. Heagerty PJ and Kurland BF. Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika* 2001; 88: 973–985.

25. Begg MD and Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat Med* 2003; 22: 2591–2602.

26. Brumback BA, Dailey AB and Zheng HW. Adjusting for confounding by neighborhood using a proportional odds model for complex survey data. *Am J Epidemiol* 2012; 175: 1133–1141.

27. Brumback BA, Dailey AB, Brumback LC et al. Adjusting for confounding by cluster using generalized linear mixed models. *Stat Probab Lett* 2010; 80: 1650–1654.

28. Neuhaus JM and Lesperance ML. Estimation efficiency in a binary mixed-effects model setting. *Biometrika* 1996; 83: 441–446.

29. Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome - when, why and how? *BMC Med Res Methodol* 2014; 14.

30. Lukociene O and Vermunt JK. A comparison of multilevel logistic regression models with parametric and nonparametric random intercepts. Technical report, Tilburg University, 2008.

31. Neyman J and Scott EL. Consistent estimates based on partially consistent observations. *Econometrica* 1948; 16: 1–32.

32. Neuhaus JM, Scott AJ, Wild CJ et al. Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics* 2014; 70: 44–52.

33. Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press, 2003. pp. 66–92, 130–166.

34. Janes H and Pepe MS. Adjusting for covariates in studies of diagnostic, screening or prognostic markers: an old concept in a new setting. *Am J Epidemiol* 2008; 168: 89–97.

35. Janes H and Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* 2009; 96: 1–12.

36. Janes H, Longton G and Pepe M. Accommodating covariates in ROC analysis. *Stata J* 2009; 9: 17–39.

37. Copas JB and Corbett P. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 2002; 89: 315–331.

38. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* New York: Springer-Verlag, 2001. pp. 94–95.

39. Bouwmeester W, Moons KGM, Kappen TH et al. Internal validation of risk models in clustered data: a comparison of bootstrap schemes. *Am J Epidemiol* 2013; 177: 1209–1217.

40. Van Oirbeek R and Lesaffre E. An application of Harrell's C-index to PH frailty models. *Stat Med* 2010; 29: 3160–3171.

41. Pepe MS, Cai T and Longton G. Combining predictors for classification using area under the receiver operating characteristic curve. *Biometrics* 2006; 62: 221–229.

42. Shapiro NI, Trzeciak S, Hollander JE et al. A prospective, multicenter derivation of a biomarker panel to assess risk of organ dysfunction, shock, and death in emergency department patients with suspected sepsis. *Crit Care Med* 2009; : 96–104.

43. Vuilleumier N, Le Gal G, Verschuren F et al. Cardiac biomarkers for risk stratification in non-massive pulmonary embolism: a multicenter prospective study. *J Thromb Haemost* 2008; 7: 391–398.

44. Guo J and Geng Z. Collapsibility of logistic regression coefficients. *J R Stat Soc Series B Stat Methodol* 1995; 57: 263–267.

45. Bouwmeester W, Twisk JWR, Kappen TH et al. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol* 2013; 13.

46. Van Klaveren D, Steyerberg EW and Vergouwe Y. Interpretation of concordance measures for clustered data. *Stat Med* 2014; 33: 714–716.

47. McIntosh MW and Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics* 2002; 58: 657–664.

48. Ioannidis JPA. Biomarker failures. *Clin Chem* 2013; 59: 202–204.