Risk Assessment with Imprecise EHR Data

Risk Assessment with Imprecise EHR Data

A dissertation presented

by

Stephanie F. Chan

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

June 2018

Dissertation Advisor: Professor Tianxi Cai Stephanie F. Chan

Risk Assessment with Imprecise EHR Data

Abstract

Electronic health records (EHRs) are electronic versions of patient charts, created to improve patient care. The adoption of EHRs in the US has increased significantly in the last decade, making it a rich resource for conducting clinical research. The breadth of the EHRs, with detailed longitudinal patient data and information on a wide range of disease conditions, allows for new opportunities for different types of clinical research.

The detailed phenotypic information on individual patients allows for simultaneously studying multiple phenotypes. A useful tool for such simultaneous assessment is the Phenome-wide association study (PheWAS), which relates a genomic or biological marker of interest to a wide spectrum of disease phenotypes, typically defined by the diagnostic billing codes. One challenge arises when the biomarker of interest is expensive to measure on the entire EMR cohort. Performing PheWAS based on supervised estimation using only subjects who have marker measurements may yield limited power. In chaper 1, we focus on the setting in a PheWAS where the marker is measured on a small fraction of the patients while a few surrogate markers such as historical measurements of the biomarker are available on a large number of patients. We propose an efficient semi-supervised estimation procedure to estimate the covariance between the biomarker and the billing code, leveraging the surrogate marker information. We employ surrogate marker values to impute the missing outcome via a two-step semi-non-parametric approach and demonstrate that our proposed estimator is always more efficient than the supervised counterpart without requiring the imputation model to be correct. We illustrate the proposed procedure by assessing the association between the C-reactive protein (CRP) and some inflammatory diseases with an EMR study of inflammatory bowel disease performed with the Partners HealthCare EMR where CRP was only measured for a small fraction of the patients due to budget constraints.

In chapters 2 and 3, we focus on the challenges in using EHRs to build risk prediction models. One major challenge is that the timing of disease onset is not readily available. Extracting clinical event times for patients requires labor intensive medical chart reviews. Additionally, since a significant proportion of clinical events may occur prior to patients' first EHR encounter or outside of the specific hospital system, the EHR may only capture partial information on the event time. For example, the domain expert would be able to determine whether a patient has experienced a clinical outcome by the end of EHR follow-up, but the exact timing may be unknown even after chart review. The time to first ICD9 billing code for the clinical condition or the first NLP mention of the condition in the notes can serve as a proxy for the true event time, but is subject to measurement error. In chapter 2, we propose a robust approach to developing a risk prediction model by synthesizing multiple imperfect sources of information on the event time of interest. Treating the partially observed outcomes as survival time subject to current status censoring and survival time measured with errors, we construct an optimally combined estimator under a flexible semi-parametric transformation model for the survival time given baseline predictors and unspecified measurement errors. Simulation studies demonstrate that the proposed estimator performs well in finite sample. We illustrate the proposed estimator by assessing the effects of genetic markers on coronary artery disease with an EHR study of rheumatoid arthritis patients performed with the Partners HealthCare EMR. In chapter 3, we propose a maximum likelihood estimator to estimate the risk of developing a disease by combining only the multiple imperfect sources of information on the event time of interest. Simulation studies demonstrate that the proposed estimator performs well in finite sample. We illustrate the proposed estimator by predicting the risk of developing type 2 diabetes based on a obesity genetic risk score in a cohort of patients from the Partners Biobank.

Contents

Acknowledgments

I would like to thank my advisor, Professor Tianxi Cai, for the endless patience and understanding she has shown me over the last 5 years. I would also like to thank my committee members, Professor Jun Liu and Professor Chirag Patel, for generously sharing their time and advice with me throughout the years. Finally, I would like to thank my family for their support and encouragement. I could not have achieved this without you.

Semi-Supervised Estimation of Covariance with Application to Phenome-wide Association Studies with EHR Data

Stephanie F. Chan

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Boris P. Hejblum

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Abhishek Chakrabortty

Department of Statistics

The Wharton School

University of Pennsylvania

Tianxi Cai

Department of Biostatistics

Harvard T.H. Chan School of Public Health

1.1 Introduction

Electronic medical records (EMRs) are a database of clinical data from a particular medical provider. They contain a range of information on patients, including demographics, medical history, test results, and billing information. There have been high hopes that this data-rich resource can be widely used to perform observational clinical association studies. One popular tool for performing discovery research with EMR is the phenome-wide association study (PheWAS) (Denny et al., 2010) where one examines the association between a genomic or biological marker and a wide range of disease phenotypes, typically defined by the International Classification of Diseases, Ninth Revision (ICD9) billing codes. This method has been used in several exploratory studies, for example to detect association between autoantibody positivity and ICD9 codes related to hypertension (Liao et al., 2010, 2013).

When the biomarker of interest is too expensive to be measured on all subjects in the EMR cohort, performing PheWAS may be challenging. For example, in an EMR study on how the co-morbidities of inflammatory bowl disease relate to inflammation conducted at Partner's Healthcare, the inflammatory marker, C-reactive Protein (CRP) was only measured on a small, randomly selected subset of the study participants. Performing PheWAS only on those with CRP measurements would have limited power. In this paper, we propose semi-supervised PheWAS methods that enable us to increase the power for such settings by leveraging additional information on surrogate markers such as historical measurements of inflammation markers. We are interested in the semi-supervised setting since the percent of missingness in the CRP measurement is approaching 100%. As such, traditional missing data approaches such as multiple imputation and inverse probability weighting do not directly apply here (Rubin, 1987; Seaman and White, 2013). Multiple imputation relies on creating a distribution for the missing outcome data and making M repeated draws from this distribution to create M complete datasets. The M estimators for each dataset are averaged together to obtain a final estimator; however, in cases where the percent of missingness is high, the required minimum M needed to accurate inference will be rather large (Kenward and Carpenter, 2007). This makes multi-

ple imputation a computationally difficult approach for our setting. Furthermore, simple imputation methods may not be effective when the imputation model is mis-specified. In this paper, we propose a semi-supervised estimator of the covariance between CRP and the ICD9 billing codes via a two-step semi-non-parametric imputation, which is robust to model mis-specification.

Semi-supervised methods have been applied to EMR data in the past (Rosales et al., 2007; Kim and Shin, 2013); however, most of these methods also focus on classification of disease status, rather than on estimation or testing (Dligach et al., 2015; Wang et al., 2012). There are no current semi-supervised methods for estimating covariance, which we can use to test for a potential association between the outcome variable and a particular disease, but recently, there has been some literature on semi-supervised estimation of the mean, which could be potentially be used in the calculation of the covariance. For example, Sokolovska *et al.* Sokolovska et al. (2008) proposed a method for estimating the conditional density for classification using a weighted likelihood estimator based on the ratio of the densities of the covariates from labeled and unlabeled data. Kawakita and Kanamori Kawakita and Kanamori (2013) extend Sokolovska *et al.*'s Sokolovska et al. (2008) method to allow for estimating the conditional mean using an estimate of the density ratio. Unfortunately, these methods require specification of the basis functions used in the density ratio model and the choice of the basis functions remain unclear. Additionally, it is unclear how to extend their methods for the estimation of the covariance which involves both first and second moment estimations. Our two-step approach uses surrogate variables to aid in the imputation of the missing outcome values. We start with a linear regression to impute the missing biomarker levels using the ICD9 codes and the surrogate variables as predictors. In the second step, we use these imputed values to calculate the individual contribution to the covariance, and then employ a calibration step via kernel smoothing to increase robustness to the misspecification in the imputation model. The remainder of the paper is organized as follows. In Section 1.2, we formulate a semi-supervised estimator for this covariance and devise a method to calculate its standard error. In Section 1.3, we perform a simulation study to explore our methods and show the results of the simulations, and in section 1.4, we apply our method to an

example dataset.

1.2 Methods

In this section, we detail our proposed semi-supervised estimator for the covariance between a biological marker of interest, denoted by Y, and a phenotype of interest, denoted by G. In EMR settings, examples for Y include inflammation markers such as CRP or autoantibodies such as anti-cyclic citrullinated peptide; while G could be the total count of ICD9 codes for a specific disease condition. Due to cost limits, Y is only measured for n patients randomly selected from an EMR cohort of size N, where G is available for all patients, where we assume that $n \ll N$ in that $\lim_{n\to\infty} n/N = 0$ as in a standard semi-supervised setting. In addition, there are often auxiliary variables, denoted by $\mathbf{S}$, potentially predictive of Y stored in the EMR for all patients, that we can use as surrogate variables for Y. For example, if Y is current CRP level, $\mathbf{S}$ could be past history of inflammation markers including CRP and erythrocyte sedimentation rate (ESR). We do not require past history to be available on all subjects or assumptions on how $\mathbf{S}$ relation to Y. For example, we may encode availability of the past measurements as one of the surrogate variables since the availability of such measurements may be predictive of Y. Suppose that the underlying full data data consists of N independent and identically distributed (iid) random vectors $\mathscr{F} = \{(Y_i, G_i, \mathbf{S}_i^\intercal)^\intercal, i = 1, ..., N\}$, while the observable data is $\mathscr{D} = \mathscr{L} \cup \mathscr{U}$ with

$$\mathscr{L} = \{(Y_i, G_i, \mathbf{S}_i^\intercal)^\intercal, i = 1, ..., n\}, \quad \text{and} \quad \mathscr{U} = \{(G_i, \mathbf{S}_i^\intercal)^\intercal, i = n+1, ..., N\}$$

as the labeled and unlabeled data, respectively. We assume that Y is missing completely at random as typically assumed in the semi-supervised setting.

1.2.1 Estimation

Our goal is to leverage all available data in $\mathscr{D}$ and provide a semi-supervised estimation of

$$\theta_0 = \text{cov}(Y_i, G_i) = E(r_i), \quad \text{where } r_i = (Y_i - \mu_y)(G_i - \mu_g), \mu_y = E(Y_i) \text{ and } \mu_g = E(G_i).$$

The standard supervised estimator is:

$$\widehat{\theta}_{\text{SL}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_{y,\text{SL}})(G_i - \widehat{\mu}_{g,\text{SSL}}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{r}_i$$

where $\widehat{\mu}_{y,\text{SL}} = n^{-1}\sum_{i=1}^{n} Y_i$, $\widehat{r}_i = (Y_i - \widehat{\mu}_{y,\text{SL}})(G_i - \widehat{\mu}_{g,\text{SSL}})$, and $\widehat{\mu}_{g,\text{SSL}} = N^{-1}\sum_{i=1}^{N} G_i$. It is well known that $\widehat{\theta}_{\text{SL}}$ is a consistent estimator of θ_0 and $n^{\frac{1}{2}}(\widehat{\theta}_{\text{SL}} - \theta_0)$ converges in distribution to a normal with mean 0 and variance $\sigma^2_{\text{SL}} = E\{(r_i - \theta_0)^2\}$.

To derive a semi-supervised estimator leveraging $\mathcal{U}$, we propose a two-step procedure. In step I, we fit a working linear model

$$E(Y_i - \mu_y \mid \mathbf{S}_i, G_i) = \boldsymbol{\beta}^{\intercal}\mathbf{W}_i, \tag{1.1}$$

where $\mathbf{W}_i$ is some basis expansions of $\mathbf{S}_i$ and G_i that include both 1 and G_i. For example, $\mathbf{W}_i$ may include 1, $\mathbf{S}_i$, G_i, as well as the interaction between $\mathbf{S}_i$ and G_i. Let

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n}\mathbf{W}_i\mathbf{W}_i^{\intercal}\right)^{-1}\sum_{i=1}^{n}\mathbf{W}_i(Y_i - \widehat{\mu}_{y,\text{SL}})$$

be the ordinary least square estimator of $\boldsymbol{\beta}$. Regardless of the adequacy of the linear model (1.1), $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\bar{\boldsymbol{\beta}}$, the solution to $E\{\mathbf{W}_i(Y_i - \mu_y - \boldsymbol{\beta}^{\intercal}\mathbf{W}_i)\} = 0$. Based on this model, we predict the unobserved r_i as

$$\widehat{R}_i = \widehat{\boldsymbol{\beta}}^{\intercal}\widehat{\mathbf{X}}_i, \quad \text{where} \quad \widehat{\mathbf{X}}_i = \mathbf{W}_i(G_i - \widehat{\mu}_{g,\text{SSL}}), \quad \text{where} \quad \widehat{\mu}_{g,\text{SSL}} = N^{-1}\sum_{i=1}^{N} G_i.$$

If the linear model (1.1) is correctly specified, $\widehat{R}_i$ is a consistent estimator of $E(r_i \mid \mathbf{W}_i)$ and hence

$$\widehat{\theta}^{\text{par}}_{\text{SSL}} = N^{-1}\sum_{i=1}^{N}\widehat{R}_i$$

consistently estimates θ_0. When (1.1) is potentially mis-specified, we show in the appendix that $\max_i |\widehat{R}_i - R_i| \to 0$ in probability, and therefore $\widehat{\theta}^{\text{par}}_{\text{SSL}}$ remains a consistent estimator of θ_0 provided that $\mathbf{W}_i$ includes 1 and G_i, where $R_i = \bar{\boldsymbol{\beta}}^{\intercal}\mathbf{W}_i(G_i - \mu_g)$. In addition, $n^{\frac{1}{2}}(\widehat{\theta}^{\text{par}}_{\text{SSL}} - \theta_0)$ converges in distribution to a normal random variable with mean 0 and variance $(\sigma^{\text{par}}_{\text{SSL}})^2 = E\{(r_i - R_i)^2\}$.

Despite its robustness, $\widehat{\theta}^{\text{par}}_{\text{SSL}}$ may not be very efficient under model mis-specification. To further improve efficiency, in step II, we propose to calibrate the conditional mean

$E(r_i \mid R_i)$ via a one-dimensional smoothing and use the calibrated estimate to construct our semi-supervised estimator. Specifically, our calibrated semi-supervised estimator of θ_0 is

$$\widehat{\theta}_{\text{SSL}} = N^{-1} \sum_{i=1}^{N} \widehat{m}(\widehat{\boldsymbol{\beta}}^{\intercal} \widehat{\mathbf{X}}_i, \widehat{\boldsymbol{\beta}}) = \int \widehat{m}(x, \widehat{\boldsymbol{\beta}}) d\widehat{\mathcal{P}}(x, \widehat{\boldsymbol{\beta}}),$$

where $\widehat{\mathcal{P}}(x, \boldsymbol{\beta}) = N^{-1} \sum_{i=1}^{N} I(\boldsymbol{\beta}^{\intercal} \widehat{\mathbf{X}}_i \leq x)$,

$$\widehat{m}(\mathbf{x}, \boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\intercal} \widehat{\mathbf{X}}_i - x) \widehat{r}_i}{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\intercal} \widehat{\mathbf{X}}_i - x)},$$

$K_h(x) = h^{-1} K(x/h)$, $K(\cdot)$ is a smooth kernel density function, $h = O(n^{-\nu})$ is the bandwidth with $\nu \in (1/4, 1/2)$. Since kernel smoothing introduces some bias to the estimate in finite samples, we add an additional bias correction term to $\widehat{\theta}_{\text{SSL}}$ and propose our final bias corrected semi-supervised estimator as

$$\widehat{\theta}^{\text{BC}}_{\text{SSL}} = \widehat{\theta}_{\text{SSL}} - \left\{ n^{-1} \sum_{i=1}^{n} \widehat{m}(\widehat{\boldsymbol{\beta}}^{\intercal} \widehat{\mathbf{X}}_i, \widehat{\boldsymbol{\beta}}) - \widehat{\theta}_{\text{SL}} \right\}.$$

To improve smoothing performance, we may also consider transformed scores. For example, we may find its percentile using the unlabeled data and smooth over the percentiles. For ease of presentation, we omit the transformation.

1.2.2 Inference

We show in the appendix that $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$ is consistent and $n^{\frac{1}{2}}(\widehat{\theta}^{\text{BC}}_{\text{SSL}} - \theta_0)$ is asymptotically normal with mean 0 and variance

$$\sigma^2_{\text{SSL}} = E[\{r_i - E(r_i \mid R_i)\}^2] = E\{\text{var}(r_i \mid R_i)\}.$$

It is straightforward to see that $\sigma^2_{\text{SSL}} < \sigma^2_{\text{SL}}$ provided that $\mathbf{W}_i$ is predictive of Y_i. Comparing to the model based estimator $\widehat{\theta}^{\text{par}}_{\text{SSL}}$, we note that when the parametric model of $E(Y_i - \mu_g \mid \mathbf{S}_i, G_i) = \boldsymbol{\beta}^{\intercal} \mathbf{W}_i$ holds, $R_i = E(r_i \mid R_i)$ and hence the $\widehat{\theta}^{\text{par}}_{\text{SSL}}$ is asymptotically equivalent to the calibrated estimator $\widehat{\theta}_{\text{SSL}}$. Under model mis-specification, we may have $P\{E(r_i \mid R_i) \neq R_i\} > 0$ in which case $(\sigma^{\text{par}}_{\text{SSL}})^2 = E\{(r_i - R_i)^2\} > \sigma^2_{\text{SSL}}$.

To estimate the variance for $\widehat{\theta}_{\text{SSL}}$, we may estimate σ^2_{SSL} empirically as $n^{-1} \sum_{i=1}^{n} \{\widehat{r}_i - \widehat{m}(\widehat{\boldsymbol{\beta}}^{\intercal} \widehat{\mathbf{X}}_i, \widehat{\boldsymbol{\beta}})\}^2$.

1.3 Simulation results

We conducted a simulation study to assess the finite sample performance of our semi-supervised estimation procedures and also compare the semi-supervised estimators to $\widehat{\theta}_{\text{SL}}$. Throughout, G_i was generated from the log of 1 plus a negative binomial$(3, 0.9)$ to mimic the number of ICD9 codes. We then generate $(V_i, \mathbf{U}_i^{\intercal})_{4\times 1}^{\intercal}$ from a multivariate normal distribution with mean $\beta G_i \mathbf{1}_{4\times 1}$ and covariance matrix $0.7+0.3\mathbf{I}_{4\times 4}$, where β is chosen to be 0 leading to $\theta_0 = 0$ and 0.3 to reflect a modest association. We consider two scenarios for generating $\mathbf{S}_i$ and Y_i:

$$\begin{aligned} \mathscr{M}_{\text{lin}}: &\quad Y_i = V_i, \quad \mathbf{S}_i = \mathbf{U}_i, \\ \mathscr{M}_{\text{nlin}}: &\quad Y_i = V_i + \beta G_i^2 - \beta G_i, \quad \mathbf{S}_i = \mathbf{U}_i - \beta G_i^2 \end{aligned}$$

For both settings, we let $\mathbf{W}_i = (1, G_i, \mathbf{S}_i^{\intercal}, \mathbf{S}_i G_i)^{\intercal}$ when fitting the imputation model. We let $N = 60000$ and consider labeled data sizes of $n = 200, 400,$ and 600. The bandwidth h was chosen as $\widehat{\tau} \times n^{-0.3}$, where $\widehat{\tau}$ is the empirical standard deviation of $\widetilde{\pi}_i$, the percentile of scores. For each configuration, we summarize results using 1000 datasets.

In Table 1.1, we summarize results for $\widehat{\theta}_{\text{SL}}$, $\widehat{\theta}_{\text{SSL}}^{\text{par}}$ and $\widehat{\theta}_{\text{SSL}}^{\text{BC}}$ along with their bias, mean squared error (MSE), and relative efficiency (RE) of the semi-supervised estimators compared to the supervised estimator. All estimators have negligible biases regardless of the adequacy of the fitted parametric model although the bias of the parametric imputation based semi-supervised estimator $\widehat{\theta}_{\text{SSL}}^{\text{par}}$ has slightly larger biases. Consistent with our theoretical results, the semi-supervised estimators $\widehat{\theta}_{\text{SSL}}^{\text{par}}$ and $\widehat{\theta}_{\text{SSL}}^{\text{BC}}$ are substantially more efficient than the supervised estimator $\widehat{\theta}_{\text{SL}}$, with relative efficiency ranging from about 2.1 to 5.2. Under $\mathscr{M}_{\text{lin}}$, $\widehat{\theta}_{\text{SSL}}^{\text{par}}$ are $\widehat{\theta}_{\text{SSL}}^{\text{BC}}$ have near identical MSEs, which is expected since they are asymptotically equivalent. Under $\mathscr{M}_{\text{nlin}}$, the fitted linear model is mis-specified and hence we would expect $\widehat{\theta}_{\text{SSL}}^{\text{BC}}$ to be more efficient than $\widehat{\theta}_{\text{SSL}}^{\text{par}}$. This is indeed reflected in the simulation results - the efficiency of $\widehat{\theta}_{\text{SSL}}^{\text{BC}}$ relative to $\widehat{\theta}_{\text{SSL}}^{\text{par}}$ is around 1.5. We also investigated the performance of our interval estimation based on the asymptotic variance. We calculated the coverage of θ_0 from the estimated 95% CIs. As shown in Figure 1.1, the empirical coverage probabilities are close to their nominal level. We note that the parametric imputation

is somewhat unstable under model mis-specifications in small samples, resulting CIs that slightly under cover when $n = 200$.

Table 1.1: Bias ($\times 100$), MSE ($\times 100$), and relative efficiency (RE) of of the semi-supervised estimators compared to the supervised estimator for $\widehat{\theta}_{\text{SL}}$, $\widehat{\theta}^{\text{par}}_{\text{SSL}}$ and $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$.

		$\mathscr{M}_{\text{lin}}$: $\theta_0 = 0$			$\mathscr{M}_{\text{lin}}$: $\theta_0 = 0.188$			$\mathscr{M}_{\text{nlin}}$: $\theta_0 = 0.907$		
	n	200	400	600	200	400	600	200	400	600
$\widehat{\theta}_{SL}$	Bias	-0.075	-0.207	-0.254	-0.156	-0.315	-0.321	-0.372	-0.632	-0.562
	MSE	0.299	0.157	0.102	0.348	0.180	0.118	1.032	0.540	0.356
$\widehat{\theta}^{par}_{SSL}$	Bias	0.020	-0.052	-0.076	0.029	-0.054	-0.074	1.198	0.608	0.330
	MSE	0.128	0.065	0.043	0.128	0.065	0.043	0.327	0.154	0.104
	RE	2.345	2.423	2.347	2.713	2.774	2.714	3.154	3.497	3.423
$\widehat{\theta}^{BC}_{SSL}$	Bias	0.012	-0.034	-0.067	-0.094	-0.097	-0.099	-0.482	-0.350	-0.266
	MSE	0.140	0.072	0.047	0.149	0.076	0.049	0.223	0.104	0.069
	RE	2.140	2.167	2.167	2.343	2.374	2.392	4.623	5.187	5.147

1.4 Application to an EMR Study of Inflammation for Inflammatory Bowel Disease

We applied the proposed method to investigate potential associations between an inflammatory marker and co-morbidities among patients suffering from Inflammatory Bowel Disease (IBD). The two main types of IBD are Crohn's disease, which causes inflammation in the digestive tract, and ulcerative colitis, which causes inflammation and ulcers in the colon and rectum (Tu et al., 2015). In response to inflammation in the body, the liver releases C-reactive protein (CRP) into the bloodstream, so higher CRP levels are an indication of inflammation in the body (Gabay and Kushner, 1999). The goal of our analysis is to examine whether inflammation (quantified by CRP levels) is related to comorbidities for IBD patients using an EMR crimson cohort of 2,048 patients from Partner's Healthcare Systems. The IBD EMR cohort originally consists of 11,001 patients who were identified as having IBD via a phenotyping algorithm as described in Ananthakrishnan et al. (2013). Out of the 11,001 patients, 2,048 contributed blood for research and we only consider the crimson cohort as the full cohort due to the discrepancy between patients who contributed blood versus those who did not.

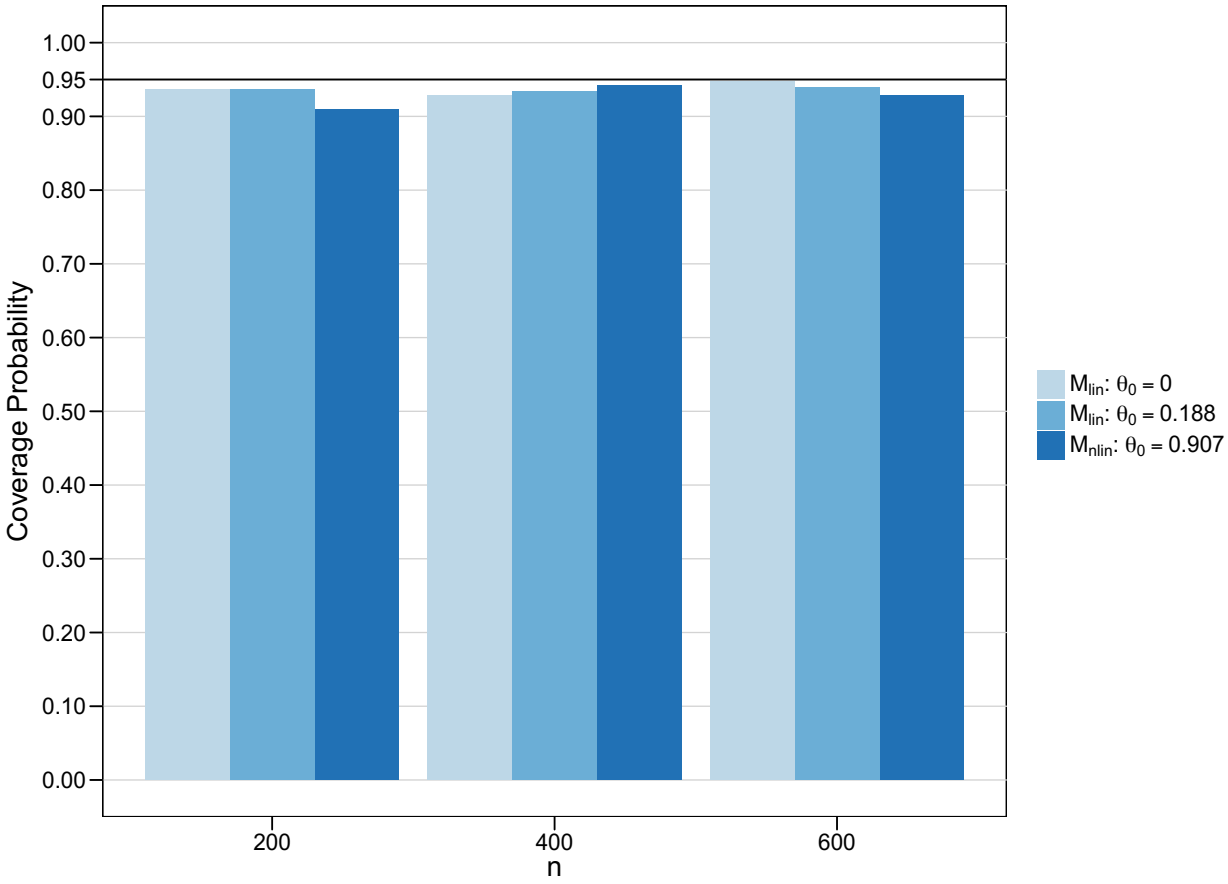


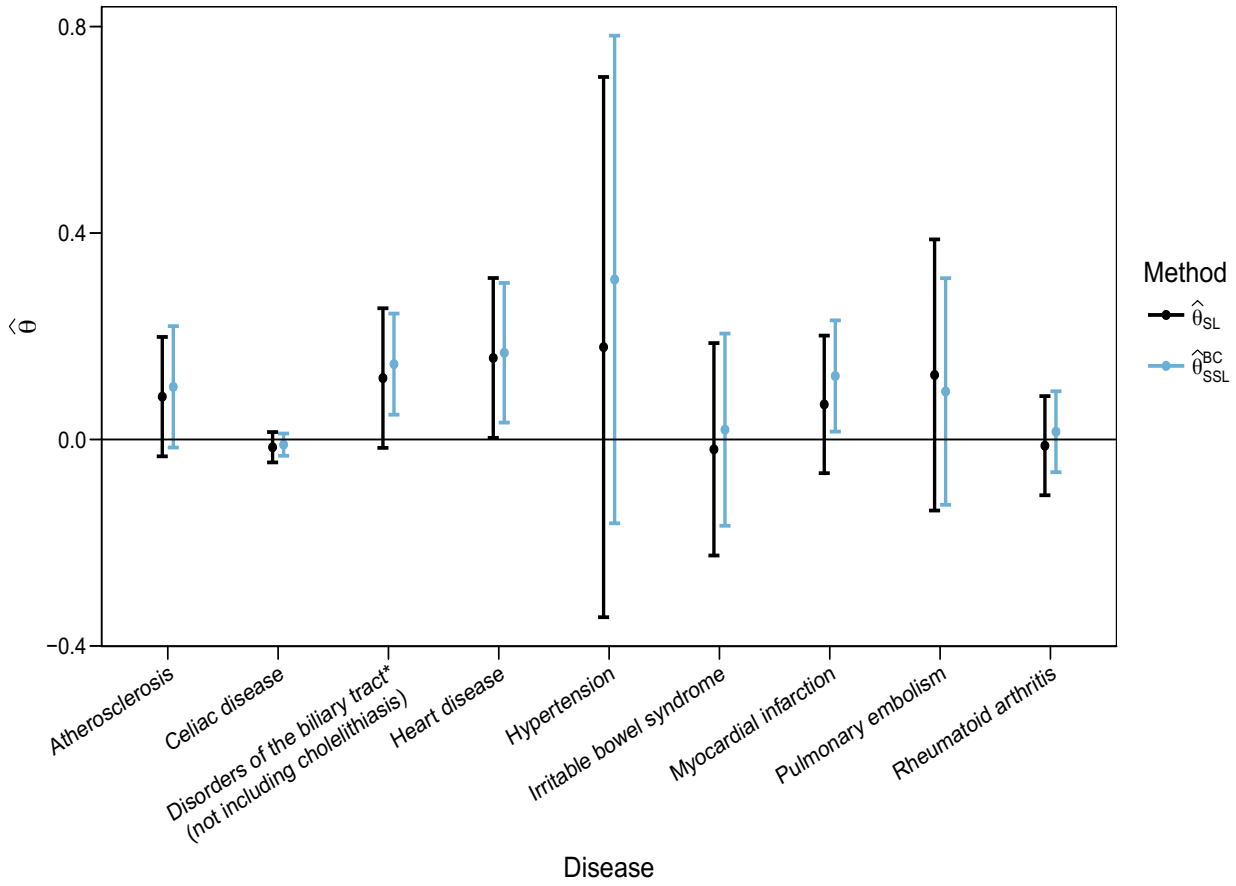
Figure 1.1: Coverage probabilities of the 95% CIs for $\widehat{\theta}^{BC}_{SSL}$ under various simulation settings

To quantify the current level of inflammation, 97 patients were randomly selected from the IBD crimson cohort to have their CRP measured. The co-morbidities are quantified by the number of PheWAS codes associated with each disease condition of interest, which is available for all subjects. In addition, 1,686 patients have previously measured CRP and/or ESR levels recorded, which we use to construct $\mathbf{S}$. Note that in addition to the previous levels of CRP and ESR, the fact that no such measurements exist for certain patients is potentially predictive of the current CRP level. We thus create $\mathbf{S}$ to include the average levels of CRP and ESR for those who have such information, the missing indicators, as well as other predictors including age, gender and race. For our analysis, we let Y be the current log CRP level and G be the $x \rightarrow \log(x+1)$ transformed PheWAS code for each disease of interest. We considered several disease conditions that are previously reported as being associated with inflammation or being a comorbiditiy of IBD including atherosclerosis, celiac disease, disorders of the biliary tract (not including cholelithiasis)[1], heart disease, hypertention, irritable bowel syndrome, mycardial infarction, pulmonary embolism and rheumatoid arthritis. The point estimators and 95% CIs for $\widehat{\theta}_{\text{SL}}$ and $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$ are shown in Figure 1.2. The results suggest that the supervised and semi-supervised estimates are reasonably consistent with each other in value, while the 95% CIs for the semi-supervised method is always smaller than the supervised method, as we expect. For example, for heart disease, the covariance is estimated as 0.158 with 95% CI [0.003,0.313] based on $\widehat{\theta}_{\text{SL}}$; as 0.168 with 95% CI [0.033, 0.303] based on $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$. In the cases of myocardial infarction and disorders of the biliary tract, a Z-test based on $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$ would reject the null hypothesis, whereas a Z-test based on $\widehat{\theta}_{\text{SL}}$ would not.

1.5 Discussion

Our semi-supervised estimate of the covariance is able to improve the supervised estimator by incorporating information from the large number of unlabeled patients with available ICD9 codes as well as surrogate variables including past measurements of biomarkers. Simulation results show that our proposed estimator is consistent and more efficient

[1]This corresponds to PheWAS code 576, as described in Denny *et al.*Denny et al. (2010)



*This corresponds to PheWAS code 576, as described in Denny *et al.* Denny et al. (2010)

Figure 1.2: Supervised and Semi-supervised estimates of the covariance for select PheWAS codes, along with the 95% CIs

than the supervised estimate, which is confirmed by the results from an EMR study. Additionally, the results indicate that our estimator is consistent regardless of the adequacy of the working model.

Our proposed covariance estimator, along with its standard error estimate, can be used to perform tests of association between the ICD9 codes and outcome of interest, for example, a Z-test. The gain in efficiency of our method over the supervised method would increase the power of association tests. Further increases in power to detect association could be achieved by selecting a portion of the labeled data to be patients with extreme values of surrogate variables. Our method can also be easily extended to account for such extreme phenotype sampling for the labeled data, by adding weights to the estimator that are inversely proportional to the probability of being selected.

1.6 Appendix

In this appendix, we will establish properties of our estimator $\widehat{\theta}_{\text{SSL}}$. Throughout, we assume that $\mathbf{W}$, which includes G as an element, is bounded with $\mathbb{C}_{WW} = E(\mathbf{W}\mathbf{W}^{\intercal})$ positive definite and the joint density of Y and $\mathbf{W}$ is twice continuously differentiable. Furthermore, we assume that $\bar{\boldsymbol{\beta}}$ in an interior point of a compact set Ω. Let $\mathbf{X}_i = \mathbf{W}_i(G_i - \mu_g)$, $R_i = \bar{\boldsymbol{\beta}}^{\intercal}\mathbf{X}_i$, $\mathcal{P}(x, \boldsymbol{\beta}) = P(\boldsymbol{\beta}^{\intercal}\mathbf{X} \leq x)$, $\dot{\mathcal{P}}(x, \boldsymbol{\beta}) = \partial\mathcal{P}(x, \boldsymbol{\beta})/\partial x$, and $m(x, \boldsymbol{\beta}) = E(r_i \mid \boldsymbol{\beta}^{\intercal}\mathbf{X}_i = x)$. Since $\widehat{\mathcal{P}}(x, \boldsymbol{\beta})$ is estimated using the entire dataset, it follows from standard empirical processes theory (Pollard, 1990) that

$$\sup_{x,\boldsymbol{\beta}\in\Omega} \left|\widehat{\mathcal{G}}(x, \boldsymbol{\beta})\right| = O_p(N^{-\frac{1}{2}}), \quad \text{where} \quad \widehat{\mathcal{G}}(x, \boldsymbol{\beta}) = N^{\frac{1}{2}}\{\widehat{\mathcal{P}}(x, \boldsymbol{\beta}) - \mathcal{P}(x, \boldsymbol{\beta})\} \tag{1.2}$$

1.6.1 Consistency of our estimators

To establish the consistency of $\widehat{\theta}^{\text{par}}_{\text{SSL}}$ and $\widehat{\theta}_{\text{SSL}}$, we first note that $\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\| = O_p(n^{-\frac{1}{2}})$,

$$\max_{1\leq i\leq N} \|\widehat{\mathbf{X}}_i - \mathbf{X}_i\| = O_p(N^{-\frac{1}{2}}), \text{ and } \max \|\widehat{R}_i - R_i\| = O_p(n^{-\frac{1}{2}}).$$

Furthermore, since $\mathbf{W}$ includes 1 and G,

$$0 = E(Y_i - \mu_y) = E(\bar{\boldsymbol{\beta}}^{\intercal}\mathbf{W}_i), \quad \text{and} \quad E((Y_i - \mu_y)G_i) = E(\bar{\boldsymbol{\beta}}^{\intercal}\mathbf{W}_i G_i).$$

It follows that

$$E(R_i) = E(\bar{\boldsymbol{\beta}}^{\intercal}\mathbf{W}_i G_i) = E((Y_i - \mu_y)G_i) = E(r_i)$$

and hence $|\widehat{\theta}^{\text{par}}_{\text{SSL}} - \theta_0| \leq \max_{1\leq i\leq N} |\widehat{R}_i - R_i| + |N^{-1}\sum_{i=1}^{N} R_i - \theta_0| \to 0$ in probability. It follows from a Taylor series expansion that

$$n^{\frac{1}{2}}(\widehat{\theta}^{\text{par}}_{\text{SSL}} - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (\mathbb{C}^{-1}_{WW}\mathbb{C}_{WG})^{\intercal} \left\{ (\mathbf{W}_i - \boldsymbol{\mu}_W)(Y_i - \mu_y) - \mathbf{W}_i \bar{\boldsymbol{\beta}}^{\intercal}\mathbf{W}_i \right\} + o_p(1)$$

where $\mathbb{C}_{WG} = E\{\mathbf{W}_i(G_i - \mu_g)\}$ and $\boldsymbol{\mu}_W = E(\mathbf{W}_i)$. Since $\mathbf{W}$ includes 1 and G, it is straightforward to see that $G_i - \mu_g = (\mathbb{C}^{-1}_{WW}\mathbb{C}_{WG})^{\intercal}\mathbf{W}_i$ and $(\mathbb{C}^{-1}_{WW}\mathbb{C}_{WG})^{\intercal}\boldsymbol{\mu}_W = 0$. It follows that

$$n^{\frac{1}{2}}(\widehat{\theta}^{\text{par}}_{\text{SSL}} - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} (r_i - R_i) + o_p(1),$$

which converges in distribution to a normal with mean zero and variance $(\sigma^{\text{par}}_{\text{SSL}})^2 = E\{(r_i - R_i)^2\}$.

1.6.2 Asymptotic properties of our estimators

To derive asymptotic properties for $\widehat{\theta}_{\text{SSL}}$, we first write $\widehat{\theta}_{\text{SSL}} - \theta_0 = \widehat{\mathcal{W}}_{\text{SSL}}(\widehat{\boldsymbol{\beta}})$, with $\widehat{\mathcal{W}}_{\text{SSL}}(\boldsymbol{\beta}) = \widehat{\theta}_{\text{SSL}}(\boldsymbol{\beta}) - \theta_0(\boldsymbol{\beta})$ and our next goal is to show that

$$\widehat{\mathcal{W}}_{\text{SSL}}(\widehat{\boldsymbol{\beta}}) - \widehat{\mathcal{W}}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) \equiv \widehat{\mathcal{E}}_1 + \widehat{\mathcal{E}}_2 + \widehat{\mathcal{E}}_3 + \widehat{\mathcal{E}}_4 = o_p(n^{-\frac{1}{2}}).$$

where $\theta_0(\boldsymbol{\beta}) = \int m(x, \boldsymbol{\beta}) d\mathcal{P}(x, \boldsymbol{\beta}) = E\{E(r_i \mid \boldsymbol{\beta}^{\intercal}\mathbf{X}_i)\} = E(r_i) = \theta_0$,

$$\widehat{\mathcal{E}}_1 = \int \{\widehat{\mathcal{W}}_m(x, \widehat{\boldsymbol{\beta}}) - \widehat{\mathcal{W}}_m(x, \bar{\boldsymbol{\beta}})\} d\widehat{\mathcal{P}}(x, \widehat{\boldsymbol{\beta}}), \quad \widehat{\mathcal{W}}_m(x, \boldsymbol{\beta}) = \widehat{m}(x, \boldsymbol{\beta}) - m(x, \boldsymbol{\beta})$$

$$\widehat{\mathcal{E}}_2 = N^{-\frac{1}{2}} \int \{m(x, \widehat{\boldsymbol{\beta}}) - m(x, \bar{\boldsymbol{\beta}})\} d\mathcal{G}(x, \widehat{\boldsymbol{\beta}}), \quad \widehat{\mathcal{E}}_3 = N^{-\frac{1}{2}} \int \widehat{m}(x, \bar{\boldsymbol{\beta}}) d\{\widehat{\mathcal{G}}(x, \widehat{\boldsymbol{\beta}}) - \widehat{\mathcal{G}}(x, \bar{\boldsymbol{\beta}})\}$$

$$\widehat{\mathcal{E}}_4 = \int \widehat{\mathcal{W}}_m(x, \bar{\boldsymbol{\beta}}) d\{\mathcal{P}(x, \widehat{\boldsymbol{\beta}}) - \mathcal{P}(x, \bar{\boldsymbol{\beta}})\}.$$

To bound $\widehat{\mathcal{E}}_1$, we note that

$$\sup_{x, \boldsymbol{\beta}} |\widehat{m}(x, \boldsymbol{\beta}) + \widehat{b}(x, \boldsymbol{\beta}) - \widetilde{m}(x, \boldsymbol{\beta})| = o_p(n^{-\frac{1}{2}}),$$

where $\widehat{b}(x, \boldsymbol{\beta}) = (\widehat{\mu}_y - \mu_y)\mu_g(x, \boldsymbol{\beta})$, $\mu_g(x, \boldsymbol{\beta}) = E(G_i \mid \boldsymbol{\beta}^{\intercal}\mathbf{X}_i = x) - \mu_g$, and

$$\widetilde{m}(x, \boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i - x) r_i}{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\intercal}\mathbf{X}_i - x)}.$$

Let $\widetilde{\mathcal{W}}_m(x,\boldsymbol{\beta}) = \widetilde{m}(x,\boldsymbol{\beta}) - m(x,\boldsymbol{\beta})$. It then follows from the convergence of (1.2), the smoothness of $\mu_g(x,\boldsymbol{\beta})$ and the root-n convergence of $\widehat{\boldsymbol{\beta}}$ that

$$\begin{aligned}\widehat{\mathcal{E}}_1 &\leq o_p(n^{-\frac{1}{2}}) + \left|\int \{\widetilde{\mathcal{W}}_m(x,\widehat{\boldsymbol{\beta}}) - \widetilde{\mathcal{W}}_m(x,\bar{\boldsymbol{\beta}})\} d\widehat{\mathcal{P}}(x,\widehat{\boldsymbol{\beta}})\right| + |\widehat{\mu}_y - \mu_y| \left|\int \{\mu_g(x,\widehat{\boldsymbol{\beta}}) - \mu_g(x,\bar{\boldsymbol{\beta}})\} d\widehat{\mathcal{P}}(x,\widehat{\boldsymbol{\beta}})\right| \\ &\leq o_p(n^{-\frac{1}{2}}) + \left|\int \{\widetilde{\mathcal{W}}_m(x,\widehat{\boldsymbol{\beta}}) - \widetilde{\mathcal{W}}_m(x,\bar{\boldsymbol{\beta}})\} d\mathcal{P}(x,\widehat{\boldsymbol{\beta}})\right|\end{aligned}$$

To bound the last term above, we next aim to show that

$$\sup_{x,\boldsymbol{\beta}} \left|\frac{\partial \widetilde{m}(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial m(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right| = o_p(1). \tag{1.3}$$

To this end, we first note that for $q = 0, 1$,

$$\widehat{e}_q(x) = n^{-1}\sum_{i=1}^n K_h(\boldsymbol{\beta}^\intercal \mathbf{X}_i - x) r_i^q - E\{K_h(\boldsymbol{\beta}^\intercal \mathbf{X}_i - x) r_i^q\} = \int r^q K_h(s-x) d\{\widehat{\mathcal{P}}_{\boldsymbol{\beta}}(s,r) - \mathcal{P}_{\boldsymbol{\beta}}(s,r)\}$$

where $\widehat{\mathcal{P}}_{\boldsymbol{\beta}}(s,r) = n^{-1}\sum_{i=1}^n I(\boldsymbol{\beta}^\intercal \mathbf{X}_i \leq s, r_i \leq r)$ and $\mathcal{P}_{\boldsymbol{\beta}}(s,r) = P(\boldsymbol{\beta}^\intercal \mathbf{X}_i \leq s, r_i \leq r)$. From the strong approximation result of Tusnády Tusnády (1977), there exists a Gaussian process $\mathbb{G}_{\mathcal{P}_n}(s,r;\boldsymbol{\beta})$ such that

$$\sup_{s,\boldsymbol{\beta}} \left\| n^{\frac{1}{2}}\{\widehat{\mathcal{P}}_{\boldsymbol{\beta}}(s,r) - \mathcal{P}_{\boldsymbol{\beta}}(s,r)\} - \mathbb{G}_{\mathcal{P}_n}(s,r;\boldsymbol{\beta})\right\| = O\{n^{-\frac{1}{2}}\log(n)^2\}, \quad \text{almost surely.}$$

It follows that

$$\widehat{e}_q(x) = n^{-\frac{1}{2}} \int r^q K_h(s-x) d\mathbb{G}_{\mathcal{P}_n}(s,r;\boldsymbol{\beta}) + O\{(nh)^{-1}\log(n)^2\} = o[\{n^{-\frac{1}{2}} + (nh)^{-1}\}n^\epsilon]$$

In the last step above, we used the fact that $\sup_{x,\boldsymbol{\beta}} \| \int r^q K_h(s-x) d\mathbb{G}_{\mathcal{P}_n}(s,r;\boldsymbol{\beta})\| = o(n^\epsilon)$ for any $\epsilon > 0$ (Bickel and Rosenblatt, 1973). Therefore, we have

$$\sup_{\boldsymbol{\beta},x} \left| n^{-1}\sum_{i=1}^n K_h(\boldsymbol{\beta}^\intercal \mathbf{X}_i - x) r_i^q - E(r_i^q \mid \boldsymbol{\beta}^\intercal \mathbf{X}_i = x)\dot{\mathcal{P}}(x,\boldsymbol{\beta})\right| = o[\{n^{-\frac{1}{2}} + (nh)^{-1}\}n^\epsilon + h^2]$$

for any $\epsilon > 0$. Similarly, for any $\epsilon > 0$ and $l = 1, ..., p$,

$$\begin{aligned}&n^{-1}\sum_{i=1}^n \dot{K}_h(\boldsymbol{\beta}^\intercal \mathbf{X}_i - x) r_i^q X_{li} - E\{\dot{K}_h(\boldsymbol{\beta}^\intercal \mathbf{X}_i - x) r_i^q X_{li}\} \\ &= n^{-1/2} \int z K_h(s-x) d\mathbb{G}_{H_{ln}^{(q)}}(s,z;\boldsymbol{\beta}) + O\{h^{-1}n^{-2/3}\log(n)^{\tilde{d}}\} = o(n^{\epsilon - 1/2}h^{-1})\end{aligned}$$

where $H_l^{(q)}(s,z;\boldsymbol{\beta}) = P(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i \le s, r_i^q X_{li} \le z)$, $\widehat{H}_l^{(q)}(s,z;\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n I(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i \le s, r_i^q X_{li} \le z)$, and $\mathbb{G}_{H_n}(s,z;\boldsymbol{\beta})$ is a Gaussian process such that

$$\sup_{s,z,\boldsymbol{\beta}} \left\| n^{\frac{1}{2}}\{\widehat{H}_l^{(q)}(s,z;\boldsymbol{\beta}) - H_l^{(q)}(s,z;\boldsymbol{\beta})\} - \mathbb{G}_{H_{ln}^{(q)}}(s,z;\boldsymbol{\beta}) \right\| = O(n^{-1/6}\log(n)^{\tilde{d}}) \quad \text{almost surely.}$$

The existence of the Gaussian process is ensured by the results of Massart Massart (1989). Furthermore, by the standard Taylor series expansion for the bias term, we have

$$\sup_{\boldsymbol{\beta},x} \left\| n^{-1}\sum_{i=1}^n \dot{K}_h(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i - x) r_i^q \mathbf{X}_i - \frac{\partial E(r_i^q \mathbf{X}_i \mid \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i = x)}{\partial \boldsymbol{\beta}} \dot{\mathcal{P}}(x,\boldsymbol{\beta}) \right\| = o(n^{\epsilon - 1/2}h^{-1} + h)$$

for any $\epsilon > 0$. It follows that

$$\sup_{x,\boldsymbol{\beta}} \left| \frac{\partial \widetilde{m}(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial m(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right| = O(n^{\epsilon-1/2}h^{-1} + h) = o_p(1)$$

for any $\epsilon > 0$ provided that $h = O(n^{-\nu})$ for $\nu \in [1/5, 1/2)$. This, together with the root-n convergence of $\widehat{\boldsymbol{\beta}}$ and (1.3) implies that $\widehat{\mathcal{E}}_1 = o_p(n^{-\frac{1}{2}})$. Since $n/N \to 0$, it is straightforward to see that $|\widehat{\mathcal{E}}_2| + |\widehat{\mathcal{E}}_3| = o_p(n^{-\frac{1}{2}})$. From the uniform convergence of $\widetilde{m}(x,\boldsymbol{\beta})$ and (1.3) and the root-n convergence of $\widehat{\boldsymbol{\beta}}$, we have $\widehat{\mathcal{E}}_4 = o_p(n^{-\frac{1}{2}})$. It follows that $\widehat{\mathcal{W}}_{\text{SSL}}(\widehat{\boldsymbol{\beta}}) - \widehat{\mathcal{W}}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) = o_p(n^{-\frac{1}{2}})$ and therefore

$$n^{\frac{1}{2}}(\widehat{\theta}_{\text{SSL}} - \theta_0) = n^{\frac{1}{2}}\widehat{\mathcal{W}}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) + o_p(1) = n^{\frac{1}{2}}\{\widehat{\theta}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) - \theta_0\} + o_p(1).$$

Next, the consistency of $\widehat{\theta}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) = \int \widehat{m}(x,\bar{\boldsymbol{\beta}}) d\widehat{\mathcal{P}}(x,\bar{\boldsymbol{\beta}})$ follows directly from the uniform consistency of $\widehat{m}(x,\bar{\boldsymbol{\beta}})$ and $\widehat{\mathcal{P}}(x,\bar{\boldsymbol{\beta}})$. To derive the asymptotic distribution of $n^{\frac{1}{2}}\widehat{\mathcal{W}}_{\text{SSL}}(\bar{\boldsymbol{\beta}})$, we write $n^{\frac{1}{2}}\widehat{\mathcal{W}}_{\text{SSL}}(\bar{\boldsymbol{\beta}}) = \mathscr{I}_1 + \mathscr{I}_2 + \mathscr{I}_3$, where $\mathscr{I}_1 = (n/N)^{1/2}\int m(x,\bar{\boldsymbol{\beta}}) d\widehat{\mathcal{G}}(x,\bar{\boldsymbol{\beta}})$,

$$\mathscr{I}_2 = (n/N)^{\frac{1}{2}} \int \{\widehat{m}(x,\bar{\boldsymbol{\beta}}) - m(x,\bar{\boldsymbol{\beta}})\} d\widehat{\mathcal{G}}(x,\bar{\boldsymbol{\beta}}), \quad \text{and} \quad \mathscr{I}_3 = n^{\frac{1}{2}} \int \{\widehat{m}(x,\bar{\boldsymbol{\beta}}) - m(x,\bar{\boldsymbol{\beta}})\} d\mathcal{P}(x,\bar{\boldsymbol{\beta}}).$$

Since $\widehat{\mathcal{G}}(x,\bar{\boldsymbol{\beta}})$ converges weakly to a zero-mean Gaussian process and $n/N \to 0$, we have $\mathscr{I}_1 = o_p(1)$. The term $\mathscr{I}_2$ can be shown as $o_p(1)$ following Lemma A.1 of Chakrabortty and Cai Chakrabortty and Cai (2017). We then write

$$\begin{aligned}
\mathscr{I}_3 &= n^{\frac{1}{2}} \int \{\widetilde{m}(x,\bar{\boldsymbol{\beta}}) - m(x,\bar{\boldsymbol{\beta}})\} d\mathcal{P}(x,\bar{\boldsymbol{\beta}}) - n^{\frac{1}{2}}(\widehat{\mu}_y - \mu_y) \int \mu_g(x,\bar{\boldsymbol{\beta}}) d\mathcal{P}(x,\bar{\boldsymbol{\beta}}) \\
&= n^{-\frac{1}{2}} \sum_{i=1}^n \int K_h(\bar{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_i - x)\{r_i - m(x,\bar{\boldsymbol{\beta}})\} dx + o_p(1)
\end{aligned}$$

$$= n^{-\frac{1}{2}} \sum_{i=1}^{n} \{r_i - m(\bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X}_i, \bar{\boldsymbol{\beta}})\} + o_p(1) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \{r_i - E(r_i \mid R_i)\} + o_p(1)$$

It then follows that $n^{\frac{1}{2}}(\widehat{\theta}_{\text{SSL}} - \theta_0)$ converges in distribution to a normal with mean 0 and variance $\sigma^2_{\text{SSL}} = E\{\text{var}(r_i \mid \bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X}_i)\}$.

For the bias corrected estimator, following similar arguments as given above, we have

$$\widehat{\theta}_{\text{SSL}} - \widehat{\theta}^{\text{BC}}_{\text{SSL}} = \int \{\widehat{m}(x, \bar{\boldsymbol{\beta}}) - m(x, \bar{\boldsymbol{\beta}})\} d\mathcal{P}(x, \bar{\boldsymbol{\beta}}) + n^{-1} \sum_{i=1}^{n} \left\{ m(\bar{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X}_i, \bar{\boldsymbol{\beta}}) - r_i \right\} + o_p(n^{-\frac{1}{2}}) = o_p(n^{-\frac{1}{2}}),$$

where $\tilde{\mathcal{P}}(x, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathsf{T}} \mathbf{X}_i)$. Thus, $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$ is asymptotically equivalent to $\widehat{\theta}_{\text{SSL}}$ and thus $n^{\frac{1}{2}}(\widehat{\theta}^{\text{BC}}_{\text{SSL}} - \theta_0)$ also converges in distribution to a normal with mean 0 and variance σ^2_{SSL}.

Risk Prediction with Imperfect Survival Outcome Information from EHRs

Stephanie F. Chan

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Xuan Wang

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Tianxi Cai

Department of Biostatistics

Harvard T.H. Chan School of Public Health

2.1 Introduction

One major problem in using EMRs to build risk prediction models is that event times are difficult to determine from the data. Labor-intensive manual chart reviews are required to extract event times; however, if the clinical event occurs before the patient enters the specific EMR system, even chart reviews will not be able to recover the event time. Phenotyping chart reviews, on the other hand, are quick and can give us partial information on the survival times through obtaining disease status. Performing these phenotyping chart reviews gives rise to current status data, a type of survival data where the occurrence of a clinical condition is only determined at a single time of examination (in our case, the last doctor's visit before the chart review is performed). The exact time to disease is still unknown; however, we know whether or not the disease occurred prior to the examination time. In addition to disease status, the wide breadth of EMR data provides us with other related information on event times. For example, the first occurrence of a International Classification of Diseases, Ninth Revision (ICD9) diagnosis code related to the disease for a patient can give us an estimate of the time to occurrence of the disease. Alternatively, we can use the first natural language processing (NLP) mention of a term related to the disease in the doctor's notes as the estimate of the survival time. These mismeasured estimates can be used in conjunction with the current status data setup to obtain more efficient estimates of the effect of baseline covariates on survival.

Several regression models have been developed to analyze current status data, or the more general interval censored data. For example, under the proportional hazards (PH) model, Huang et al. (1996) proposed a nonparametric maximum likelihood estimator (NPMLE) approach to estimate the regression parameters, and Pan (1999) extended the iterative convex minorant (ICM) algorithm for interval-censored data. Under the proportional odds (PO) model, Rossini and Tsiatis (1996) used a maximum likelihood approach, and Huang and Rossini (1997) used sieve estimation procedures. Other regression methods under the additive hazard model and accelerated failure time (AFT) models have also been proposed (Lin et al., 1998; Chen and Sun, 2010; Betensky et al., 2001; Tian and Cai, 2006). More general models, such as the semiparametric linear transformation

model proposed by Sun and Sun (2005), have also been studied. However, most of these methods require some assumptions about the censoring distribution.

In this paper, we first propose a simple robust estimator for current status data defined by a set of kernel-weighted estimating equations that does not depend on the censoring distribution, under a nonparametric transformation (NPT) model, which includes PH, PO, AFT, and additive hazards models as special cases. We then propose an estimator that incorporates the information from the mismeasured estimates of the survival time, using the derivative of a rank estimator and combining it with our current status estimator.

The rest of the paper is formatted as follows. In section 2.2, we introduce our estimators for the regression coefficients. In section 2.3, we perform a simulation study to explore our methods and present the results of our simulations, and in section 2.4, we apply our methods to an example dataset using EMR data from the Partners HeathCare System. Concluding remarks are giving in section 2.5.

2.2 Methods

In this section, we detail our proposed estimator for estimating the effect of baseline covariates, denoted by $\mathbf{Z}$, on $t-$year survival. Suppose the full cohort consists of N subjects.

The true time to disease, denoted by T, is unobserved, but through chart reviews, we can obtain information about disease status for a small subset of patients $n << N$, denoted by δ, at the time that the chart review was performed, denoted by C. We note that $\delta = I(T \leq C)$. EMRs can also provide us with surrogates for T, such as the time of the first ICD9 code related to the disease or the time of the first NLP mention of the disease in the doctor's notes, denoted by $\boldsymbol{\mathscr{T}} = (\mathscr{T}_1, ..., \mathscr{T}_K)^\intercal$. Thus, the full underlying data is $\mathscr{F} = \{(\delta_i, T_i, C_i, \mathbf{Z}_i, \boldsymbol{\mathscr{T}}_i)^\intercal, i = 1, \ldots, n\}$, the labeled data is $\mathscr{L} = \{(\delta_i, C_i, \mathbf{Z}_i^\intercal, \boldsymbol{\mathscr{T}}_i^\intercal)^\intercal, i = 1, \ldots, n\}$, the unlabeled data is $\mathscr{U} = \{(C_i, \mathbf{Z}_i^\intercal, \boldsymbol{\mathscr{T}}_i^\intercal)^\intercal, i = n+1, \ldots, N\}$, and the observable data is $\mathscr{D} = \mathscr{L} \cup \mathscr{U}$. The censoring time C is assumed to be independent of T, $\mathbf{Z}$ and $\boldsymbol{\mathscr{T}}$.

We assume the following semi-parametric transformation (ST) failure time model:

$$P(T_i \leq t \mid \mathbf{Z}_i) = g(h_0(t) + \boldsymbol{\beta}_0^\intercal \mathbf{Z}_i)$$

where $g(\cdot)$ is a known smooth probability distribution function, $h_0(t)$ is an unspecified smooth increasing function. For each of the mis-measured survival outcome $\mathscr{T}_k$, we assume that

$$\log(\mathscr{T}_{ki}) = \log T_i + \epsilon_{ki}, \quad \text{for k = 1, ...,K,}$$

where ϵ_{ki} is independent of both T_i and $\mathbf{Z}_i$ with a completely unspecified distribution function and we also leave the correlation structure among $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_K)^\intercal$ unspecified.

2.2.1 Estimation

We can use the following estimating equations derived from Van Der Laan and Robins (1998) to obtain estimates for $h_0(t)$ and β:

$$\begin{aligned} \sum_{i=1}^{n} K_h(C_i - t_j)(\delta_i - g(h_0(t_j) + \beta'\mathbf{Z}_i)) &= 0 \text{ for all } j = 1 \dots k \\ \sum_{j=1}^{k}\sum_{i=1}^{n} K_h(C_i - t_j)\mathbf{Z}_i(\delta_i - g(h_0(t_j) + \beta'\mathbf{Z}_i) &= 0 \end{aligned}$$

We can solve this system of equations iteratively by first fixing β in the first equation and obtaining an estimate for $h_0(t_j)$ for all values of t_j, call them $\widehat{h}_0(t_j)$. We can then plug in $\widehat{h}_0(C_i)$ for $h_0(C_i)$ into the simplified second equation below to get an estimate for β.

$$\sum_{i=1}^{n} \mathbf{Z}_i(\delta_i - g(h_0(C_i) + \beta'\mathbf{Z}_i) = 0$$

We will call this estimator $\widehat{\beta}_\delta$. To derive an estimator leveraging $\boldsymbol{\mathscr{T}}$, we first consider the MRC estimator, proposed by Cai and Cheng (2007), as the maximizer of

$$Q_k(\beta) = \sum_{i \neq j} I(\beta'\mathbf{Z}_i > \beta'\mathbf{Z}_j)\frac{I(X^*_{ki} < X^*_{kj})\delta^*_{ki}}{\widehat{G}(X^*_{ki})^2}$$

where $X^*_{ki} = \min(C_i, \mathscr{T}_{ki})$, $\widehat{G}(t) = \frac{1}{n}\sum_{i=1}^{N} I(X^*_{ki} \leq C_i)$, and $\delta^*_{ki} = I(\mathscr{T}_{ki} \leq C_i)$. We note that this estimator can only estimate β up to a scalar. Modifying this objective function by using a kernel to approximate the indicator function, and taking the derivative of this function with respect to β, we get the following score function:

$$S_k(\beta) = \frac{\sum_{i \neq j}(\mathbf{Z}_i - \mathbf{Z}_j)K_h(\beta'\mathbf{Z}_i - \beta'\mathbf{Z}_j)\frac{I(X^*_{ki}<X^*_{kj})\delta^*_{ki}}{\widehat{G}(X^*_{ki})^2}}{\sum_{i \neq j}\frac{I(X^*_{ki}<X^*_{kj})\delta^*_{ki}}{\widehat{G}(X^*_{ki})^2}}$$

Although $\mathscr{T}$ does not follow a NPT model, we show in the Appendix that it does follow a single index model, so $E[S_k(\widehat{\mathcal{B}}_\delta)] = 0$ still holds, where $\widehat{\mathcal{B}}_\delta = \widehat{\beta}_\delta / ||\widehat{\beta}_\delta||_2$.

We define our combined estimator as $\widehat{\beta}_{SSL} = \widehat{\beta}_\delta + w' S(\widehat{\mathcal{B}}_\delta)$, where $S(\widehat{\mathcal{B}}_\delta) = (S_1(\widehat{\mathcal{B}}_\delta), \ldots, S_K(\widehat{\mathcal{B}}_\delta))^T$. We choose $w = -\Sigma_{SS}^{-1}\Sigma_{S\beta}$ so that the variance of $\widehat{\beta}_{SSL}$ is minimized, where Σ_{SS} is the variance-covariance matrix for $S(\widehat{\mathcal{B}}_\delta)$, and $\Sigma_{S\beta}$ is the covariance matrix between $S(\widehat{\mathcal{B}}_\delta)$ and $\widehat{\beta}_\delta$. To estimate $\Sigma_{SS}^{-1}\Sigma_{S\beta}$, we use perturbation resampling to obtain P perturbed estimates of $\widehat{\boldsymbol{\beta}}^*_\delta$ and $\mathbf{S}^*(\widehat{\mathcal{B}}^*_\delta)$ and then perform a ridge regression of $\widehat{\boldsymbol{\beta}}^*_\delta$ on $\mathbf{S}^*(\widehat{\mathcal{B}}^*_\delta)$ to account for potential overfitting from using multiple surrogates. We use the solution of the following objective function as our estimate for $\Sigma_{SS}^{-1}\Sigma_{S\beta}$:

$$\min_\alpha \frac{1}{P}\sum_{p=1}^{P}\frac{1}{2}\left(\widehat{\beta}^*_{\delta p} - \alpha^T S^*_p(\widehat{\mathcal{B}}^*_{\delta p})\right)^2 + \frac{1}{2}\lambda||\alpha||_2^2$$

where λ is chosen to have the minimum testing error via cross-validation.

2.3 Simulations

We conducted a simulation study to assess the performance of our estimator in finite sample settings. Throughout, we let $h_0(t) = 3\log(t/4)$. $\mathbf{Z}_i = (Z_{i,1}, Z_{i,2}, Z_{i,3})$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = 0.2 + 0.8I$, and T_i was generated from $\mathbf{Z}_i$ and β_0 using a inverse CDF transform. We consider the following two possible values for β_0:

$$\beta_{0A} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$$
$$\beta_{0B} = (1, 1, 1, -0.5, -0.5, -0.5, 1.2, 1.2, 1.2, 0)$$

We then generate C_i from a $Uniform(0, a)$ distribution, independent of T_i, where a is chosen such that $P(\delta_i = 1) \approx 0.5$. We generate $K = 2$ surrogates, $\mathscr{T}_1$ and $\mathscr{T}_2$. ϵ_k is generated from a mixture of normal distributions. We consider the following two sets of distributions of ϵ_k in our simulations:

$$\epsilon_{1,A} \sim D_{1i}N(0, 0.1) + (1 - D_{1i})N(0.1, 0.03) \qquad \epsilon_{2,A} \sim D_{2i}N(-0.05, 0.05) + (1 - D_{2i})N(0, 0.07)$$

$$\epsilon_{1,B} \sim D_{1i}N(0.2, 0.3) + (1 - D_{1i})N(-0.1, 0.1) \quad \epsilon_{2,B} \sim D_{2i}N(0, 0.2) + (1 - D_{2i})N(0.3, 0.1)$$

where D_{ki} follows a $Bernoulli(0.5)$ distribution. The bandwidth h for the kernel is chosen to be $\widehat{\tau} \times n^{-0.3}$, where $\widehat{\tau}$ is the empirical standard error of $\widehat{\beta}_\delta^T \mathbf{Z}$. We summarize results using 500 datasets. For the standard error estimates, we use 200 perturbations. In Table 2.1, we show results for the bias$\times 100$, mean square error (MSE), and relative efficiency (RE) of our point estimators. Both estimators have negligible biases, regardless of the choice for β_0 and ϵ_k. The relative efficiency of the semi-supervised estimator, compared to the β_δ estimator, ranges from 1.31 to 1.99.

We also investigated the performance of our standard error estimates. In Table 2.2, we show how the estimated standard error compares with the empirical standard error of the point estimates across the 500 datasets, as well as the coverage of β_0 from the 95% confidence interval constructed using our standard error estimate. We note that the empirical coverage probabilities range from 0.930 to 0.964, close to the nominal level.

2.4 Application to EMR study

We applied our proposed method to investigate the risk prediction potential of 21 genes associated with low density lipoprotein (LDL) cholesterol levels on developing coronary artery disease (CAD), in a cohort of rheumatoid arthritis (RA) patients. Coronary artery disease (CAD) is the global leading cause of death, killing 7.4 million people around the world annually. Unfortunately, genome-wide association studies have not found many reproducible genetic risk factors for CAD. However, one of the major risk factors for CAD, high LDL cholesterol levels, has been reproducibly shown to be associated with approximately 20 genetic loci.

The goal of our analysis is to examine the association between CAD and 21 genetic loci associated with LDL among RA patients, using an EMR cohort from the Partners

Table 2.1: Bias ($\times 100$) of the proposed estimators as well as the efficiency of the score estimator relative to the β_δ estimator (RE) under various choices for β_0, σ_1, and σ_2.

(I) $\beta_{0A} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$

		$\epsilon_{k,A}$		$\epsilon_{k,B}$	
	$Bias_\delta$	$Bias_{score}$	RE	$Bias_{score}$	RE
Z1	0.2327	-0.6468	1.5077	-0.4327	1.4618
Z2	0.8381	0.3501	1.6551	0.4035	1.6048
Z3	-0.4285	-1.0838	1.8785	-0.7074	1.7393
Z4	-0.1550	-0.7623	1.8916	-0.1732	1.8104
Z5	0.7304	-0.1255	1.7124	0.0729	1.5454
Z6	-0.2772	-0.7956	1.6077	-0.4364	1.5104
Z7	-0.0913	-0.3493	1.6987	-0.1909	1.5904
Z8	1.3498	-0.3360	1.8290	-0.2807	1.6497
Z9	-0.1146	-1.0931	1.9423	-0.6579	1.8230
Z10	0.1858	0.1851	1.6611	0.4462	1.5783

(II) $\beta_{0B} = (1, 1, 1, -0.5, -0.5, -0.5, 1.2, 1.2, 1.2, 0)$

		$\epsilon_{k,A}$		$\epsilon_{k,B}$	
	$Bias_\delta$	$Bias_{score}$	RE	$Bias_{score}$	RE
Z1	0.0421	-0.7663	1.4408	-0.5378	1.3571
Z2	0.0551	-0.0533	1.4544	-0.1404	1.4168
Z3	0.0884	-0.5620	1.4902	-0.0675	1.4150
Z4	-0.8258	0.8385	1.7733	0.6370	1.6743
Z5	0.6851	1.3377	1.5489	1.3396	1.4494
Z6	0.4795	1.2640	1.6818	0.9752	1.4756
Z7	-0.0820	-0.4629	1.4342	-0.0722	1.3411
Z8	0.1643	-0.2784	1.3727	0.0118	1.3190
Z9	0.0811	-0.6045	1.4489	-0.3663	1.3315
Z10	0.9032	1.4520	1.9967	1.4286	1.8970

HealthCare System (Liao et al., 2010). The RA EMR cohort originally consists of 4,453 patients. Of these patients, 1,311 had available genetic data on the 21 single nucleotide polymorphisms (SNPs) previously identified to be associated with LDL levels. Gold-standard labels for CAD status were provided by a rheumatologist using manual chart reviews (Liao et al., 2015). A total of 1,307 patients had definitive CAD status labels, which we use as our final dataset. For this analysis, we let δ_i be the CAD status and C_i be the patient's age at the time of the chart review. We use 5 surrogates for event time: the ICD9 billing codes for CAD and ischemic heart disease, as well as NLP mentions of arteriosclerotic heart disease, coronary disease, and myocardial infarction in the doctors'

Table 2.2: Empirical SE (ESE), average of the estimated SEs (ASE), and empirical coverage levels of the quantile based 95% CIs (CovP) for our estimators under various choices for β_0, σ_1, and σ_2

(I) $\beta_{0A} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$

	$\epsilon_{k,A}$			$\epsilon_{k,B}$		
	ESE	ASE	CovP	ESE	ASE	CovP
Z1	0.1294	0.1244	0.9420	0.1315	0.1254	0.9300
Z2	0.1302	0.1260	0.9440	0.1322	0.1272	0.9420
Z3	0.1172	0.1247	0.9560	0.1221	0.1260	0.9600
Z4	0.1204	0.1242	0.9520	0.1233	0.1251	0.9480
Z5	0.1211	0.1247	0.9480	0.1275	0.1260	0.9500
Z6	0.1164	0.1248	0.9520	0.1203	0.1257	0.9460
Z7	0.1165	0.1240	0.9620	0.1204	0.1258	0.9560
Z8	0.1221	0.1253	0.9520	0.1286	0.1263	0.9500
Z9	0.1163	0.1244	0.9560	0.1204	0.1259	0.9520
Z10	0.1244	0.1241	0.9460	0.1275	0.1257	0.9460

(II) $\beta_{0B} = (1, 1, 1, -0.5, -0.5, -0.5, 1.2, 1.2, 1.2, 0)$

	$\epsilon_{k,A}$			$\epsilon_{k,B}$		
	ESE	ASE	CovP	ESE	ASE	CovP
Z1	0.1104	0.1101	0.9520	0.1139	0.1112	0.9460
Z2	0.1176	0.1105	0.9360	0.1192	0.1117	0.9220
Z3	0.1125	0.1105	0.9540	0.1155	0.1117	0.9440
Z4	0.0945	0.0931	0.9360	0.0975	0.0944	0.9440
Z5	0.0942	0.0931	0.9440	0.0974	0.0945	0.9520
Z6	0.0890	0.0931	0.9640	0.0954	0.0946	0.9460
Z7	0.1136	0.1190	0.9540	0.1176	0.1200	0.9460
Z8	0.1215	0.1191	0.9420	0.1240	0.1205	0.9300
Z9	0.1179	0.1188	0.9380	0.1231	0.1201	0.9320
Z10	0.0862	0.0869	0.9540	0.0886	0.0885	0.9480

notes. Our time invariant covariates $\mathbf{Z}$ are the 21 LDL SNPs, race, and sex.

The point estimators, as well as their 95% confidence intervals are shown in Table 2.3. The results show that the point estimators for the β coefficients are reasonably consistent between the two methods; however, the 95% CIs from the semi-supervised method are always smaller than the current status method. The semi-supervised method identifies rs2902940 and rs11065987 as being significantly associated with CAD ($p < 0.05$), whereas the current status method does not identify any of the LDL SNPs as being associated with CAD.

Table 2.3: Point estimates and 95% CIs of the risk prediction potential of sex, race, and SNPs associated with LDL on CAD, along with p-values from a Z-test

	β_δ	p_δ	β_{score}	p_{score}
sex	-0.8505 (-1.3332, -0.3678)	0.0006	-0.7180 (-1.1313, -0.3046)	0.0007
race	0.0562 (-0.5754, 0.6879)	0.8615	0.2434 (-0.1843, 0.6711)	0.2647
rs2479409	-0.1315 (-0.4934, 0.2305)	0.4765	-0.1648 (-0.4767, 0.1472)	0.3006
rs2131925	0.1352 (-0.1803, 0.4508)	0.4009	0.1228 (-0.1419, 0.3874)	0.3632
rs2642442	0.0335 (-0.3145, 0.3815)	0.8503	0.0211 (-0.2577, 0.2998)	0.8823
rs1367117	0.0387 (-0.3376, 0.4150)	0.8402	-0.0078 (-0.3114, 0.2958)	0.9598
rs4299376	0.0653 (-0.2592, 0.3899)	0.6932	0.0744 (-0.2335, 0.3822)	0.6360
rs6882076	-0.1204 (-0.4574, 0.2167)	0.4840	-0.1358 (-0.4589, 0.1873)	0.4101
rs12670798	-0.0316 (-0.4350, 0.3717)	0.8778	-0.2091 (-0.5040, 0.0858)	0.1646
rs2072183	0.1750 (-0.2210, 0.5711)	0.3864	0.1603 (-0.2015, 0.5221)	0.3853
rs2081687	-0.0256 (-0.3505, 0.2992)	0.8770	-0.0473 (-0.3561, 0.2614)	0.7638
rs2255141	-0.0903 (-0.4430, 0.2624)	0.6158	0.0090 (-0.2784, 0.2963)	0.9512
rs174546	0.0258 (-0.3501, 0.4017)	0.8932	0.0011 (-0.2802, 0.2823)	0.9941
rs964184	0.0726 (-0.3815, 0.5267)	0.7539	0.0493 (-0.3662, 0.4648)	0.8161
rs11220462	-0.2930 (-0.8839, 0.2980)	0.3312	-0.2354 (-0.6884, 0.2175)	0.3083
rs11065987	-0.2510 (-0.5469, 0.0449)	0.0964	-0.2846 (-0.5088, -0.0604)	0.0128
rs1169288	-0.0206 (-0.3613, 0.3201)	0.9057	-0.0050 (-0.2876, 0.2777)	0.9725
rs8017377	-0.1515 (-0.5096, 0.2066)	0.4069	-0.0379 (-0.3469, 0.2711)	0.8100
rs3764261	-0.0275 (-0.3729, 0.3179)	0.8760	-0.0554 (-0.3205, 0.2097)	0.6822
rs2000999	0.3437 (-0.0768, 0.7642)	0.1091	0.3457 (-0.0079, 0.6992)	0.0554
rs7206971	0.1983 (-0.1294, 0.5260)	0.2357	0.2619 (-0.0073, 0.5311)	0.0566
rs4420638	0.1729 (-0.2373, 0.5830)	0.4087	0.0456 (-0.2868, 0.3780)	0.7880
rs2902940	0.3091 (-0.0198, 0.6381)	0.0655	0.3392 (0.0761, 0.6023)	0.0115

2.5 Discussion

We proposed two robust estimators to analyze current status data in the EMR setting. Our proposed SSL estimator is able to incorporate the imperfect estimates of survival time available in the EMR databases to improve on the current status β_δ estimator. Simulation results show that our estimators are consistent and that the SSL estimator is more efficient, compared to the β_δ estimator, which is confirmed by applying our method to a EMR cohort of RA patients.

In practice, we note that the baseline covariates for the model may need to be time-invariant covariates, such as sex or genetics, since it may be difficult to determine baseline covariates measured prior to developing the disease. If we can determine disease status

at baseline (ex. the disease of interest has a very specific ICD9 code), then we can also use time-dependent variables measured prior to baseline.

Our proposed combined estimator, along with its standard error estimate, can be used to test the risk prediction potential of baseline covariates on developing a disease or clinical event of interest, for example, using a Z-test. The improved efficiency of our estimator would lead to increased power in such tests. Further research could be done to use our combined estimator to estimate the risk of developing a disease.

2.6 Appendix

To show that $E[S_k(\widehat{\mathcal{B}}_\delta)] = 0$, we first show that each of the $\mathscr{T}_k$ follows a single index model. By direct calculation, we find that conditional distribution function of $\mathscr{T}_{ki}$ is

$$\begin{aligned} P(\mathscr{T}_{ki} \leq t|\mathbf{Z}_i) &= P(\log \mathscr{T}_{ki} \leq \log t|\mathbf{Z}_i) = P(\log T_i + \epsilon_{ki} \leq \log t|\mathbf{Z}_i) \\ &= P(\log T \leq \log t - \epsilon_{ki}|\mathbf{Z}_i) = P(T \leq te^{-\epsilon_{ki}}|\mathbf{Z}_i) \\ &= \int g(h_0(te^{-u}) + \boldsymbol{\beta}_0'\mathbf{Z}_i) f_{\epsilon_{ki}}(u)du, \end{aligned}$$

which is still a increasing function of $\boldsymbol{\beta}_0'\mathbf{Z}_i$. Thus, by Han (1987) and Sherman (1993), the maximizer of $Q_k(\beta)$ is consistent and asymptotically normal.

Let

$$\widetilde{S}_k(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} (\mathbf{Z}_i - \mathbf{Z}_j) K_h(\beta'\mathbf{Z}_i - \beta'\mathbf{Z}_j) I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj})$$

and

$$\begin{aligned} \tau(\beta) &= E\left[(\mathbf{Z}_i - \mathbf{Z}_j) K_h(\beta'\mathbf{Z}_i - \beta'\mathbf{Z}_j) I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj})\right. \\ &\quad \left. + (\mathbf{Z}_j - \mathbf{Z}_i) K_h(\beta'\mathbf{Z}_i - \beta'\mathbf{Z}_j) I(\mathscr{T}_{kj} \leq \mathscr{T}_{ki})\right] \\ &= 2E\left[\widetilde{S}_k(\beta)\right] \end{aligned}$$

In the following, we prove that $E\left[\widetilde{S}_k(\boldsymbol{\beta}_0)\right] = 0$ by showing that $E\left[\tau(\boldsymbol{\beta}_0)\right] = 0$. The results will also hold for $S_k(\boldsymbol{\beta}_0)$ since C is independent of T, $\mathbf{Z}$, and $\boldsymbol{\mathscr{T}}$.

$K_h(\cdot)$ is symmetric, so we note that

$$\tau(\boldsymbol{\beta}_0) = E\left[(\mathbf{Z}_i - \mathbf{Z}_j) K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{Z}_j) I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj})\right.$$

$$
\begin{aligned}
& +(\mathbf{Z}_j - \mathbf{Z}_i)K_h(\boldsymbol{\beta}_0'\mathbf{Z}_j - \boldsymbol{\beta}_0'\mathbf{Z}_i)I(\mathscr{T}_{kj} \leq \mathscr{T}_{ki})] \\
= \; & E\left\{E\left[(\mathbf{Z}_i - \mathbf{Z}_j)K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{Z}_j)(I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj}) - I(\mathscr{T}_{ki} \geq \mathscr{T}_{kj}))|\mathbf{Z}_j = \mathbf{z}\right]\right\} \\
= \; & E\left\{(\mathbf{Z}_i - \mathbf{z})K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{z})E\left[I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj}) - I(\mathscr{T}_{ki} \geq \mathscr{T}_{kj})|\mathbf{Z}_j = \mathbf{z}\right]\right\} \\
= \; & E\left\{(\mathbf{Z}_i - \mathbf{z})K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{z})E\left[I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj}) - I(\mathscr{T}_{ki} \geq \mathscr{T}_{kj})|\boldsymbol{\beta}_0'\mathbf{Z}_j = \boldsymbol{\beta}_0'\mathbf{z}\right]\right\}
\end{aligned}
$$

as $\mathscr{T}_{ki}$ follows a single index model. Let the first $p-1$ components of $\mathbf{z}$ be $\mathbf{r}$ and the p^{th} component of $\mathbf{z}$ be z_p. We also let $D(u) = E\left[I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj}) - I(\mathscr{T}_{ki} \geq \mathscr{T}_{kj})|\boldsymbol{\beta}_0'\mathbf{Z}_j = u\right]$ Then,

$$
\begin{aligned}
\tau(\boldsymbol{\beta}_0) &= \int (\mathbf{Z}_i - \mathbf{z})K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{z})D(\boldsymbol{\beta}_0'\mathbf{z})f_{\mathbf{Z}}(\mathbf{z})d\mathbf{z} \\
&= \int (\mathbf{Z}_i - \mathbf{z})K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{z})D(\boldsymbol{\beta}_0'\mathbf{z})f_{\mathbf{r}}(\mathbf{r})f_{z_p|\mathbf{r}}(z_p|\mathbf{r})d\mathbf{r}dz_p \\
&= \int (\mathbf{Z}_i - \mathbf{z})K_h(\boldsymbol{\beta}_0'\mathbf{Z}_i - \boldsymbol{\beta}_0'\mathbf{z})D(\boldsymbol{\beta}_0'\mathbf{z})f_{\mathbf{r}}(\mathbf{r})f_{\boldsymbol{\beta}_0'\mathbf{z}|\mathbf{r}}(\boldsymbol{\beta}_0'\mathbf{z}|\mathbf{r})d(\boldsymbol{\beta}_0'\mathbf{z})d\mathbf{r} \\
&= \int (\mathbf{Z}_i - \mathbf{z})D(\boldsymbol{\beta}_0'\mathbf{Z}_i)f_{\mathbf{r},\boldsymbol{\beta}_0'\mathbf{Z}_i}(\mathbf{r}, \boldsymbol{\beta}_0'\mathbf{Z}_i)d\mathbf{r} + o(h) \\
&= \int (\mathbf{Z}_i - \mathbf{z})D(\boldsymbol{\beta}_0'\mathbf{Z}_i)f_{\mathbf{r}|\boldsymbol{\beta}_0'\mathbf{Z}_i}(\mathbf{r}|\boldsymbol{\beta}_0'\mathbf{Z})f_{\boldsymbol{\beta}_0'\mathbf{Z}_i}(\boldsymbol{\beta}_0'\mathbf{Z}_i)d\mathbf{r} + o(h) \\
&= (\mathbf{Z}_i - E\left[\mathbf{Z}_i|\boldsymbol{\beta}_0'\mathbf{Z}_i\right])D(\boldsymbol{\beta}_0'\mathbf{Z}_i)f_{\boldsymbol{\beta}_0'\mathbf{Z}}(\boldsymbol{\beta}_0'\mathbf{Z}_i) + o(h)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E\left[\tau(\boldsymbol{\beta}_0)\right] &= E\left[(\mathbf{Z}_i - E\left[\mathbf{Z}_i|\boldsymbol{\beta}_0'\mathbf{Z}_i\right])D(\boldsymbol{\beta}_0'\mathbf{Z}_i)f_{\boldsymbol{\beta}_0'\mathbf{Z}}(\boldsymbol{\beta}_0'\mathbf{Z}_i)\right] \\
&= E\left\{E\left[(\mathbf{Z}_i - E\left[\mathbf{Z}_i|\boldsymbol{\beta}_0'\mathbf{Z}_i\right])D(\boldsymbol{\beta}_0'\mathbf{Z}_i)f_{\boldsymbol{\beta}_0'\mathbf{Z}}(\boldsymbol{\beta}_0'\mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{z}\right]\right\} \\
&= E\left\{(\mathbf{z} - E\left[\mathbf{Z}_i|\boldsymbol{\beta}_0'\mathbf{Z}_i = \boldsymbol{\beta}_0'\mathbf{z}\right])f_{\boldsymbol{\beta}_0'\mathbf{Z}}(\boldsymbol{\beta}_0'\mathbf{z}_i)E\left[D(\boldsymbol{\beta}_0'\mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{z}\right]\right\} \\
&= E\left\{(\mathbf{z} - E\left[\mathbf{Z}_i|\boldsymbol{\beta}_0'\mathbf{Z}_i = \boldsymbol{\beta}_0'\mathbf{z}\right])f_{\boldsymbol{\beta}_0'\mathbf{Z}}(\boldsymbol{\beta}_0'\mathbf{z}_i)E\left[D(\boldsymbol{\beta}_0'\mathbf{Z}_i)|\boldsymbol{\beta}_0'\mathbf{Z}_i = \boldsymbol{\beta}_0'\mathbf{z}\right]\right\}
\end{aligned}
$$

since $\mathscr{T}_{ki}$ follows a single index model. We note that

$$
\begin{aligned}
& E\left[D(\boldsymbol{\beta}_0'\mathbf{Z}_i)|\mathbf{Z}_i = \mathbf{z}\right] \\
& = E\left[I(\mathscr{T}_{ki} \leq \mathscr{T}_{kj}) - I(\mathscr{T}_{ki} \geq \mathscr{T}_{kj})|\boldsymbol{\beta}_0'\mathbf{Z}_j = \boldsymbol{\beta}_0'\mathbf{z}, \boldsymbol{\beta}_0'\mathbf{Z}_i = \boldsymbol{\beta}_0'\mathbf{z}\right] \\
& = 0
\end{aligned}
$$

so $E\left[\tau(\boldsymbol{\beta}_0)\right] = 2E\left[\widetilde{S}_k(\beta)\right] = 0.$

Estimating Risk with Imperfect Survival Outcome Information from EHRs

Stephanie F. Chan

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Tianxi Cai

Department of Biostatistics

Harvard T.H. Chan School of Public Health

3.1 Introduction

In the previous chapter, we focused on estimating the effects of covariates on the risk of developing a disease in our risk prediction model; however, in this chapter, we will focus on predicting the actual risk of developing the disease using EHR data, which will be useful in aiding clinicians in their decision making. As mentioned previously, exact event times are difficult to obtain in an EHR setting; however, EHRs can still provide us with mismeasured estimates of event time, such as the time to the first ICD9 diagnosis code for a patient or the first NLP mention of a term related to the disease in the clinicians' notes. In lieu of the exact event times, we can use these mismeasured estimates to build our risk prediction model.

Several methods have been developed to account for measurement errors in the covariates; however, there is much less literature on methods to handle measurement error in survival outcomes. Under parametric models such as proportional hazards (PH) and accelerated failure time (AFT), Skinner and Humphreys (1999) use a bias-corrected maximum likelihood estimator to handle multiplicative for survival time, and Oh et al. (2018) extend the SIMEX method developed by Cook and Stefanski (1994) to handle measurement error in survival outcomes. Comte et al. (2017) uses a nonparametric quotient estimator to obtain estimates of the hazard function, survival function, and density function. However, none of these methods are able to handle multiple mismeasured estimates of survival time, which can be combined to obtain a more precise estimate for risk prediction.

In this paper, we propose a simple maximum likelihood estimator for $t-$year survival using multiple mismeasured estimates of survival time under a semiparametric transformation (ST) model, which includes proportion hazards and proportional odds models as special cases. The rest of the paper is formatted as follows. In section 3.2, we introduce our estimators for the regression coefficients. In section 3.3, we perform a simulation study to explore our methods and present the results of our simulations, and in section 3.4, we apply our methods to an example dataset using EMR data from the Partners Biobank. Concluding remarks are giving in section 3.5.

3.2 Methods

In this section, we detail our proposed estimator for estimating the effect of baseline covariates, denoted by $\mathbf{Z}$, on $t-$year survival. Suppose the full cohort consists of n subjects.

Let T denote the unobservable true time to disease and C denote the follow up time. EMRs can provide us with imprecise proxies for T, such as the time of the first ICD9 code related to the disease or the time of the first NLP mention of the disease in the doctor's notes. We denote these proxies by $\boldsymbol{\mathscr{T}}^* = (\mathscr{T}_1^*, ..., \mathscr{T}_K^*)^\intercal$. $\boldsymbol{\mathscr{T}}^*$ is observable through $\mathbf{X}^* = (X_1^*, ..., X_K^*)$ and $\boldsymbol{\Delta}^* = (\Delta_1^*, ..., \Delta_K^*)$, where $X_k^* = \min(\mathscr{T}_k^*, C)$ and $\Delta_k^* = I(\mathscr{T}_k^* \leq C)$. The full underlying data is $\mathscr{F} = \{(\Delta_i, T_i, C_i, \mathbf{Z}_i^\intercal, \boldsymbol{\mathscr{T}}_i^{*\intercal}, \mathbf{X}_i^{*\intercal}, \boldsymbol{\Delta}_i^{*\intercal})^\intercal, i = 1, \ldots, n\}$, and the observable data is $\mathscr{D} = \{(C_i, \mathbf{Z}_i^\intercal, \mathbf{X}_i^{*\intercal}, \boldsymbol{\Delta}_i^{*\intercal})^\intercal, i = 1, \ldots, n\}$ where n is the number of patients. The censoring time C is assumed to be independent of T, $\boldsymbol{\mathscr{T}}^*$, and $\mathbf{Z}$.

We assume the following semi-parametric transformation (ST) failure time model for $\log T$:

$$P(\log T_i \leq t \mid \mathbf{Z}_i) = g(h_0(t) + \boldsymbol{\beta}_0^\intercal \mathbf{Z}_i) = \mathcal{G}\left(\mathcal{H}(t) e^{\boldsymbol{\beta}_0^\intercal \mathbf{Z}_i}\right)$$

where $g(\cdot)$ is a known smooth probability distribution function, $h_0(t)$ is an unspecified smooth increasing function, $\mathcal{H}(t) = e^{h_0(t)}$, and $\mathcal{G}(x) = g(\log x)$. For each of the mismeasured survival outcome $\mathscr{T}_k^*$, we assume that

$$\log(\mathscr{T}_{ki}^*) = \log T_i + \epsilon_{ki}, \quad \text{for k = 1, ...,K,}$$

where ϵ_{ki} is independent of T_i, C, and $\mathbf{Z}_i$ and has a distribution function f_k known up to a parameter vector $\boldsymbol{\alpha}_k$. Thus, $\mathscr{T}_{ki}^*$ can be modeled as:

$$P(\log \mathscr{T}_{ki}^* \leq t \mid \mathbf{Z}_i) = \int \mathcal{G}\left(\mathcal{H}(t - \epsilon) e^{\boldsymbol{\beta}_0^\intercal \mathbf{Z}_i}\right) f_k(\epsilon, \boldsymbol{\alpha}_k) d\epsilon$$

3.2.1 Estimation

We are interested in estimating $\boldsymbol{\beta}_0$ and $h_0(t)$ together with the nuisance parameters $\boldsymbol{\alpha}_0 = (\alpha_1, ..., \alpha_K)$. We approximate $h_0(t)$ using a regression spline such that $\mathcal{H}(t) = \int_{-\infty}^{t} e^{\boldsymbol{\gamma}_0^\intercal \mathbf{B}(s)} ds$, where $\mathbf{B}(s)$ is composed of κ linear basis functions. Thus, for $k = 1, ..., K$, we can approximate $P(\log \mathscr{T}_{ki}^* \leq t \mid \mathbf{Z}_i)$ as $\pi_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i)$:

$$\pi_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i) = \int \mathcal{G}\left(\int_{-\infty}^{t-\epsilon} e^{\boldsymbol{\gamma}^\intercal B(s) + \boldsymbol{\beta}^\intercal \mathbf{Z}_i} ds\right) f_k(\epsilon, \boldsymbol{\alpha}_k) d\epsilon$$

We estimate $\theta = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ as the maximum composite likelihood estimator $\widehat{\theta} = \text{argmax}_\theta \, \widehat{\ell^*}(\theta)$, where $\widehat{\ell^*}(\theta)$ is the log-likelihood

$$\widehat{\ell^*}(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \Delta_{ki}^* \log \dot{\pi}_k^*(X_{ki}^*; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i) + (1 - \Delta_{ki}^*) \log(1 - \pi_k^*(X_{ki}^*; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i))$$

,

and $\dot{\pi}_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i) = \partial \pi_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i)/\partial t$. The resulting estimator for $h_0(t)$ is $\widehat{h}(t) = \log \widehat{\mathcal{H}}(t) = \log\left(\int_{-\infty}^{t} e^{\widehat{\gamma}^\intercal B(s)} ds\right)$. Thus, our estimate for the risk is:

$$\widehat{P}(\log T_i \leq t \mid \mathbf{Z}_i) = g(\widehat{h}(t) + \widehat{\boldsymbol{\beta}}^\intercal \mathbf{Z}_i)$$

3.3 Simulations

We conducted a simulation study to assess the performance of our estimator in finite sample settings. Throughout, we let $n = 5000$. $\mathbf{Z}_i = (Z_{i,1}, Z_{i,2}, Z_{i,3})$ was generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = 0.2 + 0.8I$, and $\log T_i$ was generated from $\mathbf{Z}_i$, $h_0(t)$ and β_0 using a inverse CDF transform. We consider the following three possible forms for $h_0(t)$:

$$\begin{aligned}
h_{0,linear}(t) &= -15 + 5t \\
h_{0,cubic}(t) &= t^3 \\
h_{0,probit}(t) &= t + 100\Phi\left(\frac{t}{10} - 0.2\right) - 55
\end{aligned}$$

We then generate C_i from a $Uniform(a, b)$ distribution, independent of T_i, where a and b are chosen such that $P(T_i \leq C_i) \approx 0.5$. To estimate $h_0(t)$, we use $\kappa = 6$ knots for our basis spline $B(s)$ defined at the 10th, 20th, 40th, 60th, 80th, and 90th percentiles of C. We generate $K = 2$ surrogates, $\mathscr{T}_1^*$ and $\mathscr{T}_2^*$. The distribution of the errors for our surrogates follows a normal distribution with $\epsilon_1 \sim N(0, 0.3)$ and $\epsilon_2 \sim N(0, 0.1)$.

To calculate the likelihood, we use Gaussian-Hermite quadrature to estimate integrals over ϵ in both $\pi_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i)$ and $\dot{\pi}_k^*(t; \boldsymbol{\alpha}_k, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{Z}_i)$. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is used to maximize the likelihood, with initial values for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ estimated using logistic regression.

In the following tables, we show results for the mean and standard error of the estimates using our method, compared to a naive method, where the $\boldsymbol{\mathscr{T}}^*$ are assumed to be the true $\mathbf{T}$, for different forms of $h_0(t)$

The mean of our estimates using the $h_{0,linear}(t)$ are close to the true values, and using multiple surrogates definitely improves the efficiency of our estimates. The mean of the estimates using our method with $h_{0,cubic}(t)$ show bias, especially in the estimates of $\boldsymbol{\alpha}$. The efficiency of the results is also increased when using multiple surrogates, except for $\boldsymbol{\alpha}$. The bias in the results using $h_{0,probit}(t)$ is small; however, the efficiency of the estimates for $\boldsymbol{\alpha}$ and $h(t)$ are larger when using multiple surrogates. For all settings, our method shows significantly less bias than the naive method.

3.4 Application to EMR study

We applied our proposed method to build a risk prediction model for developing type 2 diabetes from genetic markers associated with obesity. In 2011, diabetes was estimated to be the 7th leading cause of death in the US (Heron, 2015), and type 2 diabetes accounts for 90% of all diabetes cases. Type 2 diabetes is characterized by insulin resistance, a condition where cells do not respond properly to insulin, which leads to high glucose levels in the blood. A well known risk factor for type 2 diabetes is obesity; however, the exact mechanism that links obesity and type 2 diabetes is unknown (Eckel et al., 2011). The increase in the prevalence of obesity has led to an increase in diabetes cases.

The goal of our analysis is to predict the risk of developing type 2 diabetes given their obesity risk score and other demographic information, using an EMR cohort from the Partners Biobank. The Partners Biobank consists of 38,345 patients. 20,091 patients have available genetic data from which we can obtain a genetic risk score for obesity. Of these patients, 17,220 did not have any ICD9 codes or NLP mentions of diabetes within

Table 3.1: Mean and standard error of estimates, comparing our method to a naive method for one surrogate and two surrogates, using $h_{0,linear}(t)$

(I) Using one surrogate, $\mathscr{T}_1^*$ where $\epsilon_1 \sim Normal(0, 0.3)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.2620	0.0727	-	-
$h(-1)$	-20.0000	-19.0086	1.5311	-15.1606	0.2692
$h(0)$	-15.0000	-14.2550	1.1459	-11.3623	0.1967
$h(1)$	-10.0000	-9.5014	0.7613	-7.5641	0.1250
$h(2)$	-5.0000	-4.7477	0.3791	-3.7658	0.0575
$h(3)$	0.0000	0.0027	0.0478	0.0182	0.0316
β_1	2.0000	1.8964	0.1569	1.4961	0.0361
β_2	-1.0000	-0.9520	0.0866	-0.7511	0.0306
β_3	0.0000	0.0013	0.0351	0.0008	0.0276

(II) Using one surrogate, $\mathscr{T}_2^*$ where $\epsilon_2 \sim Normal(0, 0.1)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.1000	0.0937	0.0520	-	-
$h(-1)$	-20.0000	-20.1979	0.9396	-19.2326	0.3479
$h(0)$	-15.0000	-15.1486	0.7039	-14.4226	0.2556
$h(1)$	-10.0000	-10.0994	0.4686	-9.6126	0.1640
$h(2)$	-5.0000	-5.0501	0.2350	-4.8027	0.0754
$h(3)$	0.0000	-0.0004	0.0355	0.0042	0.0323
β_1	2.0000	2.0203	0.1002	1.9188	0.0395
β_2	-1.0000	-1.0119	0.0545	-0.9612	0.0326
β_3	0.0000	-0.0003	0.0309	-0.0002	0.0294

(III) Combining both $\mathscr{T}_1^*$ and $\mathscr{T}_2^*$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.2918	0.0150	-	-
	0.1000	0.0603	0.0512	-	-
$h(-1)$	-20.0000	-19.7093	0.7370	-16.8907	0.2623
$h(0)$	-15.0000	-14.7815	0.5501	-12.6648	0.1926
$h(1)$	-10.0000	-9.8537	0.3638	-8.4389	0.1236
$h(2)$	-5.0000	-4.9259	0.1794	-4.2130	0.0575
$h(3)$	0.0000	0.0009	0.0338	0.0072	0.0279
β_1	2.0000	1.9693	0.0760	1.6808	0.0334
β_2	-1.0000	-0.9873	0.0453	-0.8428	0.0281
β_3	0.0000	0.0004	0.0295	0.0004	0.0255

the first year of entering and can be assumed to not have diabetes at the time they entered the EMR system, which we use as our final dataset. We use 2 surrogates for event time: the ICD9 billing codes for type 2 diabetes, and NLP mentions of non-insulin dependent diabetes. Our covariates include the obesity risk score, age, race, and sex. The predicted risks and bootstrap confidence intervals for a higher risk group and a lower risk group are shown in Table 3.4.

As we expect, the patients in the higher risk group (high obesity GRS, non-white, male) have a higher risk of developing diabetes than the lower risk group (low obesity GRS, white, female).

3.5 Discussion

We propose an estimator for the risk of developing a disease using a maximum likelihood estimator under the semiparametric transformation model. Our proposed estimator is able to incorporate multiple estimates of survival time, subject to measurement error, available in EMR databases. Simulation results show that our estimators are consistent. This is confirmed by applying our method to an EMR dataset. Our proposed estimator can be used to aid clinicians in patient care.

In practice, we note that we need to ensure that our baseline covariates are measured prior to developing the disease. Since it may be difficult to determine exactly when a patient developed a disease from EMR data, especially if they already have the disease prior to entering the EMR system, our baseline covariates may need to be time-invariant covariates, such as sex, race, or genetics. If we can determine disease status at our chosen baseline, for example, in cases where the disease has a very specific ICD9 code associated with it, then we can additionally use time-dependent covariates that are measured prior to the baseline.

Table 3.2: Mean and standard error of estimates, comparing our method to a naive method for one surrogate and two surrogates, using $h_{0,cubic}(t)$

(I) Using one surrogate, $\mathscr{T}_1^*$ where $\epsilon_1 \sim Normal(0, 0.3)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.2798	0.0123	-	-
$h(-1)$	-1.0000	-0.9307	0.0503	-1.0569	0.0388
$h(0)$	0.0000	0.0296	0.0431	0.0258	0.0398
$h(1)$	1.0000	0.9348	0.0705	1.0438	0.0497
$h(2)$	8.0000	7.3098	1.8909	5.2352	0.5497
$h(3)$	27.0000	17.0197	8.0744	11.1284	2.9227
β_1	2.0000	1.9509	0.0535	1.6644	0.0385
β_2	-1.0000	-0.9790	0.0390	-0.8354	0.0314
β_3	0.0000	0.0002	0.0347	0.0009	0.0293

(II) Using one surrogate, $\mathscr{T}_2^*$ where $\epsilon_2 \sim Normal(0, 0.1)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.1000	0.1706	0.0128	-	-
$h(-1)$	-1.0000	-0.8660	0.0462	-0.9660	0.0399
$h(0)$	0.0000	0.0241	0.0455	0.0294	0.0443
$h(1)$	1.0000	0.9061	0.0609	0.9573	0.0517
$h(2)$	8.0000	10.0040	2.7613	7.4866	0.9582
$h(3)$	27.0000	26.3851	11.0993	17.6178	4.2130
β_1	2.0000	2.1249	0.0544	1.9294	0.0407
β_2	-1.0000	-1.0647	0.0408	-0.9668	0.0338
β_3	0.0000	-0.0006	0.0346	-0.0006	0.0314

(III) Combining both $\mathscr{T}_1^*$ and $\mathscr{T}_2^*$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.2960	0.0085	-	-
	0.1000	0.1455	0.0142	-	-
$h(-1)$	-1.0000	-0.9005	0.0439	-1.0086	0.0362
$h(0)$	0.0000	0.0270	0.0434	0.0268	0.0408
$h(1)$	1.0000	0.9159	0.0553	0.9985	0.0453
$h(2)$	8.0000	8.5285	1.6834	5.9829	0.4851
$h(3)$	27.0000	21.1134	6.9960	12.9105	2.3768
β_1	2.0000	2.0304	0.0477	1.7716	0.0360
β_2	-1.0000	-1.0181	0.0364	-0.8885	0.0302
β_3	0.0000	-0.0003	0.0332	0.0002	0.0285

Table 3.3: Mean and standard error of estimates, comparing our method to a naive method for one surrogate and two surrogates, using $h_{0,probit}(t)$

(I) Using one surrogate, $\mathscr{T}_1^*$ where $\epsilon_1 \sim Normal(0, 0.3)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.2711	0.0476	0.0000	0.0000
$h(-1)$	-17.7911	-18.5730	9.0644	-15.0358	0.9420
$h(0)$	-12.9259	-13.2049	5.1288	-10.6599	0.5482
$h(1)$	-7.9827	-7.8368	1.2610	-6.2839	0.1696
$h(2)$	-3.0000	-2.8555	0.2296	-2.2235	0.0441
$h(3)$	1.9827	1.8956	0.1723	1.4697	0.0454
β_1	2.0000	1.9133	0.1573	1.4969	0.0371
β_2	-1.0000	-0.9601	0.0854	-0.7517	0.0320
β_3	0.0000	0.0020	0.0365	0.0012	0.0288

(II) Using one surrogate, $\mathscr{T}_2^*$ where $\epsilon_2 \sim Normal(0, 0.1)$

		real		naive	
	truth	mean	sd	mean	sd
α	0.1000	0.1123	0.0589	0.0000	0.0000
$h(-1)$	-17.7911	-18.5127	2.1319	-18.6906	11.3449
$h(0)$	-12.9259	-13.3725	1.2968	-13.2870	6.3834
$h(1)$	-7.9827	-8.2323	0.5367	-7.8834	1.4357
$h(2)$	-3.0000	-3.1022	0.2199	-2.8737	0.0541
$h(3)$	1.9827	2.0433	0.1336	1.9017	0.0602
β_1	2.0000	2.0704	0.1404	1.9197	0.0441
β_2	-1.0000	-1.0342	0.0729	-0.9622	0.0370
β_3	0.0000	-0.0009	0.0329	0.0004	0.0308

(III) Combining both $\mathscr{T}_1^*$ and $\mathscr{T}_2^*$

		real		naive	
	truth	mean	sd	mean	sd
α	0.3000	0.3493	1.2935	0.0000	0.0000
	0.1000	0.1251	1.3026	0.0000	0.0000
$h(-1)$	-17.7911	-19.3502	16.9511	-16.3144	5.4570
$h(0)$	-12.9259	-13.8913	12.8506	-11.6198	3.0888
$h(1)$	-7.9827	-8.4324	9.6413	-6.9251	0.7292
$h(2)$	-3.0000	-3.2151	5.5881	-2.5139	0.0480
$h(3)$	1.9827	1.7406	3.3052	1.6618	0.0404
β_1	2.0000	1.9571	0.0973	1.6824	0.0383
β_2	-1.0000	-0.9809	0.0564	-0.8439	0.0312
β_3	0.0000	0.0008	0.0310	0.0005	0.0275

Group	10-Year Risk	20-Year Risk
high obesity GRS non-white male	17.8% (15.6%, 27.8%)	48.0% (35.3%, 66.7%)
low obesity GRS white female	8.4% (7.5%, 13.3%)	28.1% (19.1%, 44.0%)

References

ANANTHAKRISHNAN, A., CAI, T., SAVOVA, G., CHENG, S., CHEN, P., PEREZ, R., GAINER, V., MURPHY, S., SZOLOVITS, P., XIA, Z., SHAW, S., CHURCHILL, S., KARLSON, E., KOHANE, I., PLENGE, R. and LIAO, K. (2013). Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach. *Inflammatory Bowel Diseases* **19** 1411–1420.

BETENSKY, R. A., RABINOWITZ, D. and TSIATIS, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88** 703–711.

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics* 1071–1095.

CAI, T. and CHENG, S. (2007). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9** 216–233.

CHAKRABORTTY, A. and CAI, T. (2017). Efficient and adaptive linear regression in semi-supervised settings. *arXiv:1701.04889* Retrieved from `https://arxiv.org/pdf/1701.04889.pdf`.

CHEN, L. and SUN, J. (2010). A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Computational statistics & data analysis* **54** 1109–1116.

COMTE, F., SAMSON, A. and STIRNEMANN, J. J. (2017). Hazard estimation with censoring and measurement error: application to length of pregnancy. *TEST* 1–22.

COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association* **89** 1314–1328.

DENNY, J. C., RITCHIE, M. D., BASFORD, M. A., PULLEY, J. M., BASTARACHE, L., BROWN-GENTRY, K., WANG, D., MASYS, D. R., RODEN, D. M. and CRAWFORD, D. C. (2010). Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26** 1205–1210.

DLIGACH, D., MILLER, T. and SAVOVA, G. K. (2015). Semi-supervised learning for phenotyping tasks. In *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association.

ECKEL, R. H., KAHN, S. E., FERRANNINI, E., GOLDFINE, A. B., NATHAN, D. M., SCHWARTZ, M. W., SMITH, R. J. and SMITH, S. R. (2011). Obesity and type 2 diabetes: what can be unified and what needs to be individualized? *The Journal of Clinical Endocrinology & Metabolism* **96** 1654–1663.

GABAY, C. and KUSHNER, I. (1999). Acute-phase proteins and other systemic responses to inflammation. *New England Journal of Medicine* **340** 448–454.

HAN, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35** 303–316.

HERON, M. P. (2015). Deaths: leading causes for 2011 .

HUANG, J. and ROSSINI, A. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association* **92** 960–967.

HUANG, J. ET AL. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24** 540–568.

KAWAKITA, M. and KANAMORI, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine learning* **91** 189–209.

KENWARD, M. G. and CARPENTER, J. (2007). Multiple imputation: current perspectives. *Statistical methods in medical research* **16** 199–218.

KIM, J. and SHIN, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association* **20** 613–618.

LIAO, K. P., ANANTHAKRISHNAN, A. N., KUMAR, V., XIA, Z., CAGAN, A., GAINER, V. S., GORYACHEV, S., CHEN, P., SAVOVA, G. K., AGNIEL, D. ET AL. (2015). Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PloS one* **10** e0136651.

LIAO, K. P., CAI, T., GAINER, V., GORYACHEV, S., ZENG-TREITLER, Q., RAYCHAUDHURI, S., SZOLOVITS, P., CHURCHILL, S., MURPHY, S., KOHANE, I. ET AL. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* **62** 1120–1127.

LIAO, K. P., KURREEMAN, F., LI, G., DUCLOS, G., MURPHY, S., GUZMAN, P. R., CAI, T., GUPTA, N., GAINER, V., SCHUR, P. ET AL. (2013). Autoantibodies, autoimmune risk alleles and clinical associations in rheumatoid arthritis cases and non-ra controls in the electronic medical records. *Arthritis and rheumatism* **65** 571.

LIN, D., OAKES, D. and YING, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85** 289–298.

MASSART, P. (1989). Strong approximation for multivariate empirical and related processes, via kmt constructions. *The Annals of probability* 266–291.

OH, E. J., SHEPHERD, B. E., LUMLEY, T. and SHAW, P. A. (2018). Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with simex. *Statistics in medicine* **37** 1276–1289.

PAN, W. (1999). Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *Journal of Computational and Graphical Statistics* **8** 109–120.

ROSALES, R., KRISHNAMURTHY, P. and RAO, R. B. (2007). Semi-supervised active learning for modeling medical concepts from free text. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*. IEEE.

ROSSINI, A. and TSIATIS, A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association* **91** 713–721.

RUBIN, D. B. (1987). Multiple imputation for nonresponse in surveys.

SEAMAN, S. R. and WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* **22** 278–295.

SHERMAN, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society* 123–137.

SKINNER, C. J. and HUMPHREYS, K. (1999). Weibull regression for lifetimes measured with error. *Lifetime data analysis* **5** 23–37.

SOKOLOVSKA, N., CAPPÉ, O. and YVON, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th international conference on Machine learning*. ACM.

SUN, J. and SUN, L. (2005). Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics* **33** 85–96.

TIAN, L. and CAI, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93** 329–342.

TU, W., XU, G. and DU, S. (2015). Structure and content components of self-management interventions that improve health-related quality of life in people with inflammatory bowel disease: a systematic review, meta-analysis and meta-regression. *Journal of clinical nursing* **24** 2695–2709.

TUSNÁDY, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica* **8** 53–55.

VAN DER LAAN, M. J. and ROBINS, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association* **93** 693–701.

WANG, Z., SHAH, A. D., TATE, A. R., DENAXAS, S., SHAWE-TAYLOR, J. and HEMINGWAY, H. (2012). Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* **7** e30412.