Semi-supervised estimation of covariance with application to phenome-wide association studies with electronic medical records data

Stephanie F Chan, Boris P. Hejblum, Abhishek Chakrabortty, Tianxi Cai

Semi-Supervised Estimation of Covariance with Application to Phenome-wide Association Studies with Electronic Medical Records Data

Stephanie F. Chan[1*], Boris P. Hejblum[1],
Abhishek Chakrabortty[2], and Tianxi Cai[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2]Department of Statistics, The Wharton School, University of Pennsylvania
[*]corresponding author

Abstract

Electronic medical records (EMRs) data are valuable resources for discovery research. They contain detailed phenotypic information on individual patients, opening opportunities 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 this paper, we focus on the setting 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.

Keywords: Electronic medical records data, model mis-specification, phenome-wide association studies, robustness, semi-supervised estimation

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) [1] 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 [2, 3].

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 [4, 5]. 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 [6]. This makes multiple 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 [7, 8]; however, most of these methods also focus on classification of disease status, rather than on estimation or testing [9, 10]. 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.* [11] 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 [12] extend Sokolovska *et al.*'s [11] 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 2, we formulate a semi-supervised estimator for this covariance and devise a method to calculate its standard error. In Section 3, we perform a simulation study to explore our methods and show the results of the simulations, and in section 4, we apply our method to an example dataset.

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.

2.1 Estimation

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

$$\theta_0 = \mathrm{cov}(Y_i, G_i) = E(r_i)$$

where $r_i = (Y_i - \mu_y)(G_i - \mu_g)$, $\mu_y = E(Y_i)$ 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

$$E(Y_i - \mu_y \mid \mathbf{S}_i, G_i) = \boldsymbol{\beta}^{\intercal}\mathbf{W}_i, \tag{1}$$

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), $\widehat{\boldsymbol{\beta}}$ 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) 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) is potentially mis-specified, we show in Appendix A 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

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

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

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

$K_h(x) = h^{-1}K(x/h)$, $K(\cdot)$ is a smooth kernel density function, $h = O(n^{-\nu})$ is the bandwidth with $\nu \in (1/4, 1/2)$. Although the estimation of $\widehat{\theta}_{\text{SSL}}$ uses G_i in the imputation step, we show in Appendix A that the design of our imputation guarantees the consistency of $\widehat{\theta}_{\text{SSL}}$ for θ_0 regardless of the adequacy of the imputation model. The inclusion of G_i in the imputation can in fact be viewed as a calibration step to ensure that any covariance between the imputed outcome and G is reflecting the covariance between Y and G. When $\theta_0 = 0$, $\widehat{\theta}_{\text{SSL}}$ also fluctuates around 0 as also confirmed by simulation results. 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

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

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.

2.2 Inference

We show in Appendix B 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 $E(Y_i - \mu_g \mid \mathbf{S}_i, G_i) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{W}_i$ 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 $n^{-1}\sum_{i=1}^{n}\{\widehat{r}_i - \widehat{m}(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\widehat{\mathbf{X}}_i,\widehat{\boldsymbol{\beta}})\}^2$.

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 3, 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, 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 $n = 200$.

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,

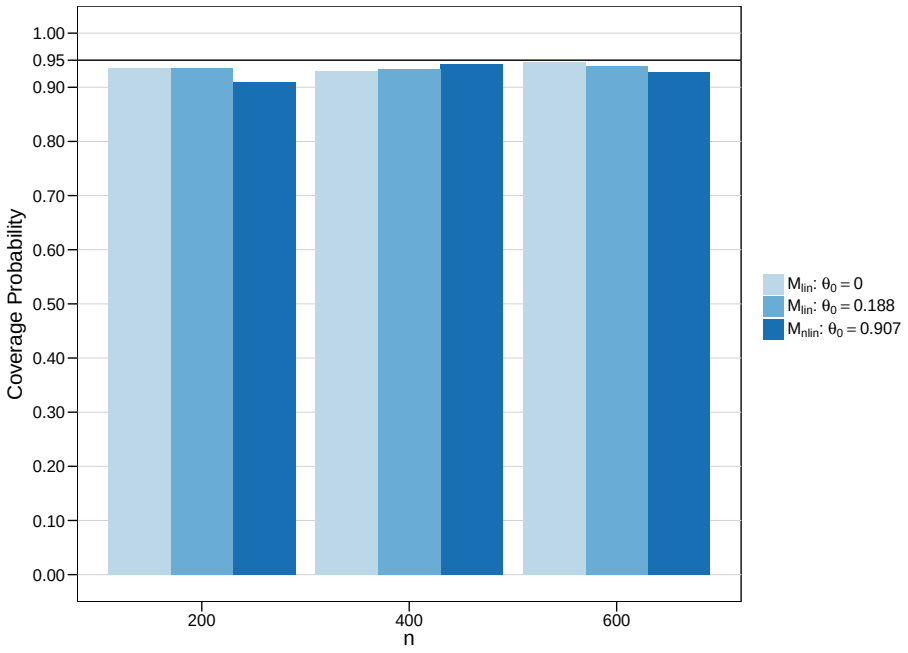


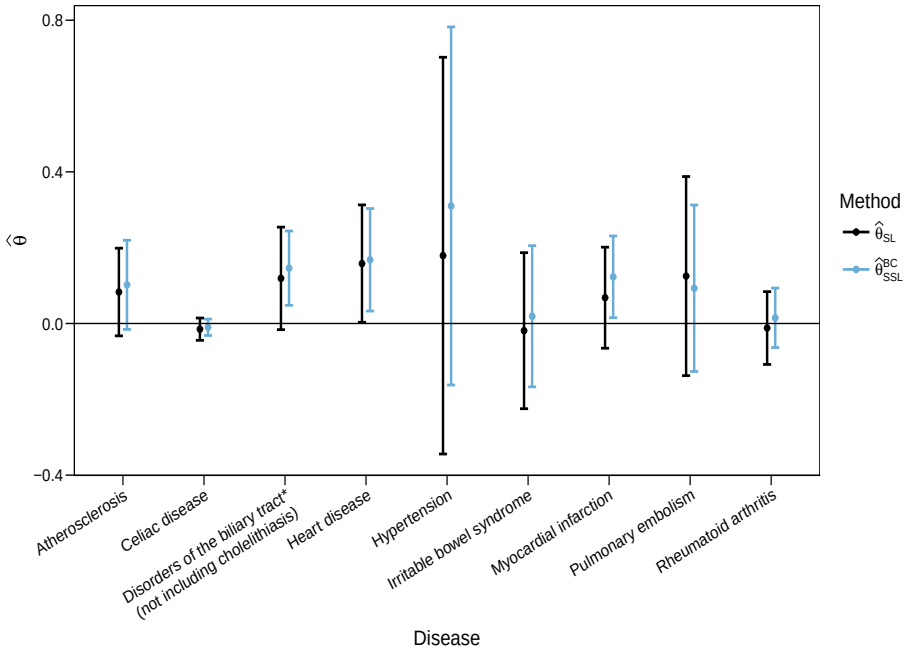
Figure 1: Coverage probabilities of the 95% CIs for $\widehat{\theta}^{\mathrm{BC}}_{\mathrm{SSL}}$ under various simulation settings

		$\mathscr{M}_{\text{lin}}$: $\boldsymbol{\theta_0 = 0}$			$\mathscr{M}_{\text{lin}}$: $\boldsymbol{\theta_0 = 0.188}$			$\mathscr{M}_{\text{nlin}}$: $\boldsymbol{\theta_0 = 0.907}$		
$\boldsymbol{n}$		**200**	**400**	**600**	**200**	**400**	**600**	**200**	**400**	**600**
$\widehat{\boldsymbol{\theta}}_{\boldsymbol{SL}}$	**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
$\widehat{\boldsymbol{\theta}}^{\boldsymbol{par}}_{\boldsymbol{SSL}}$	**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
$\widehat{\boldsymbol{\theta}}^{\boldsymbol{BC}}_{\boldsymbol{SSL}}$	**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

Table 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 $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$.

which causes inflammation and ulcers in the colon and rectum [13]. 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 [14]. The goal of our analysis is to examine whether inflammation (quantified by CRP levels) is related to comorbidities for IBD patients using a de-identified EMR crimson cohort of 2,048 patients from the Massachusetts General Hospital and Brigham and Women's Hospital of the 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 [15]. The longitudinally collected EMR data of these patients, between late 1990's to 2014, were available for analysis. Out of the 11,001 patients, 2,048 contributed blood for research and we only consider the blood cohort as the full cohort due to the discrepancy between patients who contributed blood versus those who did not.

To quantify the current level of inflammation, 97 patients were randomly selected from the IBD crimson cohort to have their CRP measured in 2015. 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 $x \to \log(x+1)$ 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 2. The results suggest that the supervised and semi-supervised estimates are reasonably consistent with each other in value, while the 95% CIs



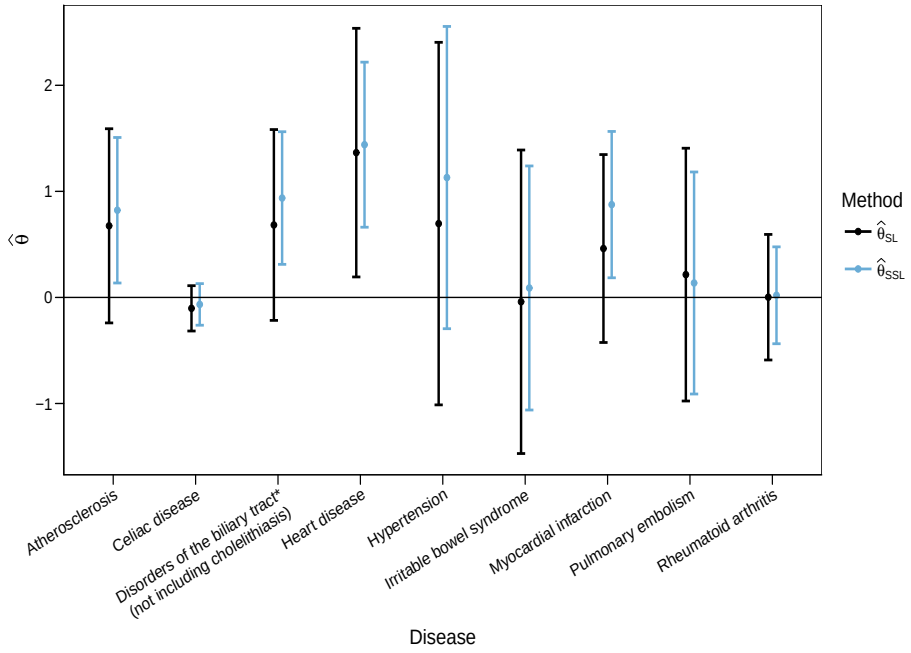
*This corresponds to PheWAS code 576, as described in Denny *et al.* [1]

Figure 2: Supervised and Semi-supervised estimates of the covariance for select PheWAS codes, along with 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.

The above analysis uses a log-transformed phecode count as the outcome, which inherently depends on the patient's healthcare utilization. As a sensitivity analysis, we also account for the healthcare utilization using the total number of ICD9 codes, H, and defined a normalized disease phenotype as $x \to \text{logit}(\frac{x+1}{H+1})$. As shown in Figure 3, we observe similar patterns for disease phenotypes that are associated with CRP. A Z-test based on $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$ would reject the null hypothesis for myocardial infarction and disorders of the biliary tract, where a Z-test based on $\widehat{\theta}_{\text{SL}}$ would not. Additionally, we see that the null hypothesis would be rejected for atherosclerosis using $\widehat{\theta}^{\text{BC}}_{\text{SSL}}$, but not for $\widehat{\theta}_{\text{SL}}$.

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



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

Figure 3: Supervised and Semi-supervised estimates of the covariance for select PheWAS codes, along with the 95% CIs, with normalization for health care utilization

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 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.

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 $\bar{\boldsymbol{\beta}}$ is 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

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

Appendix A

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

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

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 $G_i - \mu_g = (\mathbb{C}_{WW}^{-1}\mathbb{C}_{WG})^{\intercal}\mathbf{W}_i$ and $(\mathbb{C}_{WW}^{-1}\mathbb{C}_{WG})^{\intercal}\boldsymbol{\mu}_W = 0$. 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 $(\sigma^{\text{par}}_{\text{SSL}})^2 = E\{(r_i - R_i)^2\}$.

Appendix B

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

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

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

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

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

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 (2), the smoothness of $\mu_g(x,\boldsymbol{\beta})$ and the root-n convergence of $\widehat{\boldsymbol{\beta}}$ that

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

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

$$\sup_{x,\boldsymbol{\beta}} \left| \frac{\partial \widetilde{m}(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \frac{\partial m(x,\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right| = o_p(1). \tag{3}$$

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

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

where $\widehat{\mathcal{P}}_{\boldsymbol{\beta}}(s,r) = n^{-1}\sum_{i=1}^n I(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i \le s, r_i \le r)$ and $\mathcal{P}_{\boldsymbol{\beta}}(s,r) = P(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i \le s, r_i \le r)$. From the strong approximation result of Tusnády [16], 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 $\epsilon > 0$ [17]. Therefore, we have

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

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

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

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 almost surely:

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

The existence of the Gaussian process is ensured by the results of Massart [18]. 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 (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 (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 $\mathscr{I}_1 = (n/N)^{1/2}\int m(x,\bar{\boldsymbol{\beta}}) d\widehat{\mathcal{G}}(x,\bar{\boldsymbol{\beta}})$,

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

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

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 [19]. We then write

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

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

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

where $\widetilde{\mathcal{P}}(x, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}_i)$. 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}.

References

[1] Denny JC, Ritchie MD, Basford MA et al. Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010; 26(9): 1205–1210.

[2] Liao KP, Cai T, Gainer V et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research* 2010; 62(8): 1120–1127.

[3] Liao KP, Kurreeman F, Li G et al. Autoantibodies, autoimmune risk alleles and clinical associations in rheumatoid arthritis cases and non-ra controls in the electronic medical records. *Arthritis and rheumatism* 2013; 65(3): 571.

[4] Rubin DB. Multiple imputation for nonresponse in surveys, 1987.

[5] Seaman SR and White IR. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 2013; 22(3): 278–295.

[6] Kenward MG and Carpenter J. Multiple imputation: current perspectives. *Statistical methods in medical research* 2007; 16(3): 199–218.

[7] Rosales R, Krishnamurthy P and Rao RB. Semi-supervised active learning for modeling medical concepts from free text. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*. IEEE, pp. 530–536.

[8] Kim J and Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association* 2013; 20(4): 613–618.

[9] Dligach D, Miller T and Savova GK. Semi-supervised learning for phenotyping tasks. In *AMIA Annual Symposium Proceedings*, volume 2015. American Medical Informatics Association, p. 502.

[10] Wang Z, Shah AD, Tate AR et al. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 2012; 7(1): e30412.

[11] Sokolovska N, Cappé O and Yvon F. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 984–991.

[12] Kawakita M and Kanamori T. Semi-supervised learning with density-ratio estimation. *Machine learning* 2013; 91(2): 189–209.

[13] Tu W, Xu G and Du S. 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* 2015; 24(19-20): 2695–2709.

[14] Gabay C and Kushner I. Acute-phase proteins and other systemic responses to inflammation. *New England Journal of Medicine* 1999; 340(6): 448–454.

[15] Ananthakrishnan A, Cai T, Savova G et al. 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* 2013; 19(7): 1411–1420. DOI:10.1097/MIB.0b013e31828133fd.

[16] Tusnády G. A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica* 1977; 8(1): 53–55.

[17] Bickel PJ and Rosenblatt M. On some global measures of the deviations of density function estimates. *The Annals of Statistics* 1973; : 1071–1095.

[18] Massart P. Strong approximation for multivariate empirical and related processes, via kmt constructions. *The Annals of probability* 1989; : 266–291.

[19] Chakrabortty A and Cai T. Efficient and adaptive linear regression in semi-supervised settings. *arXiv:170104889* 2017; Retrieved from `https://arxiv.org/pdf/1701.04889.pdf`, `1701.04889`.