Risk prediction with imperfect survival outcome information from electronic health records

Jue Hou | Stephanie F. Chan | Xuan Wang | Tianxi Cai

Abstract

Readily available proxies for the time of disease onset such as the time of the first diagnostic code can lead to substantial risk prediction error if performing analyses based on poor proxies. Due to the lack of detailed documentation and labor intensiveness of manual annotation, it is often only feasible to ascertain for a small subset the current status of the disease by a follow-up time rather than the exact time. In this paper, we aim to develop risk prediction models for the onset time efficiently leveraging both a small number of labels on the current status and a large number of unlabeled observations on imperfect proxies. Under a semiparametric transformation model for onset and a highly flexible measurement error model for proxy onset time, we propose the semisupervised risk prediction method by combining information from proxies and limited labels efficiently. From an initially estimator solely based on the labeled subset, we perform a one-step correction with the full data augmenting against a mean zero rank correlation score derived from the proxies. We establish the consistency and asymptotic normality of the proposed semisupervised estimator and provide a resampling procedure for interval estimation. Simulation studies demonstrate that the proposed estimator performs well in a finite sample. We illustrate the proposed estimator by developing a genetic risk prediction model for obesity using data from Mass General Brigham Healthcare Biobank.

KEYWORDS

current status data, measurement error, risk prediction, semisupervised learning

1 | INTRODUCTION

Electronic health records (EHRs), containing detailed medical history of individuals in the health care system, hold immense potential for translational research (Jensen et al., 2012). In recent years, EHR data have been increasingly explored for developing risk prediction models to assist in clinical decision making (Eapen et al., 2013; Calvert et al., 2016; Jin et al., 2018, for instance). The longitudinal EHR data contain information on the occurrence time of clinical events which can be used as outcomes for risk prediction modeling. Rich clinical features, including lab measurements, medication prescriptions, and comorbidities, can be extracted as risk factors. Such clinical information can be effectively extracted from either codified data such as billing codes and procedure codes and from free-text clinical notes via natural language processing. At research institutions, EHR data have also been linked with biobanks where genetic information can be included.
in addition to clinical features to further improve risk prediction.

Although longitudinal EHR data are of great value for risk prediction modeling, precise information on clinical event time of interest, $T$, are not readily available. Timing and number of diagnostic codes and mentions of the disease in the clinical notes can serve as proxies of the true event time and status, but they are often not highly accurate. For example, at Mass General Brigham healthcare, having at least one diagnostic code of obesity only attained a sensitivity of 73% and positive predictive value of 81%; while having at least one mention of obesity in notes attained a higher sensitivity of 92% but a lower positive predictive value of 68%. It is even more challenging to approximate event time. For lung cancer recurrence, time of chemotherapy and radiation therapy initiations can only predict the true recurrence times up to 5.9 and 6.7 months on average (Uno et al., 2018). Directly using proxy event times $\mathcal{F}$ to replace the true $T$ for risk modeling can lead to substantial bias due to the measurement error in $\mathcal{F}$. On the other hand, extracting event time via manual annotation is too resource consuming for large-scale research. Furthermore, the exact timing of the event may not be precisely documented in the record, and hence it is often only feasible to annotate the current status of the event $\delta = I(T \leq C)$, that is whether the event has occurred by the end of the follow-up $C$. In this paper, we aim to develop an efficient EHR-based risk prediction procedure under a semisupervised setting with data from a small set of gold standard labels on $\delta$ and a large set of unlabeled data with mismeasured event times $\mathcal{F}$.

With the current status survival data, regression methods have been proposed for commonly used survival models including the proportional hazards (PH) and proportional odds (PO) models and the semiparametric transformation model (Huang, 1996; Rossini and Tsiatis, 1996; Carroll et al., 1997; Huang and Rossini, 1997; Van Der Laan and Robins, 1998; Sun and Sun, 2005). Estimation procedures have also been proposed for other models including the additive hazard model and accelerated failure time model (Lin et al., 1998; Betensky et al., 2001; Tian and Cai, 2006; Chen and Sun, 2010). Developing risk prediction models with mismeasured survival outcomes is a challenging problem, and few methods currently exist, in part due to the additional complexity induced by censoring. Oh et al. (2018) introduced a bias correction procedure for hazard ratio estimates in the PH model. Using a validation data where both the true and error-prone survival times are both available, Braun et al. (2018) proposed a nonparametric bias correction procedure for models using $\mathcal{F}$ as predictors. Gravel et al. (2018) proposed the model-based estimation accounting for temporally induced errors and misclassified event errors in time-to-event outcomes. Recently, Oh et al. (2021) developed a ranking and regression calibration method for data with measurement errors in both covariates and time-to-event outcomes.

No semisupervised survival regression method currently exists for the EHR setting where the true survival times are only observed up to the binary current status in a relatively small $n$ labeled set, and the mismeasured event times $\mathcal{F}$ are observed for all $N$ patients with $N$ potentially much larger than $n$. In addition, even without censoring, existing methods incorporating mismeasured outcomes largely require restrictive parametric measurement error model assumptions. To fill this gap, we propose an efficient and robust semisupervised learning (SSL) procedure for the semiparametric transformation model under flexible measurement error models without distributional assumptions. Our SSL estimation starts with an initial supervised estimator based on the current status data via solving a system of kernel-smoothed estimating equations and then constructs an augmented estimator by optimally combining the initial estimator with information from $\mathcal{F}$ derived from a rank estimation procedure. We analyze the limiting distribution of the SSL estimator and develop an inference method accordingly. When $N \gg n$, we discover the super-efficiency of the semisupervised estimator for zero coefficients, which requires special arrangements for inference.

The rest of the paper is organized as follows. We detail our SSL procedures in Section 2. In Section 3, we present results from simulation studies to examine the finite sample behavior of the SSL estimator and compare its efficiency to the initially supervised estimator. In Section 4, we apply our methods to develop an age-specific risk prediction model for obesity based on demographic and genetic information using EHR data from the Partners Biobank. Concluding remarks are given in Section 5. Technical details are presented in the Supporting Information.

## 2 | METHODS

Suppose there are a total of $N$ subjects in the EHR cohort, and a subset of $n$ subjects is randomly sampled into the labeled set to have their event status annotated via manual chart review. Let $T_i, C_i$, and $Z_i$, respectively, denote the true event time, follow-up time, and $p$-dimensional baseline covariates for the $i$th subject. The true event time $T_i$ is not observable, but $\delta_i = I(T_i \leq C_i)$ is observed for those in the labeled set. In addition, there are $K$ surrogate event times, $\mathcal{F}_i = (\mathcal{F}_1, ..., \mathcal{F}_K)^T$, that can be viewed as proxies of $T_i$. However, $\mathcal{F}$ is subject to right censoring since patients are only followed in the EHR up to time $\mathcal{E}_i$, the follow-up time for surrogate events. In many cases, $\mathcal{E}_i$ is the same as $C_i$, but there are situations $\mathcal{E}_i > C_i$ if annotation happens in the past before the observation of $\mathcal{F}_i$ ends. Thus for $\mathcal{F}$, we only
observe $\mathbf{X} = (X_1, \ldots, X_K)^T$ and $\Delta = (\Delta_1, \ldots, \Delta_K)^T$, where $\mathcal{X} = \mathcal{F} \wedge \mathcal{C}$ and $\Delta = I(\mathcal{F} \leq \mathcal{C})$. The full underlying data of the EHR cohort consist of $\{(T_i, C_i, Z_i^T, \mathcal{F}_i, \mathcal{C}_i)\}^n_{i=1}$, while the observed data consist of the labeled data $\mathcal{L} = \{(\delta_i, C_i, Z_i^T, \mathcal{F}_i, \mathcal{C}_i)\}^n_{i=1}$ and the unlabeled data $\mathcal{U} = \{(C_i, Z_i^T, \mathcal{F}_i, \mathcal{C}_i)\}^n_{i=n+1}$. Without loss of generality, we assume that both $\mathcal{C}$ and $\mathcal{F}$ have continuous distributions with twice continuously differentiable densities and finite support $(0, C_{\max})$. We predict $T_i$ with $Z_i$ via the semiparametric transformation model, which includes PH and PO models as special cases:

$$P(T_i \leq t \mid Z_i) = g(\phi_0(t) + \beta_0^T Z_i)$$

where $g(\cdot)$ is a known smooth probability distribution function, $\phi_0(t)$ is an unspecified smooth increasing function, and $\beta_0$ is the unknown regression coefficient. Under the semiparametric transformation model (1), we have

$$E(\Delta_i \mid C_i, Z_i) = P(T_i \leq C_i \mid C_i, Z_i) = g(\phi(C_i) + \beta_0^T Z_i)$$

from which we later derive our estimating equations for the initial estimator.

For each of the mismeasured survival outcome $\mathcal{F}_k$, we assume that

$$H_k(\mathcal{F}_k) = H_k(T) + \epsilon_{ki}, \quad k = 1, \ldots, K$$

where and $H_k(\cdot)$ is an unknown smooth transformation function and $\epsilon_{ki}$ is independent of $(T_i, Z_i^T, \mathcal{C}_i)^T$ with a completely unspecified distribution. We also leave the within-subject correlation structure among $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^T$ unspecified. Leaving both $\phi(\cdot)$ and the distribution of $\epsilon_{ki}$ unspecified allow a wide range of measurement error models, including both additive and multiplicative measurement errors. With the finite observation window ending at $C_{\max}$, the truncated mismeasured survival outcome $\mathcal{F}_{ki}^* = \mathcal{F}_{ki} \wedge C_{\max}$ is of greater practical interest, as no event beyond $C_{\max}$ is observable. We show in Appendix B of the Supporting Information that $\mathcal{F}_{ki}^*$ follows a generalized regression model as defined in Han (1987). It follows from Han (1987) and Sherman (1993) that the direction of the true parameter $\beta_0$, $B_0 = \frac{\beta}{\|\beta\|_2}$, is identified by the maximizer of the population rank correlation:

$$B_0 = \arg\max_{B \in \Omega_p} Q_k(B)$$

where $Q_k(B) = P\left(B^T Z_i \geq B^T Z_j \mid \mathcal{F}_{ki} < \mathcal{F}_{kj}^*\right)$

$$\Omega_p = \{B \in \mathbb{R}^p : \|B\|_2 = 1\}$$

and $\|\cdot\|_2$ denotes the $L_2$ norm. From the first-order Karush–Kuhn–Tucker condition,

$$S_k(B_0) = 0,$$

where $S_k(B) = \mathbb{E}(Z_i - Z_j \mid B^T Z_i = B^T Z_j, \mathcal{F}_{ki} < \mathcal{F}_{kj}^*)$.

Based on (5), we derive our score for the full cohort, through which we update the initial estimator with the information from the mismeasured survival outcomes.

### 2.1 Estimation

To estimate $\beta_0$, we first obtain an initial estimator using the labeled data $\mathcal{L}$ according to the moment equation (2). Specifically, we adopt the kernel smoothed estimation procedure and estimate $\{\phi(\cdot), \beta_0\}$ by solving

$$\sum_{i=1}^n Z_i[\delta_i - g(\phi(C_i) + \beta_0^T Z_i)] = 0$$

$$\sum_{i=1}^n k_h(C_i - t)[\delta_i - g(\phi(t) + \beta_0^T Z_i)] = 0$$

where $k_h(t) = h^{-1}k(t/h)$, $k(\cdot)$ is a smooth symmetric probability density function, and $h = n^{-\nu}$ with $\nu \in (1/4, 1/2)$. Semiparametric estimation like (6) has been well studied in survival analysis. With smooth $g$ and $\phi$, $\sqrt{n}$-consistent and asymptotically normal estimation of parametric component $\beta$ is achieved by choosing a small bandwidth $h$ (Carroll et al., 1997). Let $\hat{\beta}_0$ denote the resulting estimate for $\beta_0$. We present the asymptotic properties of $\hat{\beta}_0$ in Appendix B of the Supporting Information.

To improve the efficiency of $\hat{\beta}_0$, we can leverage the mismeasured survival outcomes of the full cohort, we consider the one-step correction maximizing the rank correlation $Q_k(B)$, or equivalently solving its score equation (5). However, the usual one-step correction is not applicable due to difficulty in estimation of the derivatives of $Q_k(B)$. If $Z$ is continuous, there is no empirical estimator for the score $S_k(B)$ because $B^T Z_i = B^T Z_j$ occurs with probability zero. To this end, we allowed some tolerance in $B^T Z_i = B^T Z_j$ via kernel smoothing

$$\hat{S}_k(B) = \frac{\sum_{i=1}^N \sum_{j=1}^N (Z_i - Z_j)k_h(B^T Z_i - B^T Z_j) I(X_{ki} < X_{kj}) \Delta_{ki}}{\sum_{i=1}^N \sum_{j=1}^N I(X_{ki} < X_{kj}) \Delta_{ki} \hat{g}(X_{ki})^2}$$

for $k = 1, \ldots, K$.

where $h' = N^{-\nu'}$ for $\nu' \in (1/4, 1/2)$ and $\hat{g}(t)$ is the empirical estimate of $G(t) = P(\mathcal{C} \geq t)$. Here, we accounted for the censoring through the inverse probability of censoring
weighting
\[
\mathbb{E}\left\{ \frac{I(\mathcal{X}_{k_i} < \mathcal{X}_{k_j}) \Delta_{k_i}}{G(\mathcal{X}_{k_i})^2} \mid \mathcal{T}_{k_i}, \mathcal{T}_{k_j} \right\} = I(\mathcal{T}_{k_i}^* < \mathcal{T}_{k_j}^*). \tag{8}
\]

The denominator of \(\hat{S}_k(B)\) normalized the total weights to promote finite sample stability. We accounted for censoring through the inverse probability of censoring weighting. To combine multiple surrogate events, we stack the scores evaluated at \(\hat{B}_s\) across all \(k = 1, \ldots, K\) as \(\hat{S}(\hat{B}_s)\), where \(\hat{S}(B) = [\hat{S}_1(B)^\top, \ldots, \hat{S}_K(B)^\top]^\top\). Since \(\hat{S}_k(B)\) is a sample score and \(\hat{B}_s\) is a regular \(n\)-estimator, the plug score \(\hat{S}_k(\hat{B}_s)\) with \(\hat{B}_s = \hat{B}_s/\|\hat{B}_s\|_2\) should also be mean zero and asymptotically normal at \(n^{-1/2}\) order. We derive in Appendix D of the Supporting Information the asymptotic distribution of \(\hat{S}(\hat{B}_s)\). There, we also show that \(\sqrt{n}\hat{S}(\hat{B}_s)\) has a rank-deficient asymptotic variance matrix because its asymptotic limit \(S(B_0)\), the derivatives of scale invariant \(Q_k(B)\) at \(B = B_0\), is collinear.

### 2.1.1 Optimal combination

Our final estimator will be the optimal combination of \(\hat{B}_s\) with \(\hat{S}(\hat{B}_s)\) that minimizes the asymptotic variance,
\[
\hat{B}_s - W^\top \hat{S}(\hat{B}_s), \quad \hat{B}_s = \hat{B}_s/\|\hat{B}_s\|_2. \tag{9}
\]

When \(K = 1\) and \(W\) is the second derivative of \(Q_1(B_0)\), the combination is the usual one-step correction. Hence, the optimal combination can be seen as the generalization of the usual one-step estimator that aggregates multiple scores. The optimality is achieved when \(W^\top \hat{S}(\hat{B}_s)\) is the projection of \(\hat{B}_s\) onto the space spanned by \(\hat{S}(\hat{B}_s)\),
\[
W_{opt} = \left[ \text{Var}[\sqrt{n}\hat{S}(\hat{B}_s)] + \text{Cov}[\sqrt{n}\hat{S}(\hat{B}_s), \sqrt{n}\hat{B}_s] \right]^{-1} \times \text{Cov}[\sqrt{n}\hat{S}(\hat{B}_s), \sqrt{n}\hat{B}_s], \tag{10}
\]
where \(\text{Var}\) and \(\text{Cov}\) denote asymptotic variance and covariance and \(A^{-}\) denotes the generalized inverse. Here we used the generalized inverse because \(\text{Var}[\sqrt{n}\hat{S}(\hat{B}_s)]\) is not invertible. Each row of \(W_{opt}\) is the linear regression coefficient
\[
w_j = \arg\min_{w \in \mathbb{R}^{K_P}} \text{Var}[\sqrt{n}(\hat{B}_{s,j} - \hat{B}_{0,j}) - w^\top \sqrt{n}\hat{S}(\hat{B}_s)]. \tag{11}
\]

Since the generalized inverse is not unique, the solution for \(W_{opt}\) along with \(w_j\) is not unique. We choose the \(W_{opt}(P)\) based on the spectral decomposition
\[
\text{Var}[\sqrt{n}\hat{S}(\hat{B}_s)] = \Sigma_S = P^\top DP, \quad \text{Diag}(\sigma_1^2, \ldots, \sigma_r^2),
\]
\[
r = \text{rank}(\Sigma_S) \tag{12}
\]
whose details are given in Appendix E of the Supporting Information.

Inspired by the Jin et al. (2001), we propose the perturbation resampling estimator for \(W_{opt}\), which estimates the asymptotic moments \(\text{Cov}(\sqrt{n}\hat{S}(\hat{B}_s), \sqrt{n}\hat{B}_s)\) and \(\text{Var}(\sqrt{n}\hat{S}(\hat{B}_s))\) by their empirical means in the perturbation sample. Specifically, for \(b = 1, \ldots, B\) with \(B \gg n\), we generate a vector of \(N\) independent and identically distributed random variables \(\{Y_i^{(b)} \mid i = 1, \ldots, N\}\) with \(E(Y_i) = \text{Var}(Y_i) = 1\), for example, \(Y_i/4 \sim \text{Beta}(0.5, 1.5)\), and obtain \(\hat{B}_s^{(b)}\) as the solution to
\[
\sum_{i=1}^n K_i(C_i - t) \delta_i - g(\phi_0(t) + \beta^\top Z_i) Y_i^{(b)} = 0,
\]
\[
\sum_{i=1}^n Z_i \delta_i - g(\phi_0(C_i) + \beta^\top Z_i) Y_i^{(b)} = 0. \tag{13}
\]

Then we obtain the perturbed counterpart of \(\hat{S}_k(B)\) as
\[
\hat{S}_k^{(b)}(B) = \frac{\sum_{1 \leq i < j \leq N} Y_i^{(b)} Y_j^{(b)} (Z_i - Z_j) k_i n_{ij}}{\sum_{1 \leq i < j \leq N} Y_i^{(b)} Y_j^{(b)}} \frac{(\mathcal{X}_{i} < \mathcal{X}_{j}) \Delta_{i,j}}{\text{Var}(\mathcal{X}_{i})^2} + \frac{\sum_{1 \leq i < j \leq N} Y_i^{(b)} Y_j^{(b)} (Z_i - Z_j) k_i n_{ij}}{\sum_{1 \leq i < j \leq N} Y_i^{(b)} Y_j^{(b)}} \frac{(\mathcal{X}_{i} < \mathcal{X}_{j}) \Delta_{i,j}}{\text{Var}(\mathcal{X}_{i})^2} \tag{14}
\]
for \(k = 1, \ldots, K\),

where \(\hat{\mathcal{G}}(b)(t) = \{(\sum_{i=1}^N I(\mathcal{X}_{i} \geq t) Y_i^{(b)}) / (\sum_{i=1}^N Y_i^{(b)})\} \times \hat{S}_k^{(b)}(B)\). Subsequently, we obtain \(\hat{B}_s = \hat{B}_s^{(b)} / \|\hat{B}_s^{(b)}\|_2\) and \(\hat{S}(\hat{B}_s)\) can be approximated by \(\hat{B}_s^{(b)} - \hat{B}_s\) and \(\hat{S}^{(b)}(\hat{B}_s)\). By (11), we can estimate \(w_j\) by regressing \(\sqrt{n}(\hat{B}_s^{(b)} - \hat{B}_s)\) onto \(\sqrt{n}S^{(b)}(\hat{B}_s)\). Recall that the asymptotic variance of \(\sqrt{n}S^{(b)}(\hat{B}_s)\) is rank deficient, which implies that \(\sqrt{n}S^{(b)}(\hat{B}_s)\) becomes collinear with large \(n\). We performed the principal component analysis (principal component analysis) to handle the collinearity of \(\sqrt{n}S^{(b)}(\hat{B}_s)\). Let \(\hat{P}\) be the loading matrix and \(\hat{\sigma}_1, \ldots, \hat{\sigma}_{K_P}\) be the singular values from principal component analysis. We consider the adaptive ridge regression
\[
\hat{w}_j = \arg\min_{w \in \mathbb{R}^r} \frac{1}{B} \sum_{b=1}^B \left\{ \left( \frac{(\hat{B}_s^{(b)} - \hat{B}_s) - w^\top \hat{S}^{(b)}(\hat{B}_s)}{\hat{P}^\top \hat{S}^{(b)}(\hat{B}_s)} \right)^2 \right\} + \lambda \sum_{h=1}^{K_P} \frac{w_h^2}{\hat{\sigma}_h^{2a}}, \tag{15}
\]
for some \(a > 3/2\) and \(n^{1/4-\alpha/2} \ll \lambda \ll n^{-1/2}\). The adaptive penalty in (15) shrinks the coefficients for loadings associated with small singular values to zero but keeps
intact coefficients for loadings associated with nonzero singular values. We assemble the estimated coefficients \(\tilde{w}_1, \ldots, \tilde{w}_p\) to construct estimated projection matrix

\[
\tilde{W} = \hat{\beta}^T(\tilde{w}_1, \ldots, \tilde{w}_p) = \mathbb{W}_{\text{opt}}(P) + O_p(n^{-1/2}).
\]  

(16)

We show that \(\tilde{W}\) is \(\sqrt{n}\)-consistent for \(\mathbb{W}_{\text{opt}}(P)\) in Appendix E of the Supporting Information.

We construct the semisupervised estimator by

\[
\hat{\beta}_{SSL} = \hat{\beta} - \tilde{W}^T S(\hat{\beta}).
\]

(17)

We show in Appendix F of the Supporting Information that \(\hat{\beta}_{SSL}\) is \(\sqrt{n}\)-consistent for \(\beta_0\) and asymptotically norm.

### 2.2 Inference

To quantify the estimation error, it is natural to consider the perturbed semisupervised estimator

\[
\hat{\beta}_{SSL}^{(b)} = \hat{\beta}^{(b)} - \tilde{W}^{(b)} S^{(b)}(\hat{\beta}^{(b)}), \quad b = 1, \ldots, B.
\]

By Jin et al. (2001), \(S^{(b)}(\hat{\beta}^{(b)})\) has the same asymptotic distribution of \(S(\hat{\beta})\). From (A.7), we have \(\tilde{S}^{(b)}(\hat{\beta}) = O_p(n^{-1/2})\). Since \(\tilde{S}^{(b)}(\hat{\beta})\) converges to zero, the variability from estimating \(\mathbb{W}_{\text{opt}}(P)\) by \(\tilde{W}\) is negligible on the asymptotic distribution of \(\tilde{W}\). Hence, \(\tilde{W}\) can be used directly in the perturbed semisupervised estimator without an extra layer of perturbation.

In the \(N \times n\) scenario, we can directly use \(v^T \hat{\beta}_{CI} = v^T \hat{\beta}_{SSL}\) and \(v^T \hat{\beta}_{CI}^{(b)} = v^T \hat{\beta}_{SSL}^{(b)}\) for interval estimation. We estimate the variance of \(v^T \hat{\beta}_{CI}\) by the empirical variance of \(v^T \hat{\beta}_{CI}\) :

\[
\var(v^T \hat{\beta}_{CI}) = \frac{1}{B} \sum_{b=1}^{B} \left( v^T \hat{\beta}_{CI}^{(b)} - v^T \hat{\beta}_{CI} \right)^2.
\]

(19)

We construct the \((1 - \alpha) \times 100\%\) confidence interval (CI) for \(v^T \beta_0\) by

\[
\left[ v^T \hat{\beta}_{CI} - Z_{\alpha/2} \sqrt{\var(v^T \hat{\beta}_{CI})}, v^T \hat{\beta}_{CI} + Z_{\alpha/2} \sqrt{\var(v^T \hat{\beta}_{CI})} \right],
\]

(20)

where \(Z_{\alpha/2}\) is the \(1 - \alpha/2\) quantile of standard Normal distribution.

In the \(N \gg n\) scenario, we discover that \(v^T \hat{\beta}_{SSL}\) is super-efficient for \(v^T \beta_0\) if \(v^T \beta_0 = 0\),

\[
v^T \hat{\beta}_{SSL} = O_p(n^{-1}).
\]

(21)

We give the details in Appendix F of the Supporting Information. The phenomenon agrees with the result of Cai and Cheng (2007) that the standardized coefficient \(B_0 = \beta_0/\|\beta_0\|_2\) can be estimated at \(N^{-1/2}\) rate based on rank correlation \(Q_k(B)\) alone. The parameter \(v^T \beta_0 = 0\) requires no information on the coefficient size \(\|\beta_0\|_2\) from labeled data, so it is possible to estimate it at a faster rate with large unlabeled data. A typical example is the zero element in \(\beta_0\) indexed by \(j \notin \text{supp}(\beta_0)\), which can be represented with \(j\)th natural basis \(e_j\),

\[
e_j^T \beta = \beta_{0,j} = 0, \quad e_j^T \hat{\beta}_{SSL} = \hat{\beta}_{SSL,j} = o_p(n^{-1/2}).
\]

(22)

Since \(\sqrt{n}(\hat{\beta}_{SSL,j} - 0)\) converges to zero asymptotically, it is unknown from the work of Jin et al. (2001) if the perturbation samples \(\sqrt{n}(\hat{\beta}_{SSL,j} - 0)\) may capture its asymptotic distribution. To provide a valid inference for \(\beta_{0,j}\), we perform the norm preserving soft-thresholding on \(\hat{\beta}_{SSL}\),

\[
\hat{\beta}_{soft} = \left( \hat{\beta}_{soft,1}, \ldots, \hat{\beta}_{soft,p} \right)^T, \quad \hat{\beta}_{soft,j} = \text{sign}(\hat{\beta}_{SSL,j}) \max \left\{ |\hat{\beta}_{SSL,j}| - \lambda_{soft}/|\hat{\beta}_{SSL,j}|, 0 \right\},
\]

(23)

\[
\hat{\beta}^{(b)} = \left( \hat{\beta}_{soft,1}^{(b)}, \ldots, \hat{\beta}_{soft,p}^{(b)} \right)^T, \quad \hat{\beta}_{soft,j}^{(b)} = \text{sign}(\hat{\beta}_{SSL,j}^{(b)}) \max \left\{ |\hat{\beta}_{SSL,j}^{(b)}| - \lambda_{soft}/|\hat{\beta}_{SSL,j}^{(b)}|, 0 \right\},
\]

(24)

\[
\hat{\beta}_{CI} = \hat{\beta}_{soft} \frac{\|\hat{\beta}_{SSL}\|_2}{\|\hat{\beta}_{soft}\|_2}, \quad \hat{\beta}_{CI}^{(b)} = \hat{\beta}_{soft}^{(b)} \frac{\|\hat{\beta}_{SSL}^{(b)}\|_2}{\|\hat{\beta}_{soft}^{(b)}\|_2}
\]

(25)

for \(n^{-1/2} \gg \lambda_{soft} \gg n^{-2}\). For nonzero coefficients \(\beta_{0,j} \neq 0\), the thresholding is much smaller than \(n^{-1/2}\) thus asymptotically negligible, and the CI

\[
\left[ \hat{\beta}_{CI,j} - Z_{\alpha/2} \sqrt{\text{Var}(\hat{\beta}_{CI,j})}, \hat{\beta}_{CI,j} + Z_{\alpha/2} \sqrt{\text{Var}(\hat{\beta}_{CI,j})} \right],
\]

\[
\text{Var}(\hat{\beta}_{CI,j}) = \frac{1}{B} \sum_{b=1}^{B} \left( \hat{\beta}_{CI,j}^{(b)} - \hat{\beta}_{CI,j} \right)^2
\]

(26)

must achieve asymptotically the nominal coverage. The CI for \(\beta_{0,j} = 0\), however, would have \(100\%\) coverage rate asymptotically as the \(\hat{\beta}_{CI,j}\) and \(\hat{\beta}_{0,j}^{(b)}\), \(b = 1, \ldots, B\) are all shrunk to zero with large probability. We recommend select \(\lambda_{soft}\) through a cross-validation scheme targeting the aggregated rank correlation of mismeasured survival times

\[
Q(\hat{\beta}) = \sum_{k=1}^{K} Q_k(\hat{\beta}) = \sum_{k=1}^{K} \sum_{l=1}^{L} I(\hat{\beta}^T X_k > \hat{\beta}^T X_l) (X_k < X_l) \Delta_{kl} \frac{\hat{\beta}^T X_k^2}{\hat{\beta}^T X_k^2}.
\]

(27)
3 | SIMULATIONS

We conducted simulation studies to assess the performance of our estimator in finite sample settings. Throughout, we set the labeled sample size as \( n = 500 \) and consider two sizes for total sample \( N = 500 \) and \( N = 10,000 \). The two total sample sizes characterize the \( n \approx N \) and \( n \ll N \) scenarios, respectively. We first generated a \( p = 10 \) dimensional \( Z \) from a zero-mean multivariate normal with unit variance and correlation 0.2. We set the true coefficients as

\[
\boldsymbol{\beta}_0 = (0.7, 0.7, 0.7, -0.5, -0.5, -0.5, 0.3, 0.3, 0.3, 0) \quad (28)
\]

and generated \( T_i \) from the PO model and PH model

- PO: \( P(T_i \leq t \mid Z_i) = \left[1 + \exp \left\{ \beta_0^T Z_i + 3 \log(t/4) \right\} \right]^{-1} \)
- PH: \( P(T_i \leq t \mid Z_i) = 1 - \exp \left\{ -\exp \left\{ \beta_0^T Z_i + 3 \log(t/4) \right\} \right\} \)

The censoring time \( C_i = \mathcal{C}_i \) was generated from Uniform(0,12). The event rate was 58% under the PO model and 53% under the PH model. We consider \( K = 2 \) surrogates generated from \( \log \mathcal{T}_{ki} = \log T_i + \epsilon_{ki} \), where \( \epsilon_{ki} \) is generated from a normal mixture \( \sim D_{ki}N(\mu_{ki}, \sigma_{ki}^2) + (1 - D_{ki})N(\mu_{ki}^+, \sigma_{ki}^+)^2 \) and \( D_{ki} \sim \text{Bernoulli}(0.5) \). The normal mixture measurement error distribution was chosen to allow for more heterogeneity in how their records appear in the EHR system. We considered two scenarios for the error distribution: (A) low measurement error with \((\mu_{1-}, \sigma_{1-}, \mu_{1+}, \sigma_{1+}) = (0, 0.5, 0.5, 0.15)\) and \((\mu_{2-}, \sigma_{2-}, \mu_{2+}, \sigma_{2+}) = (-0.25, 0.35, 0, 0.45)\); (B) high measurement error with \((\mu_{1-}, \sigma_{1-}, \mu_{1+}, \sigma_{1+}) = (1, 1.5, -0.5, 0.5)\) and \((\mu_{2-}, \sigma_{2-}, \mu_{2+}, \sigma_{2+}) = (0, 1.5, 0.5)\). We illustrate the sizes of measurement errors in Figure A1 in the Supporting Information. The measurement errors were quite large even for the setting (A) and very large for setting (B). The surrogate event rates were 49% for \( \mathcal{T}_1 \) and 61% for \( \mathcal{T}_2 \) under setting (A); 52% for \( \mathcal{T}_1 \) and 34% for \( \mathcal{T}_2 \) under setting (B).

The bandwidth \( h \) for estimating \( \hat{\beta}_S \) was chosen to be \( \hat{\sigma}(\sum_{i=1}^N \delta_i)^{-0.26} \), where \( \hat{\sigma} \) is the empirical standard deviation of \( C \). Here we chose a relatively conservative plug-in \( h \) with an effective sample size of \( \sum_{i=1}^N \delta_i \) to ensure the stability of the initial estimation with a relatively small even number in the labeled subset. The bandwidth \( h'_k \) for \( \hat{S}_k \) is chosen to be \( \hat{\sigma}(\sum_{i=1}^N \Delta_{ki})^{-0.3} \), where \( \hat{\sigma} \) is the empirical standard error of \( \hat{S}_k Z \). The bandwidth for rank correlation makes very little difference given the large sample size in the unlabeled data, and our chose of \( v' = -0.3 \) is an arbitrary choice from the permitted range. For each scenario, we summarize results using 500 datasets. For the standard error estimates via resampling, we use \( B = 1000 \) replications.

We report the comparative performance of \( \hat{\beta}_S, \hat{\beta}_{SSL} \), and two unsupervised estimators \( \hat{\beta}_{UL,1} \) and \( \hat{\beta}_{UL,1} \) obtained by using \((X_1, \Delta_1)\) and \((X_2, \Delta_2)\) as if they were the true survival outcome without measurement error procedures in Tables 1 and 2 under the PO and PH models. Both \( \hat{\beta}_S \) and \( \hat{\beta}_{SSL} \) have negligible biases across all settings. When \( N \approx n \), \( \hat{\beta}_{SSL} \) is more efficient than \( \hat{\beta}_S \) under the low measurement error model with relative efficiency ranging from 121% to 150% under the PO model and 121% to 130% under the PH model. The efficiency gain is minimal for the high measurement error setting when \( N \gg n \). \( \hat{\beta}_{SSL} \) is substantially more efficient than \( \hat{\beta}_S \) with relative efficiency ranging from 265% to 2357% under the PO model and from 206% to 2387% under the PH model for the low measurement error setting and from 247% to 894% under the PO model and from 195% to 714% under the PH model for the high measurement error setting under the two models. The efficiency gain is the highest for the zero coefficient as expected from our theoretical findings on the supersufficiency of \( \hat{\beta}_{SSL} \) on zero coefficients. These results also suggest that the semisupervised learning improves efficiency even with large measurement error in surrogates when the unlabeled data are large. On the other hand, the unsupervised estimators ignoring the measurement errors had substantial biased toward zero. Both \( \hat{\beta}_{UL,1} \) and \( \hat{\beta}_{UL,2} \) have very poor efficiency with large measurement error and were still much poorer than \( \hat{\beta}_{SSL} \) with small measurement error.

We report the performance of the proposed inference procedures with \( \hat{\beta}_{C1} \) in Tables 3 and 4. For individual components of the regression coefficients, the average estimated standard errors are close to the corresponding empirical standard errors and the empirical coverage levels of the 95% CIs are close to the nominal level for nonzero coefficients. For the zero-coefficient \( \hat{\beta}_{10} \), the empirical coverage probabilities are generally higher than the nominal level especially for large \( N \). This is expected due to the supersufficiency.

4 | APPLICATION TO DEVELOPING EHR-BASED OBESITY GENETIC RISK PREDICTION MODEL

As one of the most serious public health problems in the twenty-first century, obesity affects about 12% of adults globally (GBD 2015 Obesity Collaborators, 2017). Although there are lifestyle changes that can be made to prevent it, there is a strong genetic component to obesity, which is a risk factor for many other conditions such as cardiovascular disease and diabetes. We applied our proposed
TABLE 1  Bias (×100) of $\hat{\beta}_2$, semisupervised $\hat{\beta}_{SSL}$ and unsupervised $\hat{\beta}_{UL,1}$, $\hat{\beta}_{UL,2}$ as well as relative efficiency (RE) of $\hat{\beta}_{SSL}$, $\hat{\beta}_{UL,1}$ and $\hat{\beta}_{UL,2}$ compared to $\hat{\beta}_3$ under the PO model when the measurement errors in $T_1$ and $T_2$ are (A) moderate and (B) high

| $n = 500, N = 500$, (A) Small | $n = 500, N = 500$, (B) High | $n = 500, N = 10,000$, (A) Small | $n = 500, N = 10,000$, (B) High |
|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| $\beta_1 = 0.7$               | $\beta_1 = 0.7$               | $\beta_1 = 0.7$               | $\beta_1 = 0.7$               |
| $\beta_2 = 0.7$               | $\beta_2 = 0.7$               | $\beta_2 = 0.7$               | $\beta_2 = 0.7$               |
| $\beta_3 = 0.7$               | $\beta_3 = 0.7$               | $\beta_3 = 0.7$               | $\beta_3 = 0.7$               |
| $\beta_4 = 0.7$               | $\beta_4 = 0.7$               | $\beta_4 = 0.7$               | $\beta_4 = 0.7$               |
| $\beta_5 = 0.7$               | $\beta_5 = 0.7$               | $\beta_5 = 0.7$               | $\beta_5 = 0.7$               |
| $\beta_6 = 0.7$               | $\beta_6 = 0.7$               | $\beta_6 = 0.7$               | $\beta_6 = 0.7$               |
| $\beta_7 = 0.7$               | $\beta_7 = 0.7$               | $\beta_7 = 0.7$               | $\beta_7 = 0.7$               |
| $\beta_8 = 0.7$               | $\beta_8 = 0.7$               | $\beta_8 = 0.7$               | $\beta_8 = 0.7$               |
| $\beta_9 = 0.7$               | $\beta_9 = 0.7$               | $\beta_9 = 0.7$               | $\beta_9 = 0.7$               |
| $\beta_{10} = 0$              | $\beta_{10} = 0$              | $\beta_{10} = 0$              | $\beta_{10} = 0$              |

(Continues)
TABLE 1  (Continued)

| n = 500, N = 10,000, (B) High | Bias | BiasSSL | RESSL | BiasUL1 | REUL1 | BiasUL2 | REUL2 |
|-------------------------------|------|---------|-------|---------|-------|---------|-------|
| \( \beta_1 = 0.3 \)          | −1.34| −0.25   | 7.21  | −18.70  | 0.64  | −17.43  | 0.74  |
| \( \beta_3 = 0.3 \)          | −0.33| −0.11   | 6.33  | −18.72  | 0.66  | −17.72  | 0.73  |
| \( \beta_5 = 0.3 \)          | −1.59| −0.16   | 6.45  | −18.56  | 0.64  | −17.90  | 0.69  |
| \( \beta_{10} = 0 \)         | −0.18| 0.04    | 8.94  | 0.09    | 37.43 | 0.06    | 37.77 |

TABLE 2  Bias (×100) of \( \hat{\beta}_3 \), semisupervised \( \hat{\beta}_{SSL} \) and unsupervised \( \hat{\beta}_{UL1}, \hat{\beta}_{UL2} \) as well as relative efficiency (RE) of \( \hat{\beta}_{SSL}, \hat{\beta}_{UL1} \) and \( \hat{\beta}_{UL2} \) compared to \( \hat{\beta}_3 \) under the PH model when the measurement errors in \( T_1 \) and \( T_2 \) are (A) moderate and (B) high

| n = 500, N = 500, (A) Small | Bias | BiasSSL | RESSL | BiasUL1 | REUL1 | BiasUL2 | REUL2 |
|-------------------------------|------|---------|-------|---------|-------|---------|-------|
| \( \beta_1 = 0.7 \)          | −4.39| −4.62   | 1.30  | −23.99  | 0.26  | −22.34  | 0.28  |
| \( \beta_2 = 0.7 \)          | −5.48| −5.64   | 1.30  | −24.86  | 0.26  | −23.71  | 0.28  |
| \( \beta_3 = 0.7 \)          | −3.92| −4.39   | 1.21  | −24.18  | 0.24  | −22.88  | 0.26  |
| \( \beta_4 = −0.5 \)         | 3.99 | 4.75    | 1.30  | 17.07   | 0.40  | 16.36   | 0.39  |
| \( \beta_5 = −0.5 \)         | 2.69 | 4.31    | 1.29  | 17.32   | 0.38  | 15.44   | 0.46  |
| \( \beta_6 = −0.5 \)         | 3.37 | 4.69    | 1.35  | 17.71   | 0.40  | 16.54   | 0.41  |
| \( \beta_7 = 0.3 \)          | −1.98| −1.78   | 1.34  | −10.58  | 0.63  | −9.31   | 0.64  |
| \( \beta_8 = 0.3 \)          | −1.63| −1.54   | 1.49  | −10.29  | 0.68  | −9.36   | 0.71  |
| \( \beta_9 = 0.3 \)          | −1.92| −1.47   | 1.67  | −10.53  | 0.67  | −8.96   | 0.71  |
| \( \beta_{10} = 0 \)         | 0.49 | 1.07    | 1.44  | 0.74    | 1.19  | 0.54    | 1.08  |

| n = 500, N = 10,000, (B) High | Bias | BiasSSL | RESSL | BiasUL1 | REUL1 | BiasUL2 | REUL2 |
|-------------------------------|------|---------|-------|---------|-------|---------|-------|
| \( \beta_1 = 0.7 \)          | −4.39| −4.58   | 1.06  | −49.70  | 0.07  | −52.17  | 0.06  |
| \( \beta_2 = 0.7 \)          | −5.48| −5.75   | 1.03  | −50.43  | 0.07  | −52.46  | 0.07  |
| \( \beta_3 = 0.7 \)          | −3.92| −4.15   | 1.09  | −49.78  | 0.07  | −52.38  | 0.06  |
| \( \beta_4 = −0.5 \)         | 3.99 | 5.25    | 0.95  | 35.31   | 0.11  | 37.48   | 0.10  |
| \( \beta_5 = −0.5 \)         | 2.69 | 4.37    | 0.95  | 35.93   | 0.11  | 37.57   | 0.10  |
| \( \beta_6 = −0.5 \)         | 3.37 | 4.65    | 1.02  | 35.27   | 0.12  | 36.97   | 0.11  |
| \( \beta_7 = 0.3 \)          | −1.98| −1.62   | 1.10  | −21.55  | 0.25  | −22.58  | 0.24  |
| \( \beta_8 = 0.3 \)          | −1.63| −1.45   | 1.10  | −21.37  | 0.27  | −22.24  | 0.26  |
| \( \beta_9 = 0.3 \)          | −1.92| −1.49   | 1.19  | −21.05  | 0.25  | −22.20  | 0.24  |
| \( \beta_{10} = 0 \)         | 0.49 | 1.09    | 1.08  | 0.12    | 1.98  | 0.64    | 2.07  |

| n = 500, N = 10,000, (A) Small | Bias | BiasSSL | RESSL | BiasUL1 | REUL1 | BiasUL2 | REUL2 |
|-------------------------------|------|---------|-------|---------|-------|---------|-------|
| \( \beta_1 = 0.7 \)          | −4.96| −3.22   | 2.32  | −23.79  | 0.31  | −21.41  | 0.38  |
| \( \beta_2 = 0.7 \)          | −4.03| −3.26   | 2.06  | −23.73  | 0.29  | −21.61  | 0.34  |
| \( \beta_3 = 0.7 \)          | −5.10| −2.86   | 2.47  | −23.60  | 0.33  | −21.08  | 0.41  |
| \( \beta_4 = −0.5 \)         | 1.69 | 1.74    | 3.80  | 16.95   | 0.51  | 15.15   | 0.64  |
| \( \beta_5 = −0.5 \)         | 3.43 | 2.04    | 3.83  | 17.16   | 0.51  | 15.50   | 0.62  |
| \( \beta_6 = −0.5 \)         | 3.80 | 1.86    | 4.02  | 16.87   | 0.53  | 15.38   | 0.63  |
| \( \beta_7 = 0.3 \)          | −1.50| −1.37   | 8.45  | −10.30  | 1.24  | −9.15   | 1.55  |
| \( \beta_8 = 0.3 \)          | −1.70| −1.25   | 7.86  | −10.18  | 1.24  | −9.22   | 1.49  |
| \( \beta_9 = 0.3 \)          | −2.88| −1.46   | 7.87  | −10.26  | 1.31  | −9.17   | 1.60  |
TABLE 3 Root mean squared error (rMSE) of norm-preserving soft-thresholded semisupervised estimator $\hat{\beta}_C$ under the PH model. Shown also are the average of the estimated SEs (ASE) from perturbation $\hat{\beta}_C^{(b)}$, empirical coverage levels (CovP) of the 95% CIs

| $n = 500, N = 500$ | (A) Small | (B) High |
|---------------------|-----------|----------|
| $\beta_1$ | 0.1332 | 0.1358 | 0.9540 | 0.1502 | 0.1442 | 0.9380 |
| $\beta_2$ | 0.1330 | 0.1347 | 0.9420 | 0.1471 | 0.1433 | 0.9340 |
| $\beta_3$ | 0.1419 | 0.1356 | 0.9340 | 0.1492 | 0.1439 | 0.9460 |
| $\beta_4$ | 0.1270 | 0.1283 | 0.9580 | 0.1446 | 0.1373 | 0.9340 |
| $\beta_5$ | 0.1301 | 0.1287 | 0.9560 | 0.1474 | 0.1382 | 0.9420 |
| $\beta_6$ | 0.1319 | 0.1279 | 0.9280 | 0.1437 | 0.1374 | 0.9320 |
| $\beta_7$ | 0.1273 | 0.1237 | 0.9460 | 0.1420 | 0.1339 | 0.9380 |
| $\beta_8$ | 0.1250 | 0.1231 | 0.9500 | 0.1447 | 0.1340 | 0.9340 |
| $\beta_9$ | 0.1172 | 0.1234 | 0.9580 | 0.1333 | 0.1333 | 0.9360 |
| $\beta_{10}$ | 0.1172 | 0.1202 | 0.9560 | 0.1276 | 0.1310 | 0.9560 |

TABLE 4 Root mean squared error (rMSE) of $\hat{\beta}_C$ under the PH model. Shown also are the average of the estimated SEs (ASE) from perturbation $\hat{\beta}_C^{(b)}$, empirical coverage levels (CovP) of the 95% CIs

| $n = 500, N = 500$ | (A) Small | (B) High |
|---------------------|-----------|----------|
| $\beta_1$ | 0.1160 | 0.1180 | 0.9600 | 0.1283 | 0.1253 | 0.9580 |
| $\beta_2$ | 0.1202 | 0.1174 | 0.9440 | 0.1349 | 0.1247 | 0.9460 |
| $\beta_3$ | 0.1168 | 0.1190 | 0.9560 | 0.1235 | 0.1256 | 0.9520 |
| $\beta_4$ | 0.1074 | 0.1060 | 0.9600 | 0.1256 | 0.1148 | 0.9360 |
| $\beta_5$ | 0.1084 | 0.1061 | 0.9460 | 0.1264 | 0.1153 | 0.9300 |
| $\beta_6$ | 0.1094 | 0.1055 | 0.9460 | 0.1259 | 0.1152 | 0.9220 |
| $\beta_7$ | 0.1001 | 0.0987 | 0.9440 | 0.1102 | 0.1088 | 0.9500 |
| $\beta_8$ | 0.0971 | 0.0986 | 0.9420 | 0.1129 | 0.1089 | 0.9260 |
| $\beta_9$ | 0.0878 | 0.0989 | 0.9720 | 0.1039 | 0.1087 | 0.9460 |
| $\beta_{10}$ | 0.0883 | 0.0928 | 0.9500 | 0.1017 | 0.1044 | 0.9380 |

| $n = 500, N = 10,000$ | (A) Small | (B) High |
|---------------------|-----------|----------|
| $\beta_1$ | 0.0908 | 0.0850 | 0.9300 | 0.0998 | 0.0939 | 0.9280 |
| $\beta_2$ | 0.0921 | 0.0865 | 0.9400 | 0.0956 | 0.0950 | 0.9440 |
| $\beta_3$ | 0.0903 | 0.0857 | 0.9120 | 0.0924 | 0.0943 | 0.9340 |
| $\beta_4$ | 0.0656 | 0.0707 | 0.9480 | 0.0761 | 0.0819 | 0.9660 |
| $\beta_5$ | 0.0670 | 0.0686 | 0.9380 | 0.0783 | 0.0802 | 0.9400 |
| $\beta_6$ | 0.0670 | 0.0691 | 0.9500 | 0.0774 | 0.0805 | 0.9420 |
| $\beta_7$ | 0.0465 | 0.0534 | 0.9680 | 0.0604 | 0.0682 | 0.9720 |
| $\beta_8$ | 0.0470 | 0.0542 | 0.9660 | 0.0639 | 0.0688 | 0.9720 |
| $\beta_9$ | 0.0476 | 0.0530 | 0.9700 | 0.0625 | 0.0680 | 0.9540 |
| $\beta_{10}$ | 0.0161 | 0.0217 | 1.0000 | 0.0327 | 0.0370 | 0.9940 |
method to develop a genetic risk prediction model for obesity using data from Mass General Brigham Biobank where both genetic and EHR data are available for 27,369 participating patients. We extracted the birth date, gender, single nucleotide polymorphism (SNP) data, timings of obesity diagnosis codes, and mention of obesity in medical notes for the analysis.

Among the subjects, 282 patients were randomly sampled from a total of 27,369 patients to have their obesity status δ annotated by domain experts via manual chart review. A total of 150 patients out of the 282 patients were annotated to have developed obesity with the event rate of 53%. The observation time for the current status C was calculated as the difference in years from their birth date to the date of annotation. The genetic risk score (GRS) for obesity was constructed based on published log odds ratio information on 55 SNPs previously identified as significantly associated with obesity (Speliotes et al., 2010; Hung et al., 2015). For this analysis, we used patient age as timescale. We used two potentially censored surrogates: age at the first diagnostic code for obesity represented by \( \{X_1^*, \Delta_1^*\} \) and age at the first mention of obesity in medical notes represented by \( \{X_2^*, \Delta_2^*\} \). The observation time for surrogate events \( E \) was calculated as the difference in years from their birth date to the last date of EHR codes/notes. However, both surrogates have low specificity in determining the obesity status. Among the annotated 282 patients, 44 patients had false-positive surrogate events, that is codes or mention in medical notes but were labeled as no obesity. To this end, we additionally considered a binary classification \( D \) derived from a phenotyping algorithm (Liao et al., 2019), which has a very high specificity for classifying \( \delta \). We augment the two surrogates with \( D \) and define the final two censored surrogates as \( \{X_k = DX_k^* + (1 - D)E, \Delta_k = D\Delta_k^*, k = 1, 2\} \). In addition to the GRS, we included gender, and the first five principal components of 128 SNPs associated with the ancestry informative genetic markers (Kosoy et al., 2009) to adjust for population stratification. We let \( g(\cdot) \) be the logistic link, and hence \( \beta \) corresponds to the coefficients of the PO model. We set the bandwidths as \( h = 4.32, h'_1 = 0.06, \) and \( h'_2 = 0.05 \). The soft-thresholding parameter \( \lambda \) was selected as \( 1.13 \times 10^{-4} \) by cross-validation.

The point estimators of \( \beta \), as well as their 95% CIs, are shown in Table 5. The results show that the supervised and semisupervised point estimators are reasonably consistent with each other. Our analysis confirms that higher GRS is significantly associated with an elevated risks of developing obesity, with estimated effect 0.796 and \( p \)-value \( < 0.001 \) from the semisupervised method. The negative association between male gender and risk of obesity in our analysis, with estimated effect \(-0.842 \) and \( p \)-value \( 0.05 \) from the semisupervised method, is also consistent with the findings from the literature (Wang and Beydoun, 2007). Moreover, the 95% CIs from the semisupervised estimator are always smaller than the supervised estimator, leading to the positive findings on GRS and male which the supervised method failed to identify. The GRS has estimated standard error 0.334 and \( p \)-value 0.08 in the supervised estimator, while the semisupervised estimator reduced the estimated standard error to 0.221 (relative efficiency 2.28). The male has estimated standard error 0.602 and \( p \)-value 0.30 in the supervised estimator, while the semisupervised estimator reduced the estimated standard error to 0.423 (relative efficiency 2.03). This again demonstrates the benefit of leveraging the information from the mismeasured event times in our proposed procedure.

### Table 5

| \( \beta_s \) | \( \beta_{SSL} \) |
|---|---|
| **Estimate** | **Standard error** | **95% CI** | **P Value** | **Estimate** | **Standard error** | **95% CI** | **P Value** |
| Male | −0.624 | 0.602 | [−1.803,0.555] | 0.30 | −0.842 | 0.423 | [−1.671,0.012] | 0.05 |
| GRS | 0.588 | 0.334 | [−0.067,1.243] | 0.08 | 0.796 | 0.221 | [0.363,1.228] | < 0.001 |
| PC1 | 0.946 | 0.353 | [0.255,1.637] | 0.007 | 0.855 | 0.261 | [0.343,1.367] | 0.001 |
| PC2 | 0.692 | 0.623 | [−0.530,1.914] | 0.27 | 0.274 | 0.336 | [−0.384,0.932] | 0.42 |
| PC3 | 0.107 | 0.472 | [−0.818,1.032] | 0.82 | 0.049 | 0.315 | [−0.568,0.666] | 0.88 |
| PC4 | 0.495 | 0.326 | [−0.143,1.134] | 0.13 | 0.296 | 0.243 | [−0.179,0.771] | 0.22 |
| PC5 | 0.327 | 0.339 | [−0.339,0.992] | 0.34 | 0.114 | 0.239 | [−0.354,0.582] | 0.63 |

5 | **DISCUSSION**

We proposed a robust semisupervised risk prediction procedure in the EHR setting. Our proposed semisupervised estimator is able to effectively integrate two sets of imperfect information on the survival time in the EHR.
database: (i) widely available but noisy surrogate event times; and (ii) current status information manually annotated for a limited set of patients. The semisupervised method can efficiently estimate the risk model without requiring precise information on the event time for any patients in the EHR, which greatly improves the feasibility of performing risk prediction modeling using noisy EHR data. Our numerical results demonstrate that the SSL approach can significantly improve the efficiency of the estimation with large unlabeled data compared to the supervised estimator that only uses the current status data.

In the obesity risk-modeling example, we used age as the timescale and hence baseline is defined at birth. In such a case, only time-invariant covariates such as sex and genetic information can be used as risk factors. However, our method is not restricted to such settings provided that a valid baseline can be defined and subjects are free of the event of interest at baseline. For example, the baseline can be set as a year after the first encounter with the EHR. For rare conditions, patients without any diagnosis code and/or mention in notes of the disease of interest can be classified as free of event at baseline since the absence of codified or natural language processing mention tends to have near-perfect negative predictive value. For such cases, standard risk factors beyond genetics such as lifestyle information can be included as covariates. However, it is generally difficult to precisely determine the presence or absence of a condition in the EHR due to the imperfect sensitivity and specificity of EHR-derived classification rules, and hence only time-invariant covariates should be considered for most cases.

In general, gain in efficiency tends to be little when substantial measurement errors exist in the surrogates, as in most measurement error literature. However, our setting differs from existing literature in two ways: (1) the unlabeled size is substantially larger than the labeled size, and certain estimation errors diminish which in turn improves the gain of the semisupervised estimator; (2) we currently focus on the direction of the coefficients rather than the full $\hat{\beta}_0$ with the mismeasured outcome. We have shown that the surrogate event times follow a single index model with the same linear prediction in Appendix B of the Supporting Information. Hence, we should be able to learn the direction of the coefficient $B_0 = \beta_0 / \| \beta_0 \|_2$ solely from the unlabeled data (Cai and Cheng, 2007). When the sizes of unlabeled data increase, the semisupervised method gains improvement on the estimation precision of $B_0$ which contributes to the overall efficiency gain.

Thresholding has been previously used for the inference of irregular estimators whose asymptotic distribution is not the normal distribution (Chakraborty et al., 2010). When the unlabeled data are much larger than the labeled data $N \gg n$, our semisupervised estimator for zero coefficient is irregular in this sense, which also converges to zero superefficiently at $n^{-1}$ rate. Since we only need to shrink the very small superefficient estimator to zero, our soft-thresholding should work even with a relatively small sample size $n$. With a more involved investigation of the irregular estimators, a better inference procedure can be developed (Laber and Murphy, 2011).

Time-dependent covariates are often considered in the analysis of time-to-event data. Though our method cannot directly incorporate time-dependent covariates, extension is technically possible by applying our idea to the work of Zeng et al. (2016) on semiparametric transformation models for interval-censored data and Antolini et al. (2005) on rank correlation with time-dependent covariates. However, the extension would considerably increase the complexity of the method. For risk prediction of a future event, incorporation of time-dependent covariates has also conceptual issues. Most time-dependent features are often observed progressively during follow-up, so they cannot be used to forecast a new individual with baseline information alone. Moreover, relevant time-dependent features and themselves dependent on the history of the event of interest. In such a case, sequential modeling via structured nested models (Vansteelandt and Joffe, 2014) would be more appropriate than including the feature as a covariate.

We developed our own program solving $\hat{\beta}_2$ and calculating $\hat{S}(\hat{B})$, as well as creating their perturbation samples. We solved the adaptive ridge regression for $\hat{W}$ using “glmnet” package in R with penalty parameter selected by fivefold cross-validation. We invented an efficient iterative algorithm for $\hat{\beta}_2$ that updates the parameter of interest $\beta$ and the nuisance parameter $\phi(C_i)$ separately while sharing the gradient information, which produced the solution within 0.5 s for $n = 500$. We are currently working on a separate manuscript on the algorithm with applications on semiparametric transformation models for the current status data among others. The total computation time for each simulation repeat was around 20 min for $N = 500$ and 3 h for $N = 10,000$. The computation cost lies mostly in $\hat{S}(\hat{B})$ and its perturbation, taking about 15 s for each repeat with $N = 10,000$. The computational efficiency can be improved if reliably closed form estimators for $\hat{W}$ and estimation variance becomes available. The R codes for the simulation are available in the online Supporting Information.

**ACKNOWLEDGMENTS**

Authors Jue Hou and Stephanie F. Chan made equal contributions to the paper and are co-first authors.
DATA AVAILABILITY STATEMENT
The data that support the findings in this paper are available from Mass General Brigham Biobank. Restrictions apply to the availability of these data, which were used under license in this paper. Data are available from the authors with the permission of Mass-General-Brigham.

ORCID
Jue Hou https://orcid.org/0000-0002-9015-1827
Xuan Wang https://orcid.org/0000-0002-0011-5663

REFERENCES
Antolini, L., Boracchi, P., and Biganzoli, E. (2005) A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24, 3927–3944.

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

Braun, D., Gorfine, M., Katki, H.A., Zigos, A., and Parmigiani, G. (2018) Nonparametric adjustment for measurement error in time to event data. *Journal of the American Statistical Association*, 113, 11–25.

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

Calvert, J., Mao, Q., Hoffman, J.L., Jay, M., Desautels, T., Mohamadlou, H., et al. (2016) Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Annals of Medicine and Surgery*, 11, 52–57.

Carroll, R.J., Fan, J., Gijbels, I., and Wand, M.P. (1997) Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 477–489.

Chakraborty, B., Murphy, S., and Strecher, V. (2010) Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19, 317–343.

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.

Eapen, Z.J., Liang, L., Fonarow, G.C., Heidenreich, P.A., Curtis, L.H., Peterson, E.D., et al. (2013) Validated, electronic health record deployable prediction models for assessing patient risk of 30-dayrehospitalization and mortality in older heart failure patients. *JACC: Heart Failure*, 1, 245–251.

GBD 2015 Obesity Collaborators, (2017) Health effects of overweight and obesity in 195 countries over 25 years. *New England Journal of Medicine*, 377, 13–27.

Gravel, C.A., Dewanji, A., Farrell, P.J., and Krewski, D. (2018) A validation sampling approach for consistent estimation of adverse drug reaction risk with misclassified right-censored survival data. *Statistics in Medicine*, 37, 3887–3903.

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

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

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.

Hunsofn., Breen, G., Czumara, D., Corre, T., Wolf, C., Kloiber, S., et al. (2015) A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder. *BMC Medicine*, 13, 86.

Jensen, P.B., Jensen, L.J., and Brunak, S. (2012) Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13, 395.

Jin, B., Che, C., Liu, Z., Zhang, S., Yin, X., and Wei, X. (2018) Predicting the risk of heart failure with EHR sequential data modeling. *IEEE Access*, 6, 9256–9261.

Jin, Z., Ying, Z., and Wei, L.J. (2001) A simple resampling method by perturbing the minimim. *Biometrika*, 88, 381–390.

Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., et al. (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human Mutation*, 30, 69–78.

Laber, E.B. and Murphy, S.A. (2011) Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106, 904–913.

Liao, K.P., Sun, J., Cai, T.A., Link, N., Hong, C., Huang, J., et al. (2019) High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *Journal of the American Medical Informatics Association*, 26, 1255–1262.

Lin, D.Y., Oakes, D., and Ying, Z. (1998) Additive hazards regression with current status data. *Biometrika*, 85, 289–298.

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 Siemex. *Statistics in Medicine*, 37, 1276–1289.

Oh, E.J., Shepherd, B.E., Lumley, T., and Shaw, P.A. (2021) Raking and regression calibration: methods to address bias from correlated covariate and time-to-event error. *Statistics in Medicine*, 40, 631–649.

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.

Sherman, R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, 61, 123–137.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42, 937.

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.

Uno, H., Ritzwoller, D.P., Cronin, A.M., Carroll, N.M., Hornbrook, M.C., and Hassett, M.J. (2018) Determining the time of cancer recurrence using claims or electronic medical record data. *JCO Clinical Cancer Informatics*, 2, 1–10.
Van DerLaan, 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.

Vansteelandt, S. and Joffe, M. (2014) Structural nested models and G-estimation: The partially realized promise. *Statistical Science*, 29, 707–731.

Wang, Y. and Beydoun, M.A. (2007) The obesity epidemic in the United States—gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiologic Reviews*, 29, 6–28.

Zeng, D., Mao, L., and Lin, D.Y. (2016) Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103, 253–271.

**SUPPORTING INFORMATION**

Web Appendices A-F and Figures A1-A2 referenced in Sections 2, 3 and 5, as well as the simulation codes, are available with this paper at the Biometrics website on Wiley Online Library.

---

**How to cite this article:** Hou, J., Chan, S.F., Wang, X., Cai, T. (2023) Risk prediction with imperfect survival outcome information from electronic health records. *Biometrics*, 79, 190–202. https://doi.org/10.1111/biom.13599