Semi-supervised Calibration of Risk with Noisy Event Times (SCORNET) Using Electronic Health Record Data

Yuri Ahuja, Liang Liang, Selena Huang, Tianxi Cai

January 9, 2021

Abstract

Leveraging large-scale electronic health record (EHR) data to estimate survival curves for clinical events can enable more powerful risk estimation and comparative effectiveness research. However, use of EHR data is hindered by a lack of direct event times observations. Occurrence times of relevant diagnostic codes or target disease mentions in clinical notes are at best a good approximation of the true disease onset time. On the other hand, extracting precise information on the exact event time requires laborious manual chart review and is sometimes altogether infeasible due to a lack of detailed documentation. Current status labels – binary indicators of phenotype status during follow up – are significantly more efficient and feasible to compile, enabling more precise survival curve estimation given limited resources. Existing survival analysis methods using current status labels focus almost entirely on supervised estimation, and naive incorporation of unlabeled data into these methods may lead to biased results. In this paper we propose Semi-supervised Calibration of Risk with Noisy Event Times (SCORNET), which yields a consistent and efficient survival curve estimator by leveraging a small size of current status labels and a large size of imperfect surrogate features. In addition to providing theoretical justification of SCORNET, we demonstrate in both simulation and real-world EHR settings that SCORNET achieves efficiency akin to the parametric Weibull regression model, while also exhibiting non-parametric flexibility and relatively low empirical bias in a variety of generative settings.

1 Introduction

The Electronic Health Record (EHR) has in recent years become an increasingly available source of data for clinical and translational research (Kohane *and others*, 2012; Hripcsak and Albers, 2012; Miotto *and others*, 2016). Comprising heterogeneous clinical encounters including diagnostic and procedural billing codes, lab tests, prescriptions, and free text clinical notes for millions of patients, these rich data offer abundant opportunities for in silico epidemiological analysis. One application that has garnered recent interest is estimation of population disease risk within EHR patient cohorts, which can enable more powerful and precise estimation of real-world disease risks as well as comparative effectiveness analysis of alternative treatment strategies (Hodgkins *and others*, 2017; Dean *and others*, 2003; Liu *and others*, 2018; Panahiazar *and others*, 2015; Steele *and others*, 2018). Several studies have had success estimating time to death within rule-defined disease cohorts (Panahiazar *and others*, 2015; Steele *and others*, 2018). However, estimating the temporal risk of developing a disease is a more challenging task due to EHR's lack of direct observations of either disease status or the timing of disease onset. Convenient proxies of disease status or onset time based on readily available features such as International Classification of Disease (ICD) codes often exhibit low specificity and systematic temporal biases, potentially yielding highly biased disease risk estimators if used as event time labels (Cipparone *and others*, 2015; Uno *and others*, 2018). On the other hand, extracting precise information on disease outcomes requires labor-intensive manual chart review, which is particularly challenging for event times since the event may occur outside of the hospital system and only be mentioned during follow-up visits. It is thus only practically feasible to annotate the current status $\Delta = I(T \leq C)$ of the event time T, where C is the follow up time.

In this paper, we consider the problem of estimating the disease risk $F(t) = P(T \leq t)$ when only a small number of labels on Δ and a large quantity of unlabeled EHR features $\mathbf{W}$, including proxies of T, are available. Supervised survival curve estimation with current status data on $\{\Delta, C\}$ is well established in the statistical literature with several available parametric, semi-parametric and non-parametric procedures (Vardi, 1982; Huang,

1996; van der Laan and Robins, 1998; van der Laan and Jewell, 2003; Lin *and others*, 2019, e.g.). For example, van der Laan and Robins (1998) proposed a non-parametric, locally efficient estimator via inverse probability of censoring weighting (IPCW), assuming that (1) T and C are conditionally independent given some informative baseline covariates $\mathbf{Z_0} \subset \mathbf{W}$ (e.g. age, sex, etc.) and (2) a consistent estimator for the conditional density of $C \mid \boldsymbol{Z}_0$ is available. However, these existing estimators do not leverage unlabeled EHR feature information such as time to first surrogate ICD code, which may greatly improve risk estimation precision.

Since $\mathbf{W}$ may be highly predictive of T, the estimation of $S(t)$ can potentially be improved via semi-supervised learning (SSL) leveraging both the small set of Δ observations in the labeled set and the EHR features $\mathbf{W}$ in the unlabeled set. SSL has been shown to significantly mitigate bias and/or improve efficiency for various risk prediction applications (Chai *and others*, 2017; Liang *and others*, 2016; Bair and Tibshirani, 2004; Golub *and others*, 1999). For example, several studies employ semi-parametric models to impute event times in the unlabeled set for subsequent input into an outcome survival model alongside labeled data (Chai *and others*, 2017; Liang *and others*, 2016; Zhao *and others*, 2014; Uno *and others*, 2018; Hassett *and others*, 2017; Chubak *and others*, 2012; Choi *and others*, 2015; Kaji *and others*, 2019; Ruan *and others*, 2019; Ahuja *and others*, 2020*a*). While such imputation approaches may improve efficiency under correct specification of the imputation model, they are subject to significant bias if the imputation model is misspecified. In addition, these existing methods do not allow for use of current status labels for training. Other general augmented inverse probability weighting methods in the missing data literature (Seaman and White, 2013; Rotnitzky and Robins, 2014, e.g.) are not directly applicable here since the probabilities of labels being observed tend to zero in the SSL setting.

We address this shortcoming by proposing Semi-supervised Calibration of Risk with Noisy Event Times (SCORNET) for estimation of $S(t)$. SCORNET utilizes current status labels while also employing a robust semi-supervised imputation approach on the extensive unlabeled set to maximize survival estimation efficiency. To mitigate imputation bias and maximize efficiency gain from the unlabeled data, SCORNET utilizes a highly flexi-

ble semi-non-parametric kernel regression model with EHR features as covariates, which ensures the validity of the resulting risk estimator without requiring the imputation model to hold. In addition to providing theoretical justifications for the SCORNET estimator, we illustrate via simulation studies that SCORNET substantially outperforms existing methods with regards to the bias-variance tradeoff. The rest of the paper is organized as follows. In Section 2, we detail the SCORNET procedure along with its associated inference procedures. In Section 3, we report risk estimation performance relative to existing methods in diverse simulation studies. To further illustrate the utility of SCORNET in clinical applications, we apply it to a real-world EHR study estimating the risk of heart failure among rheumatoid arthritis patients in Section 4. Finally, in Section 5 we briefly discuss the strengths, weaknesses, and potential applications of SCORNET.

2 Methods

2.1 Setup

Let T denote the event time for which we are interested in estimating a cumulative distribution function $F(t) = P(T \leq t)$ and survival function $S(t) = 1 - F(t)$. In the EHR study we do not observe T but rather $\Delta = I(T \leq C)$ for a small labeled subset, where C is the follow up time with finite support $[0, \tau_c]$. For all subjects, we also observe a set of baseline covariates $\boldsymbol{Z}_0$ and longitudinal EHR features $\boldsymbol{Z}$. Since codes used in the EHR are often highly sensitive but not specific, there often exists some filter variable $\mathbb{F} \in \{0, 1\}$ such that $\Delta_i \mid (\mathbb{F}_i = 0, \mathbf{W}_i) = 0$ almost surely, where $\mathbf{W}_i = (\boldsymbol{Z}_{0,i}^\intercal, \boldsymbol{Z}_i^\intercal)^\intercal$. Moreover, we assume that $(T, \boldsymbol{Z})C \mid \boldsymbol{Z}_0$. We assume that data for analysis consist of a small set of n current-status-labeled observations randomly selected among those with $\mathbb{F} = 1$ along with a larger set of N unlabeled observations: $\mathbb{D} = \{\mathbf{D}_i = (C_i, V_i\Delta_i, \mathbf{W}_i, V_i, \mathbb{F}_i)^\intercal, i = 1, ..., N\} = \mathbb{L} \cup \mathbb{U}$, where $\mathbb{L} = \{(C_i, \Delta_i, \mathbf{W}_i, 1, 1)^\intercal : \mathbb{F}_i = 1, V_i = 1, i = 1, ..., n\}$ and $\mathbb{U} = \{(C_i, 0, \mathbf{W}_i, 0, \mathbb{F}_i)^\intercal : V_i = 0, i = n + 1, ..., N\}$ with $\log(N)/\log(n) \to \nu_0 > 3/2$ as $n \to \infty$.

Since the censoring C may depend on $\mathbf{Z}_0$, we follow the IPCW strategy of van der Laan and Robins (1998) to weight observations by

$$\omega_{t,b}(C \mid \mathbf{Z}_0) = \frac{K_b(C-t)}{f_c(t \mid \mathbf{Z}_0)}$$

where $K_b(s) = K(s/b)/b$, $f_c(t \mid \mathbf{Z}_0) = dF_c(t \mid \mathbf{Z}_0)/dt$, $F_c(t \mid \mathbf{Z}_0) = P(C \leq t \mid \mathbf{Z}_0)$, $K(\cdot)$ is some symmetric density function, and $0 < b = O(n^{-\nu})$ is the bandwidth thereof with $\nu \in (1/5, 1/3]$. IPCW enables consistent estimation of functionals of $T \leq t$ and $\mathbf{W}$ since for any reasonable choice of function $q(\cdot)$ and $a, d \in \{0,1\}$,

$$E\left\{\Delta_i^d q(\mathbf{W}_i)\mathbb{F}_i^a \omega_{t,b}(C_i \mid \mathbf{Z}_{0,i})\right\} = E\left\{I(T_i \leq t)^d q(\mathbf{W}_i)\mathbb{F}_i^a\right\} + O(b^2). \tag{1}$$

The IPCW estimator for $\pi(t) = P(T_i \leq t \mid \mathbb{F}_i = 1)$ proposed by van der Laan and Robins (1998) essentially corresponds to

$$(t) = \frac{\sum_{i=1}^n \mathbb{F}_i \Delta_i \omega_{t,b}(C_i \mid \mathbf{Z}_{0,i})}{\sum_{i=1}^n \mathbb{F}_i \omega_{t,b}(C_i \mid \mathbf{Z}_{0,i})}$$

with $f_c(t \mid \mathbf{Z}_0)$ in $\omega_{t,b}(C_i \mid \mathbf{Z}_{0,i})$ replaced by a consistent estimator that converges faster than $n^{-1/4}$, which is not difficult to achieve under reasonable modeling assumptions since $C_i \mid \mathbf{Z}_{0,i}$ can be estimated using the full data $\mathbb{D}$. To this end, we propose to derive an estimator for the conditional density $f_c(t \mid \mathbf{Z}_0) = \lambda_c(t \mid \mathbf{Z}_{0,i})S_c(t \mid \mathbf{Z}_{0,i})$ by imposing a semi-parametric model for $C \mid \mathbf{Z}_0$. Although many commonly employed models can be used since once again C is fully observed for all patients, we illustrate our proposal by focusing on the Cox proportional hazards model (Cox, 1972) under which

$$\lambda_c(t \mid \mathbf{Z}_{0,i}) = \lambda_{0c}(t)e^{\boldsymbol{\gamma}^\intercal \mathbf{Z}_{0,i}} \quad \text{and} \quad S_c(t \mid \mathbf{Z}_{0,i}) \equiv 1 - F_c(t \mid \mathbf{Z}_{0,i}) = \exp\left\{-\Lambda_{0c}(t)e^{\boldsymbol{\gamma}^\intercal \mathbf{Z}_{0,i}}\right\}, \tag{2}$$

where $\lambda_c(t \mid \mathbf{Z}_{0,i})$ is the conditional hazard function of $C_i \mid \mathbf{Z}_{0,i}$, $\lambda_{0c}(t)$ is the unknown baseline hazard function, $\Lambda_{0c}(t) = \int_0^t \lambda_{0c}(s)ds$, and $\boldsymbol{\gamma}$ is the vector of unknown covariate effects.

2.2 SCORNET Estimation

As outlined in Figure 1, SCORNET consists of three steps: (1) estimating the conditional censoring distribution $h(t \mid \mathbf{Z}_0)$ using $\mathbb{D}$; (2) fitting an imputation *working* model for $\pi(t \mid \mathbf{W}) \equiv P(T \leq t \mid \mathbf{W}, \mathbb{F} = 1)$ using $\mathbb{L}$, denoting

the estimate of $\pi(t \mid \mathbf{W})$ as $(t \mid \mathbf{W})$; and (3) estimating $S(t)$ by marginalizing $(t \mid \mathbf{W})\mathbb{F} + \Delta\ (1 - \mathbb{F}) = (t \mid \mathbf{W})\mathbb{F}$ via IPCW.

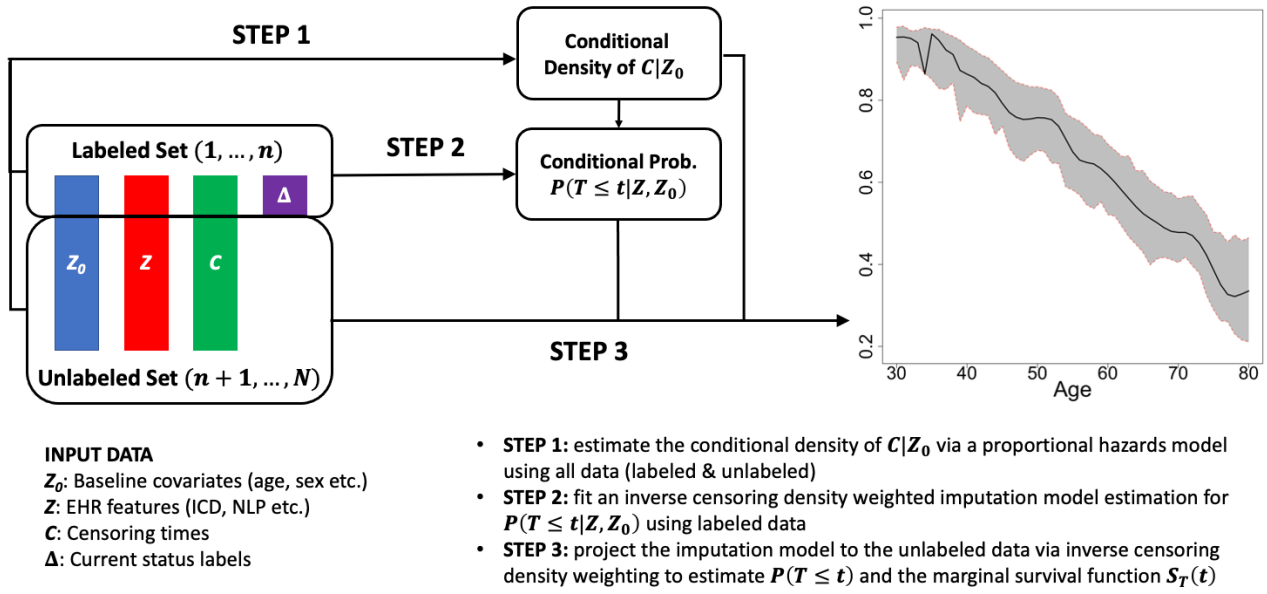

Figure 1: Schematic of the SCORNET algorithm.

2.2.1 Step 1: Estimate $f_c\ (t \mid \mathbf{Z}_0)$ Under the Cox Model for $C \mid \mathbf{Z}_0$

To estimate $f_c(t \mid \mathbf{Z}_0)$, we fit the Cox model (2) to the full data $\mathbb{D}$ to obtain the partial likelihood estimator $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$. We subsequently estimate $\Lambda_0(t)$ and $\lambda_{0c}(t)$ respectively as the standard Breslow estimator $\hat{\Lambda}_{0c}(t)$ and the kernel-smoothed Breslow estimator $\hat{\lambda}_{0c}(t)$ (Basha and Hoxha, 2019), where

$$\hat{\Lambda}_{0c}(t) = \sum_{j=1}^{N} \frac{I(C_j \leq t)}{\sum_{i=1}^{N} I(C_i \geq C_j)\exp\left(\hat{\boldsymbol{\gamma}}^{\top}\mathbf{Z}_{0,i}\right)}, \quad \hat{\lambda}_{0c}(t) = \sum_{j=1}^{N} \frac{K_{a_N}\left(C_j - t\right)}{\sum_{i=1}^{N} I(C_i \geq C_j)\exp\left(\hat{\boldsymbol{\gamma}}^{\top}\mathbf{Z}_{0,i}\right)},$$

and $a_N = O(N^{-\nu_c})$ with $\nu_c \in (1/5, 1/3]$. We then obtain

$$\hat{\lambda}_c(t \mid \mathbf{Z}_0) = \hat{\lambda}_{0c}(t)e^{\hat{\boldsymbol{\gamma}}^{\top}\mathbf{Z}_{0,i}} \text{ and } \hat{S}_c(t \mid \mathbf{Z}_0) = \exp\left\{-\hat{\Lambda}_{0c}(t)e^{\hat{\boldsymbol{\gamma}}^{\top}\mathbf{Z}_{0,i}}\right\},$$

and we estimate $f_c(t \mid \mathbf{Z}_{0,i})$ as $\hat{f}_c(t \mid \mathbf{Z}_{0,i}) = \hat{\lambda}_c(t \mid \mathbf{Z}_{0,i})\hat{S}_c(t \mid \mathbf{Z}_{0,i})$. Following standard asymptotic results for non-parametric kernel regression (Pagan and Ullah, 1999), it is not difficult to show that $\sup_{\mathbf{Z}_{0,t}} |\hat{f}_c(t \mid$

$\boldsymbol{Z}_0) - f_c(t \mid \boldsymbol{Z}_0)| = O_p\{\log(N)^{1/2}(Na_N)^{-1/2}\} = o_p()$. We denote the resulting estimate for the censoring weight as $\hat{\omega}_{b_n}(t \mid \boldsymbol{Z}_{0,i}) = K_{b_n}(C_i - t)/\hat{f}_c(t \mid \boldsymbol{Z}_{0,i})$.

2.2.2 Step 2: Estimate an Imputation Model $\pi(t \mid \mathbf{W}_i) \equiv P\,(T_i \leq t \mid \mathbf{W}_i, \mathbb{F}_i = 1)$

To leverage the unlabeled data, we fit a flexible imputation *working* model

$$\pi(t \mid \mathbf{W}_i) = g\left\{\alpha(t) + \boldsymbol{\beta}_0(t)^\intercal \vec{\boldsymbol{Z}}_{0,i} + \boldsymbol{\beta}(t)^\intercal \boldsymbol{Z}_i\right\} = g\left\{\boldsymbol{\theta}(t)^\intercal \vec{\mathbf{W}}_i\right\} \tag{3}$$

where $\boldsymbol{Z}_i$ denotes the EHR surrogate features, $\boldsymbol{\theta}(t) = \left(\alpha(t), \boldsymbol{\beta}_0(t)^\intercal, \boldsymbol{\beta}(t)^\intercal\right)^\intercal$, and $\vec{\mathbf{W}}_i = \left(1, \vec{\boldsymbol{Z}}_{0,i}^\intercal, \boldsymbol{Z}_i^\intercal\right)^\intercal$. Under (3), $P(T_i \leq t \mid \mathbf{W}_i, \mathbb{F}_i = 1) = g\left\{\boldsymbol{\theta}(t)^\intercal \vec{\mathbf{W}}_i\right\}$, and hence we may estimate $\boldsymbol{\theta}(t)$ as $\hat{\boldsymbol{\theta}}(t) = \left(\hat{\alpha}(t), \hat{\boldsymbol{\beta}}_0(t)^\intercal, \hat{\boldsymbol{\beta}}(t)^\intercal\right)^\intercal$, the solution to the IPCW estimating equation evaluated with $\mathbb{L}$,

$$\sum_{i=1}^{n} \hat{\omega}_{t,b_n}(C_i \mid \boldsymbol{Z}_{0,i}) \mathbb{F}_i \vec{\mathbf{W}}_i \left\{\Delta_i - g\left(\boldsymbol{\theta}^\intercal \vec{\mathbf{W}}_i\right)\right\} = 0,$$

where $b_n = O(n^{-\nu})$ with $\nu \in (1/5, 1/2)$. In practice, b_n can be chosen via either standard cross-validation or heuristic plug-in values. For a future observation with filter status $\mathbb{F}_i = 1$ and covariates $\mathbf{W}_i$, we impute $I(T_i \leq t)$ as the conditional risk $\hat{\pi}(t \mid \mathbf{W}_i) = g\left\{\hat{\boldsymbol{\theta}}(t)^\intercal \vec{\mathbf{W}}_i\right\}$.

It is not difficult to show that $\hat{\boldsymbol{\theta}}(t)$ converges in probability to $\bar{\boldsymbol{\theta}}(t)$, the solution to the limiting estimating equation

$$E\left[\vec{\mathbf{W}}_i \left\{I(T_i \leq t) - g(\boldsymbol{\theta}^\intercal \vec{\mathbf{W}}_i)\right\} \mid \mathbb{F}_i = 1\right] = 0,$$

which ensures that

$$E\left\{\bar{\pi}(t \mid \mathbf{W}_i) \mid \mathbb{F}_i = 1\right\} = P(T_i \leq t \mid \mathbb{F}_i = 1), \quad \text{where} \quad \bar{\pi}(t \mid \mathbf{W}_i) = g\{\bar{\boldsymbol{\theta}}(t)^\intercal \vec{\mathbf{W}}_i\}, \tag{4}$$

regardless of the adequacy of the imputation model (3).

2.2.3 Step 3: Estimate $F(t)$ by Marginalizing Imputed Risks

Finally, we marginalize the imputed values $\hat{\pi}_{it} = \hat{\pi}(t \mid \mathbf{W}_i)\ \forall\, \mathbb{F}_i = 1$ and $\Delta_i = 0\ \forall\, \mathbb{F}_i = 0$ to estimate $F(t)$. Since $\mathbb{F}_i$ depends on C_i, we again employ IPCW to marginalize $\hat{\pi}_{it}\mathbb{F}_i + \Delta_i(1 - \mathbb{F}_i) = \hat{\pi}_{it}\mathbb{F}_i$ and thereby construct our

final estimator for $F(t)$:

$$\hat{F}(t) = \frac{\sum_{i=1}^{N}\{\hat{\pi}_{it}\mathbb{F}_i + \Delta_i(1-\mathbb{F}_i)\}\hat{\omega}_{t,a_N}(C_i \mid \mathbf{Z}_{0,i})}{\sum_{i=1}^{N}\hat{\omega}_{t,a_N}(C_i \mid \mathbf{Z}_{0,i})} = \frac{\sum_{i=1}^{N}\hat{\pi}_{it}\mathbb{F}_i\hat{\omega}_{t,a_N}(C_i \mid \mathbf{Z}_{0,i})}{\sum_{i=1}^{N}\hat{\omega}_{t,a_N}(C_i \mid \mathbf{Z}_{0,i})}.$$

2.3 Inference for $\hat{F}(t)$

Following standard theory for non-parametric kernel regression (Pagan and Ullah, 1999), we show in the Supplementary Materials that $\hat{F}(t) \to P(T_i \le t, \mathbb{F}_i = 1) + P(T_i \le t, \mathbb{F}_i = 0) = P(T_i \le t) = F(t)$ in probability under mild regularity conditions and correct specification of the censoring model regardless of the adequacy of the imputation model. Here, we note that for any $t \in [0, \tau_c]$, $0 = P(\Delta_i = 0 \mid \mathbb{F}_i = 0, C_i = t, \mathbf{W}_i)$ implies that $P(T_i \le t \mid \mathbb{F}_i = 0) = 0$. Furthermore,

$$\begin{aligned}(nb_n)^{1/2}\left\{\hat{F}(t) - F(t)\right\} &= \left(\frac{b_n}{n}\right)^{1/2}\sum_{i=1}^{n}\hat{\omega}_{t,b_n}(C_i \mid \mathbf{Z}_{0,i})\mathbb{F}_i\left\{\Delta_i - \bar{\pi}(t \mid \mathbf{W}_i)\right\} + o_p(1)\\ &= \left(\frac{b_n}{n}\right)^{1/2}\sum_{i=1}^{n}\omega_{t,b_n}(C_i \mid \mathbf{Z}_{0,i})\mathbb{F}_i\left\{\Delta_i - \bar{\pi}(t \mid \mathbf{W}_i)\right\} + o_p(1)\end{aligned}$$

since $\sup_t |\hat{f}_c(t \mid \mathbf{Z}_0) - f_c(t \mid \mathbf{Z}_0)| = o_p()$. It follows that $(nb_n)^{1/2}\{\hat{F}(t) - F(t)\}$ is asymptotically normal with mean 0 and variance

$$\sigma^2(t) = R(K)E\{\mathbb{V}(t \mid \mathbf{Z}_{0,i})/f_c(t \mid \mathbf{Z}_{0,i})\}, \tag{5}$$

where

$$\mathbb{V}(t \mid \mathbf{Z}_{0,i}) = E[\mathbb{F}_i\{I(T_i \le t) - \bar{\pi}(t \mid \mathbf{W}_i)\}^2 \mid \mathbf{Z}_{0,i}]$$

and $R(K) = \int K(x)^2 dx$. Our derivation for the asymptotic distribution of $\hat{F}(t)$ can effectively ignore the variability associated with the estimation of censoring weights, which simplifies the asymptotic variance $\sigma^2(t)$. Importantly, $\sigma^2(t)$ decreases as the imputation model approximates $\pi(t \mid \mathbf{W}_i)$ better since $\mathbb{V}(t \mid \mathbf{Z}_{0,i}) = E[\mathbb{F}_i\{I(T_i \le t) - \pi(t \mid \mathbf{W}_i)\}^2 \mid \mathbf{Z}_{0,i}] + E[\mathbb{F}_i\{\pi(t \mid \mathbf{W}_i) - \bar{\pi}(t \mid \mathbf{W}_i)\}^2 \mid \mathbf{Z}_{0,i}]$ decreases. To estimate $\sigma^2(t)$ in practice, one may construct a plug-in estimator,

$$\hat{\sigma}^2(t) = \frac{b_n}{n}\sum_{i=1}^{n}\hat{\omega}_{t,b_n}(C_i \mid \mathbf{Z}_{0,i})^2\mathbb{F}_i\left\{\Delta_i - g\left(\hat{\boldsymbol{\theta}}(t)^{\intercal}\vec{\mathbf{W}}_i\right)\right\}^2.$$

By contrast, the supervised IPCW estimator that incorporates filter negative patients takes the form

$$\hat{F}(t) = \frac{\sum_{i=1}^{N}\{\hat{\pi}(t)\mathbb{F}_i + \Delta_i(1-\mathbb{F}_i)\}\hat{\omega}_{t,a_N}(C_i \mid \boldsymbol{Z}_{0,i})}{\sum_{i=1}^{N}\hat{\omega}_{t,a_N}(C_i \mid \boldsymbol{Z}_{0,i})} = \frac{\sum_{i=1}^{N}\hat{\pi}(t)\mathbb{F}_i\hat{\omega}_{t,a_N}(C_i \mid \boldsymbol{Z}_{0,i})}{\sum_{i=1}^{N}\hat{\omega}_{t,a_N}(C_i \mid \boldsymbol{Z}_{0,i})}.$$

The asymptotic variance of $(nb_n)^{1/2}\{\hat{F}(t) - F(t)\}$ is then $\sigma^2(t) = R(K)E\{\mathbb{V}(t \mid \boldsymbol{Z}_{0,i})/f_c(t \mid \boldsymbol{Z}_{0,i})\}$, where

$$\mathbb{V}(t \mid \boldsymbol{Z}_{0,i}) = E\left[\mathbb{F}_i\{I(T_i \leq t) - \pi(t)\}^2 \mid \boldsymbol{Z}_{0,i}\right].$$

The variance $\sigma^2(t)$ is equivalent to that of SCORNET if and only if the feature set $\mathbf{W}$ is uninformative for T (i.e. $T\mathbf{W}$). Supervised IPCW is otherwise less efficient, with relative efficiency controlled by the relative magnitudes of the marginal error $E[\{I(T_i \leq t) - F(t)\}^2 \mid \mathbb{F}_i = 1, \boldsymbol{Z}_{0,i}]$ and the conditional error $E[\{I(T_i \leq t) - \bar{\pi}(t \mid \mathbf{W}_i)\}^2 \mid \mathbb{F}_i = 1, \boldsymbol{Z}_{0,i}]$.

3 Simulation Study

We conduct extensive simulation experimentation to evaluate the finite sample performance of the proposed SCORNET estimator in realistic settings with $n \in \{100, 200\}$ observed labels within the set of filter-positive patients, defining the filter to have 99% sensitivity and 88% specificity for Δ. We compare SCORNET to three existing survival function estimators with current status data: 1) parametric Weibull Accelerated Failure Time (AFT) regression with interval event times (Lin *and others*, 2019), 2) semi-parametric Cox Proportional Hazards regression with interval event times and Breslow baseline hazard estimation (Huang, 1996; Cox, 1972; Breslow, 1972), and 3) non-parametric IPCW estimation (van der Laan and Robins, 1998). We incorporate the filter in the Weibull and Cox models by setting $\Delta_i \mid (\mathbb{F}_i = 0) = 0$ and weighting the n labeled filter-positive patients by $\frac{1}{n}\sum_{i=1}^{N}\mathbb{F}_i$. Weibull and Cox are implemented using the icenReg package in R, while IPCW is implemented per the algorithm detailed in van der Laan and Robins (1998), estimating $C \mid \boldsymbol{Z}_0$ using $\mathbb{D}$ under the Cox model. We note that estimating the censoring distribution using $\mathbb{L}$ yields similar asymptotic performance to using $\mathbb{D}$, but in finite sample settings the latter offers higher efficiency.

Setting	$\boldsymbol{Z}_0 \sim$	$C \mid \boldsymbol{Z}_0 \sim$	$T \mid \boldsymbol{Z}_0 \sim$	$\boldsymbol{Z} \sim$
1	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10e^{-0.5\boldsymbol{Z}_0}, \frac{3}{2}\right)$	$\text{Weibull}\left(15e^{-0.3\boldsymbol{Z}_0}, \frac{5}{2}\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$
2	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10e^{-0.5\boldsymbol{Z}_0}, \frac{3}{2}\right)$	$\text{Weibull}\left(15, \frac{5}{2}\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$
3	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10e^{-0.5\boldsymbol{Z}_0}, \frac{3}{2}\right)$	$\text{Weibull}\left(15e^{-0.3\boldsymbol{Z}_0^2}, \frac{5}{2}\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$
4	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10e^{-0.5\boldsymbol{Z}_0}, \frac{3}{2}\right)$	$\text{Logistic}\left(15 - 4\boldsymbol{Z}_0, 3\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$
5	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10, \frac{3}{2}\right)$	$\text{Weibull}\left(15e^{-0.3\boldsymbol{Z}_0}, \frac{5}{2}\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$
6	$\text{Unif}(-1, 1)$	$\text{Weibull}\left(10e^{-0.5\boldsymbol{Z}_0^2}, \frac{3}{2}\right)$	$\text{Weibull}\left(15e^{-0.3\boldsymbol{Z}_0}, \frac{5}{2}\right)$	$\text{Normal}\{T+1, \sigma(C)/2\}$

Table 1: Generative parameters employed in our simulation study.

We consider 6 diverse generative mechanisms as detailed in Table 3, including cases where Weibull-distributed accelerated failure time of $T \mid \boldsymbol{Z}_0$, proportional hazards of $T \mid \boldsymbol{Z}_0$, and proportional hazards of $C \mid \boldsymbol{Z}_0$ are respectively violated, as well as cases where SCORNET's imputation model is and is not misspecified. In settings 1, 2, and 5, we consider various cases where SCORNET and all comparator methods are correctly specified. In setting 1 we consider the specific case where C and T both depend on $\boldsymbol{Z}_0$, and both $C \mid \boldsymbol{Z}_0$ and $T \mid \boldsymbol{Z}_0$ are Weibull-distributed satisfying accelerated failure time and proportional hazards. In setting 2, by contrast, we consider a case where $T\boldsymbol{Z}_0$ to assess robustness to over-parametrization of this relationship, and in setting 5 we consider a case where $C\boldsymbol{Z}_0$ to evaluate robustness to over-parametrization thereof. In settings 3 and 4 we assess the benefit of SCORNET and IPCW's robustness to the distribution of $T \mid \boldsymbol{Z}_0$ when this distribution satisfies neither Weibull accelerated failure time nor proportional hazards. We evaluate SCORNET's sensitivity to misspecification of the imputation model in settings 1, 3, and 5, as compared to correct specification thereof in settings 2 and 4. Finally, in setting 6 we assess the sensitivity of SCORNET and IPCW to misspecification of the conditional censoring model $C \mid \boldsymbol{Z}_0$. For each given configuration, we compute the empirical bias, standard error, and root mean squared error (RMSE) of all estimators for $F(t)$ based on their average performance on

500 simulated datasets evaluated at 100 equally-spaced time points $t \in [Q_C(0.1) + b_n, Q_C(0.9) - b_n]$, where Q denotes the quantile function of C under the configuration. We used plug-in bandwidths of $b_n = \hat{s}(C)n^{-1/4}$ and $a_N = \hat{s}(C)N^{-1/4}$ for the imputation (Step 2) and marginalization (Step 3) steps of SCORNET respectively, where $\hat{s}$ is the empirical standard deviation of observed C. We present the performance of the estimators averaged over the selected time points using $n = 200$ labels in Figure 2. The performance at each time point can be found in Supplementary Figure 1, and time-averaged performance using $n = 100$ labels can be found in Supplementary Table 1 of the Supplementary Materials.

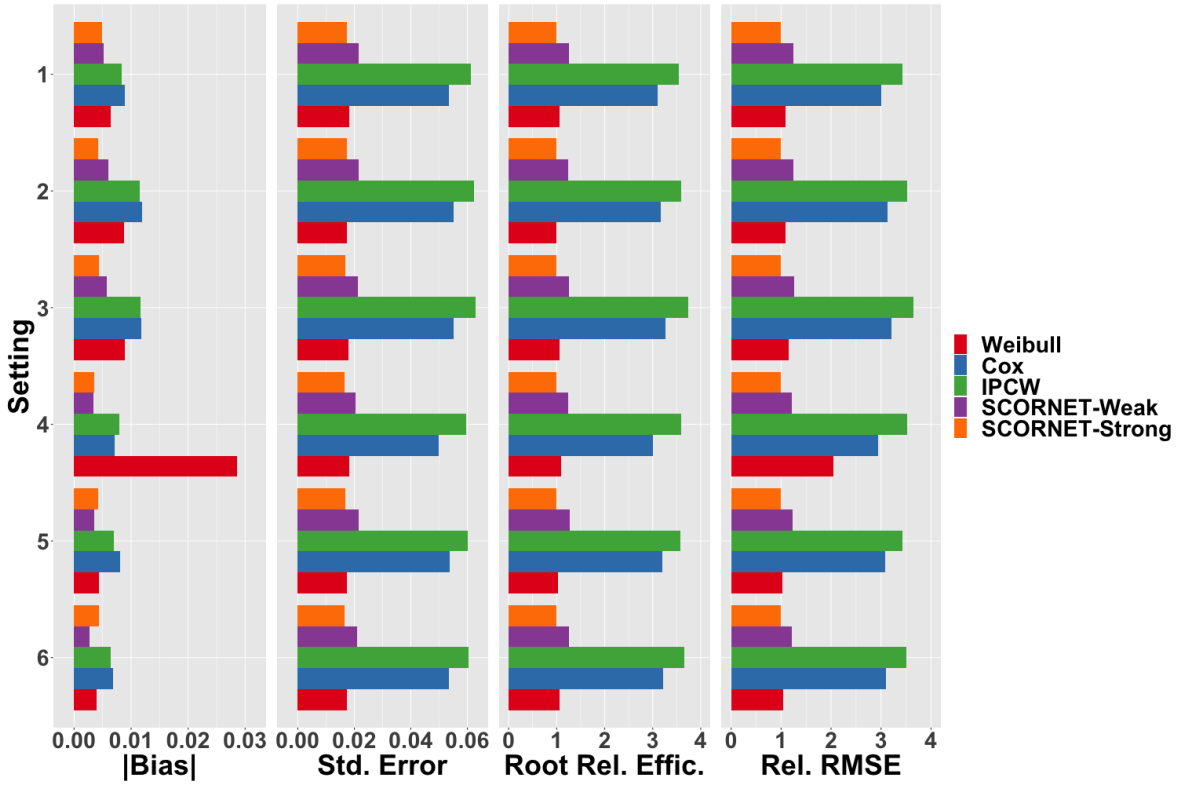

Figure 2: Time-averaged empirical absolute biases (left), standard errors (second from left), root relative efficiencies (second from right), and relative RMSEs (right) of the Weibull Accelerated Failure Time (red), Cox Proportional Hazards w/ Breslow baseline (blue), supervised IPCW (green), and SCORNET estimators using weakly informative (purple) and strongly informative (orange) surrogates, in various simulated settings with $n = 200$ observed current status labels.

As Figure 2 demonstrates, imputing using a strongly informative feature $\boldsymbol{Z}$ (SCORNET-Strong) results in consistently higher efficiency than just using the weakly informative baseline $\boldsymbol{Z}_0$ (SCORNET-Weak), which in turn is markedly more efficient than not leveraging the unlabeled set at all (IPCW). SCORNET makes minimal assumptions regarding the distribution of $T \mid \boldsymbol{Z}_0$, settling for non-parametric efficiency in exchange for enhanced flexibility. By contrast, the Weibull regression model fully parametrizes $T \mid \boldsymbol{Z}_0$, and the Cox model assumes proportional hazards thereof, increasing efficiency at the expense of bias in the case of misspecification. As expected, Weibull consistently achieves higher empirical efficiency than Cox, which in turn is more efficient than IPCW across settings. Notably, SCORNET consistently achieves empirical efficiency comparable to Weibull and significantly higher than Cox despite being far more flexible than both, again highlighting the efficiency gained by leveraging auxiliary information to impute unobserved risks. At the same time, SCORNET is much less susceptible to model misspecification bias than Weibull, as demonstrated by the latter's significantly higher bias and RMSE in Setting 4. Indeed, SCORNET achieves relatively low mean absolute bias across settings, with MSE apparently dominated by variance rather than bias in the setting of 100-200 labels. Consistent with the theory, SCORNET is robust to misspecification of the imputation model in settings 1, 3, and 5, achieving equivalently insignificant bias as in setting 2 and marginally but not meaningfully higher bias than in setting 4. That said, correctness of the imputation model in settings 2 and 4 does not yield any meaningful change in relative efficiency, likely because inherent variability functionally dominates imputation model bias given so few labels. Reassuringly, SCORNET (and IPCW) appear insensitive to misspecification of $C \mid \boldsymbol{Z}_0$ in setting 6, achieving functionally equivalent bias to the correctly-specified Weibull and Cox models. Altogether, these results corroborate the assertion that SCORNET's semi-supervised utilization of informative feature data to impute risks in the unlabeled set improves estimation efficiency without introducing bias regardless of the validity of the imputation model. Moreover, they suggest that SCORNET is particularly valuable in settings where (1) flex-

ibility is desired with regard to the distribution of $T \mid \mathbf{Z}_0$, and (2) there exists a large set of unlabeled patients with associated EHR data – both commonplace in retrospective observational studies.

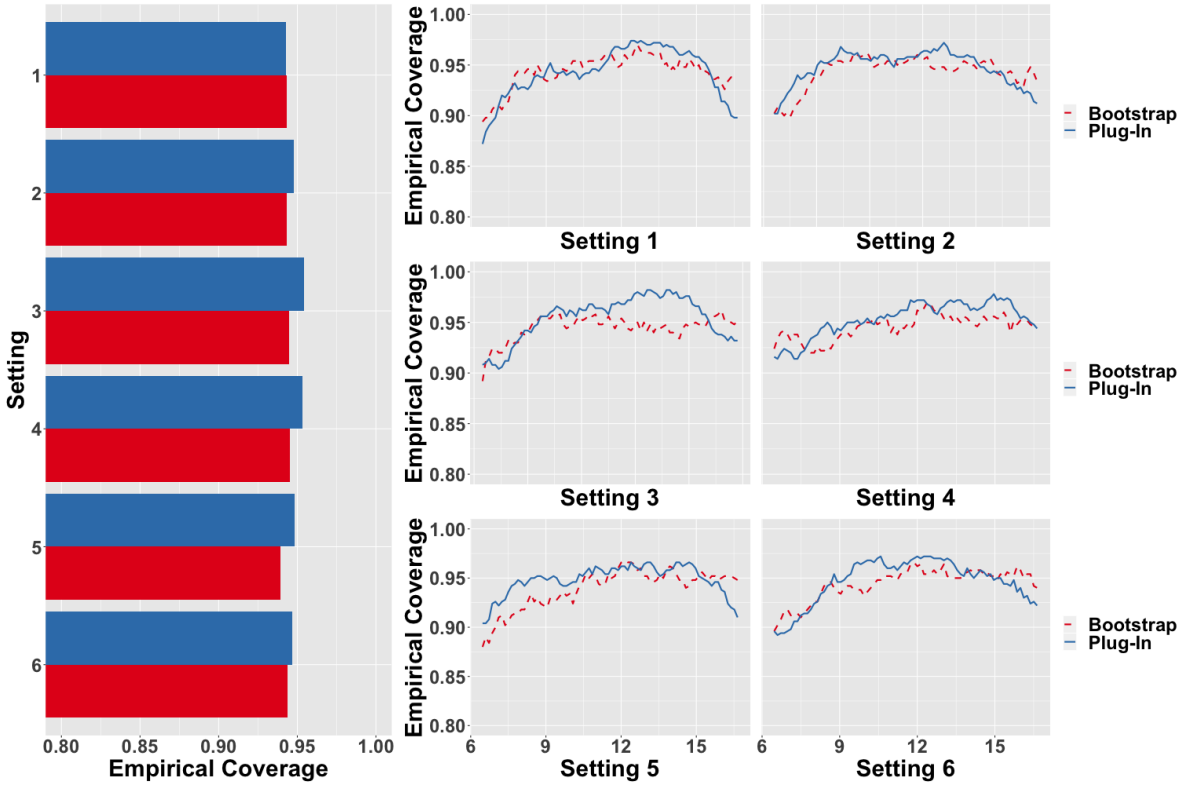

Figure 3: Empirical coverage probabilities averaged over time (left) and plotted over time (right) of SCORNET-Strong's 95% confidence intervals constructed with the bootstrap (red) and plug-in (blue) standard error estimators in various simulated settings with $n = 200$ observed current status labels. See Table 1 for details of the generative mechanism employed in each setting.

To assess the finite sample performance of the proposed interval estimation procedures, we obtain standard error estimates both using the proposed plug-in estimator $\hat{\sigma}(t)$ and via bootstrap with 500 replicates. In Figure 3 we demonstrate empirical coverage probabilities of SCORNET's 95% Wald confidence intervals constructed using each standard error estimator, both averaged over the selected timepoints (left) and at each timepoint (right). Reassuringly, we find that the 95% confidence intervals using both plug-in and bootstrap estimators achieve nearly 95% mean coverage across settings. Coverage only drops below 90% at the tails of the event time

support due to moderately increased bias from kernel smoothing thereabout. The plug-in estimator achieves marginally lower coverage than the bootstrap estimator at the right tail due to underestimation of the true standard error, likely because of overfitting of the imputation model given low local censoring density (and thus low effective N). Notably, we do not observe this trend in setting 4, wherein correct specification of the imputation model obviates overfitting. Thus, we posit that the plug-in estimator can be reliably used for finite sample problems with $n \in [100, 200]$ labels as long as the conditional censoring density $f_c(C \mid \mathbf{Z}_0)$ is sufficiently high and the timepoints evaluated are sufficiently far from the tails of the event time support.

4 Application to Assessing Heart Failure Risk Among Rheumatoid Arthritis Patients

Rheumatoid arthritis (RA), a chronic inflammatory disease that affects approximately 1% of the general population, is associated with dramatically increased risk of heart failure (HF) morbidity and mortality (Kaplan, 2010; Nicola *and others*, 2005, 2006; Ahlers *and others*, 2020). One study estimated that RA patients have a 1.9-fold lifetime risk of developing HF compared to matched RA-negative controls (Nicola *and others*, 2005), while another estimated that HF accounts for 13% of excess mortality among RA patients (Nicola *and others*, 2006). Ongoing interest lies in estimating the risk of developing HF subtypes in RA cohorts and quantifying the risk modifying effect of various RA treatments (Ahlers *and others*, 2020). Due to the increased availability of electronic health record (EHR) data, it is now possible to assess HF risk for a broader RA population using these data. For example, at Mass General Brigham we previously established an EHR cohort of $N_0 = 16,358$ RA patients (Huang *and others*, 2020). This large RA cohort can potentially be used to study the longitudinal risk of HF among RA patients.

However, such analysis is not straightforward as HF status is not readily available within the RA cohort. We propose to estimate HF risk among RA patients by leveraging (1) n current status labels on HF status obtained

via manual chart review, and (2) informative yet unlabeled EHR data, including time to first ICD code for HF, as surrogate variables $\boldsymbol{Z}$. We estimate both the age-specific HF risk, $F_{\text{age}}(\cdot)$, and the risk of developing HF after the patient's incident ICD code for RA (714), $F_{\text{RA+}}(\cdot)$, among patients with at least 6 months of follow up whose incident RA codes occur after the age of 16 to select for adult-onset as opposed to juvenile RA. Among filter-positive patients, defined as having at least 1 ICD code for HF, we have $n = 300$ labels on censoring time HF status Δ for age-specific HF risk, and we have $n = 126$ for post-RA HF risk. We let the baseline covariates $\boldsymbol{Z}_0$ include sex and decade of first EHR event for $F_{\text{age}}(\cdot)$, and sex, decade of first RA code, and age at first RA code for $F_{\text{RA+}}(\cdot)$. We obtain HF risk estimators based on SCORNET as well as the aforementioned comparator estimators. For the imputation model in Step 2 of SCORNET, we consider three EHR-derived surrogate risk predictors for $\boldsymbol{Z}$: (1) the predicted Δ based on the unsupervised Multimodal Automated Phenotyping (MAP) algorithm, which uses the total counts of HF ICD codes and mentions of HF in clinical notes, as well as the total count of all ICD codes as a healthcare utilization measure (Liao *and others*, 2019), (2) the predicted Δ based on the unsupervised Surrogate-guided Ensemble Latent Dirichlet Allocation (sureLDA) algorithm, which leverages the features used in MAP as well as 121 additional manually-selected EHR features including counts of relevant medications, ICD codes, and concept unique identifiers (CUIs) in clinical notes (Ahuja *and others*, 2020*b*); and (3) the time to first HF ICD code. As in our simulation, we select plug-in bandwidths of $b_n = \hat{s}(C)n^{-1/4}$ and $a_N = \hat{s}(C)N^{-1/4}$ for the imputation and marginalization steps of SCORNET respectively, and we evaluate risk at 100 timepoints $t \in [Q_C(0.1) + b_n, Q_C(0.9) - b_n]$. We again compare the performance of SCORNET to that of Weibull, Cox, and IPCW, incorporating the filter in the Weibull and Cox models by propensity weighting as we do in the simulation study.

In Figure 4, we show the estimated HF risk curves along with their standard errors. Reassuringly, all methods appear to agree rather closely for estimation of both age-specific HF risk and HF risk after RA diagnosis. For the latter quantity, however, Weibull and Cox appear to underfit while IPCW appears to overfit relative to

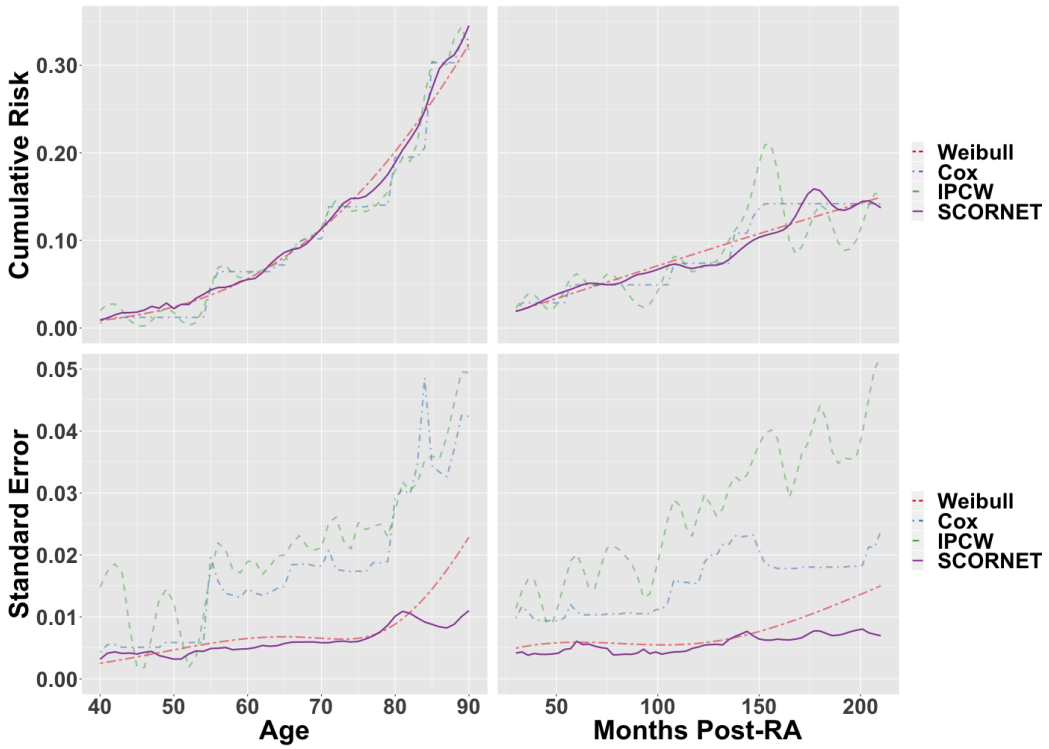

Figure 4: Estimated age-specific and post-RA cumulative risks of heart failure (top) and bootstrap standard errors thereof (bottom) over time of the Weibull Accelerated Failure Time (red, short-long-dashed), Cox Proportional Hazards w/ Breslow baseline (blue, dot-dashed), supervised IPCW (green, dashed), and SCORNET (purple, solid) estimators.

the SCORNET estimator, which appears to achieve a reasonable middle ground. Moreover, SCORNET once again attains standard errors comparable to those of the Weibull estimator and meaningfully lower than those of the Cox and IPCW estimators. This suggests that while the Weibull and Cox models potentially fail to capture the complexity of the post-RA HF risk function, and IPCW is too unstable for a limited labeled set of size $n = 126$, SCORNET offers an attractive balance of efficiency and flexibility and is thus well conditioned for such a scenario. As shown in Figure 5, averaged over the selected timepoints, the root relative efficiency of SCORNET is 1.11, 2.55, and 3.31 compared to the Weibull, Cox, and IPCW estimators respectively for estimation of age-specific risk, and 1.34, 2.32, and 3.85 respectively for estimation of HF risk after RA diagnosis. Once again, the fact that SCORNET achieves efficiency moderately higher than the relatively inflexible Weibull model and sig-

nificantly higher than the Cox and IPCW estimators reflects the value of leveraging available information from the EHR to bolster risk estimation efficiency.

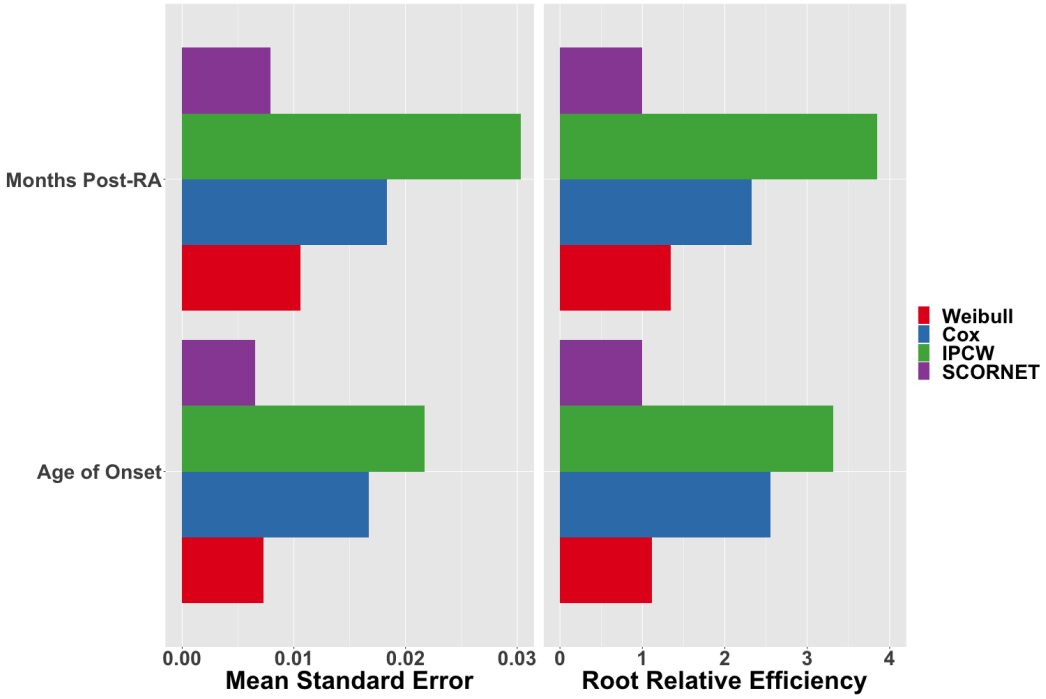

Figure 5: Time-averaged bootstrap standard errors (left) and empirical root relative efficiencies (right) of the Weibull Accelerated Failure Time (red), Cox Proportional Hazards w/ Breslow baseline (blue), IPCW (green), and SCORNET (purple) estimators for estimation of (1) age-specific HF risk (left), and (2) HF risk after RA diagnosis (right), among RA patients in the Partners EHR database.

5 Discussion

By leveraging a sizeable unlabeled data set containing imperfect surrogates of the true event times and a small set with observed current status labels, the SCORNET estimator serves as a robust and efficient alternative to existing model-free survival estimators with current status data. The semi-supervised nature of SCORNET makes it well-conditioned to EHR-based survival estimation in settings where only a limited number of labels are available or readily obtainable. Moreover, by only requiring current status labels rather than the precise

timing of event onset, SCORNET greatly reduces the burden of chart review and increases the feasibility of studying disease risk using EHR data.

To allow for covariate-dependent censoring, which is frequently present in observational settings, SCORNET requires additional assumptions on the distribution of $C \mid \boldsymbol{Z}_0$. Although we choose the proportional hazards model for illustration, any standard semi-parametric model will yield similar properties for the resulting estimator. Since $\{C, \boldsymbol{Z}_0\}$ are observed for all subjects, one can potentially allow for more flexible (i.e. non-parametric) censoring models. That said, our simulation results suggest that SCORNET is relatively insensitive to misspecification of the model for $C \mid \boldsymbol{Z}_0$. Even under mild misspecification, it achieves consistently lower mean squared errors than existing estimators.

When interest lies in assessing how risk differs across different patient sub-populations, it is straightforward to extend SCORNET to estimate subgroup-specific risks for a small number of subgroups. However, future research is warranted to estimate covariate-specific risks for a general set of covariates.

6 Software

An R package, including a sample use case and complete documentation, is available at https://cran.r-project.org/web/packages/SCORNET/index.html. Source code can be found at https://github.com/celehs/SCORNET.

Funding

This work was supported by the U.S. National Institutes of Health Grants T32-AR05588512, T32-GM7489714, and R21-CA242940.

Acknowledgements

The authors declare no conflicts of interest.

References

Ahlers, Michael J., Lowery, Brandon D., Farber-Eger, Eric, Wang, Thomas J., Bradham, William, Ormseth, Michelle J., Chung, Cecilia P., Stein, C. Michael and Gupta, Deepak K. (2020). Heart failure risk associated with rheumatoid arthritis-related chronic inflammation. *Journal of the American Heart Association*, 9.

Ahuja, Yuri, Hong, Chuan, Xia, Zongqi and Cai, Tianxi. (2020*a*). Samgep: A novel method for prediction of phenotype event times using the electronic health record. *Preprint*.

Ahuja, Yuri, Zhou, Doudou, He, Zeling, Sun, Jiehuan, Castro, Victor M, Gainer, Vivian, Murphy, Shawn N, Hong, Chuan and Cai, Tianxi. (2020*b*). surelda: A multidisease automated phenotyping method for the electronic health record. *Journal of the American Medical Informatics Association* **27**(8), 1235–1243.

Bair, Eric and Tibshirani, Robert. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology* **2**(4), E108.

Basha, Lule and Hoxha, Fatmir. (2019). Kernel estimation of the baseline function in the cox model. *European Scientific Journal* **15**(6), 105–118.

Breslow, Norman E. (1972). Discussion of professor cox's paper. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.

Chai, Hua, Li, Zi-na, Meng, De-yu, Xia, Liang-yong and Liang, Yong. (2017). A new semi-supervised learning model combined with cox and sp-aft models in cancer survival analysis. *Scientific Reports* **7**(13053).

Choi, Edward, Du, Nan, Chen, Robert, Song, Le and Sun, Jimeng. (2015). Constructing disease network and temporal progression model via context-sensitive hawkes process. IEEE Computer Society. pp. 101–108.

Chubak, Jessica, Yu, Onchee, Pocobelli, Gaia, Lamerato, Lois, Webster, Joe, Prout, Marianna N, Yood, Marianne Ulcickas, Barlow, William E and Buist, Dianna SM. (2012). Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *Journal of the National Cancer Institute* **104**(12), 931–940.

Cipparone, Charlotte W, Withiam-Leitch, Matthew, Kimminau, Kim S, Fox, Chet H, Singh, Ranjit and Kahn, Linda. (2015). Inaccuracy of icd-9 codes for chronic kidney disease: A study from two practice-based research networks (pbrns). *The Journal of the American Board of Family Medicine* **28**(5), 26094.

Cox, David R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B* **34**, 187–220.

Dean, Bonnie B, Lam, Jessica, Natoli, Jaime L, Butler, Qiana, Aguilar, Daniel and Nordyke, Robert J. (2003). Use of electronic medical records for health outcomes research: A literature review. *Medical Care Research and Review* **31**(6), 611–638.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. *and others*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**(5439), 531–537.

Hassett, Michael J, Uno, Hajime, Cronin, Angel M, Carroll, Nikki M, Hornbrook, Mark C and Ritzwoller, Debra. (2017). Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Medical Care* **55**(12), e88–e98.

Hodgkins, Adam J, Bonney, Andrew, Mullan, Judy, Mayne, Darren John and Barnett, Stephen. (2017). Survival analysis using primary care electronic health record data: A systematic review of the literature. *Health Information Management Journal* **47**(1), 6–16.

Hripcsak, George and Albers, David J. (2012). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* **20**(1), 117–121.

Huang, Jian. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics* **24**(2), 540–568.

Huang, Sicong, Huang, Jie, Cai, Tianrun, Dahal, Kumar P, Cagan, Andrew, He, Zeling, Stratton, Jacklyn, Gorelik, Isaac, Hong, Chuan, Cai, Tianxi *and others*. (2020). Impact of icd10 and secular changes on electronic medical record rheumatoid arthritis algorithms. *Rheumatology*.

Kaji, Deepak A, Zech, John R, Kim, Jun S, Cho, Samuel K, Dangayach, Neha S, Costa, Anthony B and Oermann, Eric K. (2019). An attention based deep learning model of clinical events in the intensive care unit. *PLoS One* **14**(2), e0211057.

Kaplan, Mariana J. (2010). Cardiovascular complications of rheumatoid arthritis - assessment, prevention. and treatment. *Rheumatic Disease Clinics of North America* **36**(2), 405–426.

Kohane, Isaac S, Churchill, Susanne E and Murphy, Shawn N. (2012). A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association* **19**(2), 181–185.

Liang, Yong, Chai, Hua, Liu, Xiao-Ying, Xu, Zong-Ben, Zhang, Hai and Leung, Kwong-Sak. (2016). Cancer survival analysis using semi-supervised learning method based on cox and aft models with l1/2 regularization. *BMC Medical Genomics* **9**(11), 11.

Liao, Katherine P, Sun, Jiehuan, Cai, Tianrun A, Link, Nicholas, Hong, Chuan, Huang, Jie, Huffman, Jennifer E, Gronsbell, Jessica, Zhang, Yichi, Ho, Yuk-Lam, Castro, Victor, Gainer, Vivian, Murphy, Shawn N, O'Donnell, Christopher J, Caziano, J Michael, Cho, Kelly, Szolovits, Peter, Kohane, Isaac S, Yu, Sheng *and others*. (2019). High-throughput multimodal automated phenotyping (map) with application of phewas. *Journal of the American Medical Informatics Association* **26**(11), 1255–1262.

Lin, Hung-Mo, Williamson, John M and Kim, Hae-Young. (2019). Firth adjustment for weibull current-status survival analysis. *Communications in Statistics - Theory and Methods* **49**(18), 4587–4602.

Liu, Bin, Li, Ying, Sun, Zhaonan, Ghosh, Soumya and Ng, Kenney. (2018). Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In: *The 32nd AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence. pp. 101–108.

Miotto, Riccardo, Li, Li, Kidd, Brian A and Dudley, Joel T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports* **6**(6), 26094.

Nicola, Paulo J., Crowson, Cynthia S., Maradit-Kremers, Hilal, Ballman, Karla V., Roger, Veronique L., Jacobsen, Steven J. and Gabriel, Sherine E. (2006). Contribution of congestive heart failure and ischemic heart disease to excell mortality in rheumatoid arthritis. *Arthritis Rheumatology* **54**(1), 60–67.

Nicola, Paulo J., Maradit-Kremers, Hilal, Roger, Veronique L., Jacobsen, Steven J., Crowson, Cynthia S., Ballman, Karla V. and Gabriel, Sherine E. (2005). The risk of congestive heart failure in rheumatoid arthritis: a population-based study over 46 years. *Arthritis Rheumatology* **52**(2), 412–420.

Pagan, Adrian and Ullah, Aman. (1999). *Nonparametric econometrics*. Cambridge university press.

Panahiazar, Maryam, Taslimitehrani, Vahid, Pereira, Naveen and Pathak, Jyotishman. (2015). Using ehrs and machine learning for heart failure survival analysis. *Studies in Health Technology and Informatics* **216**, 40–44.

ROTNITZKY, ANDREA AND ROBINS, JAMES M. (2014). Inverse probability weighting in survival analysis. *Wiley StatsRef: Statistics Reference Online*.

RUAN, TONG, LEI, LIQI, ZHOU, YANGMING, ZHAI, JIE, ZHANG, LE, HE, PING AND GAO, JU. (2019). Representation learning for clinical time series prediction tasks in electronic health records. *BMC Medical Informatics and Decision Making* **19**(259).

SEAMAN, SHAUN R AND WHITE, IAN R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* **22**(3), 278–295.

STEELE, ANDREW J, DENAXAS, SPIROS C, SHAH, ANOOP D, HEMINGWAY, HARRY AND LUSCOMBE, NICHOLAS M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* **13**(8), e0202344.

UNO, HAJIME, RITZWOLLER, DEBRA P, CRONIN, ANGEL M, CARROLL, NIKKI M, HORNBROOK, MARK C AND HASSETT, MICHAEL J. (2018). Determining the time of cancer recurrence using claims or electronic medical record data. *JCO Clinical Cancer Informatics* **2**, 1–10.

VAN DER LAAN, MARK J AND JEWELL, NICHOLAS P. (2003). Current status and right-censored data structures when observing a marker at the censoring time. *The Annals of Statistics* **31**(2), 512–535.

VAN DER LAAN, MARK J AND ROBINS, JAMES M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association* **93**(442), 693–701.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics* **10**, 178–203.

ZHAO, YUE, HERRING, AMY H, ZHOU, HAIBO, ALI, MIRZA W AND KOCH, GARY G. (2014). A multiple imputation method for sensitivity analyses of time-to-event data with possibly informative censoring. *Journal of Biopharmaceutical Statistics* **24**(2), 229–253.