The Canadian Journal of Statistics
Vol. 00, No. 00, 2020, Pages 000–000
La revue canadienne de statistique

Estimation in the Cox cure model with covariates missing not at random, with application to disease screening/prediction

Lisha GUO[1], Yi XIONG[2], and X. Joan HU[2*]

[1]*School of Mathematics and Statistics, South-Central University for Nationalities, Wuhan, China*
[2]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada*

Key words and phrases: Mixture model; pseudo-likelihood estimation; right-censored event time; semiparametric regression analysis; supplementary information.

MSC 2010: Primary 62N01; secondary 62N02.

Abstract: In an attempt to provide a statistical tool for disease screening and prediction, we propose a semiparametric approach to analysis of the Cox proportional hazards cure model in situations where the observations on the event time are subject to right censoring and some covariates are missing not at random. To facilitate the methodological development, we begin with semiparametric maximum likelihood estimation (SPMLE) assuming that the (conditional) distribution of the missing covariates is known. A variant of the EM algorithm is used to compute the estimator. We then adapt the SPMLE to a more practical situation where the distribution is unknown and there is a consistent estimator based on available information. We establish the consistency and weak convergence of the resulting pseudo-SPMLE, and identify a suitable variance estimator. The application of our inference procedure to disease screening and prediction is illustrated via empirical studies. The proposed approach is used to analyze the tuberculosis screening study data that motivated this research. Its finite-sample performance is examined by simulation. *The Canadian Journal of Statistics* 00: 000–000; 2020 © 2020 Statistical Society of Canada

Résumé: Dans le but de fournir des outils pour la détection et la prévision de maladies, les auteures proposent une méthode semi-paramétrique pour l'analyse du modèle de cure aux risques proportionnels de Cox lorsque les observations des temps aux événements sont censurés à droite et que certaines covariables sont manquantes de façon non aléatoire. Afin de faciliter le développement méthodologique, elles utilisent d'abord l'estimateur au maximum de vraisemblance semi-paramétrique (EMVSP) sous l'hypothèse que la distribution conditionnelle des valeurs manquantes est connue. Elles exploitent une version de l'algorithme EM pour calculer l'estimateur, puis adaptent l'EMVSP à une situation plus plausible où cette distribution est inconnue et où il existe un estimateur convergent basé sur l'information connue. Les auteures établissent la convergence en probabilité et la convergence faible du pseudo-EMVSP résultant, puis identifient un estimateur approprié de sa variance. Elles illustrent leur procédure empiriquement sur les données réelles de détection de la tuberculose à l'origine de cet article. Elles examinent également la performance de la méthode par des simulations. *La revue canadienne de statistique* 00: 000–000; 2020 © 2020 Société statistique du Canada

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Author to whom correspondence may be addressed.*
E-mail: joanh@stat.sfu.ca

1. INTRODUCTION

Cook, Hu & Swartz (2011) present a regression analysis of data from a tuberculosis (TB) study conducted at the Centre for Disease Control of British Columbia (BCCDC, www.bccdc.ca). They assume that the time to TB onset of a TB contact since the physical contact with an active infectious TB patient (source) follows the Cox proportional hazards (PH) model conditional on the potential covariates. They deal with the covariate missing not at random (MNAR) in the study data using supplementary information on the covariate.

However, the times to TB onset in the study were heavily censored: out of 7,921 study subjects, only 65 TB cases were observed (less than 0.8%); the estimated survivor function of the TB time under the Cox PH model has a long and literally unchangeable right tail as shown in Table 1B and Figure 1. As pointed out in Sy & Taylor (2000), this may be taken as empirical evidence of a fraction of *nonsusceptible* subjects in the study. Conventional survival analysis approaches regard all the subjects as *susceptible*. In addition, studies of infectious diseases, such as the TB study, typically aim to distinguish between susceptible and nonsusceptible subjects and then to predict times to disease onset for susceptible subjects. These considerations motivated the research presented in this article.

Many studies in medical research, reliability and other areas encounter situations where a substantial proportion of the study individuals never experience the event of interest. A familiar example is one where some subjects are cured of the disease of interest, and their survival times are not observable even with extended follow-up time. A commonly used model for such phenomena is the two-component mixture cure model (Berkson & Gage, 1952), which assumes the whole population to be a mixture of *susceptible* and *nonsusceptible* subjects. The cure model has been investigated by many authors, and various inferential methods have been proposed and studied both asymptotically and numerically; see, for example, Farewell (1982, 1986), Kuk & Chen (1992), Taylor (1995), Maller & Zhou (1996), Sy & Taylor (2000), Peng & Dear (2000), Lu & Ying (2004), Fang, Li & Sun (2005) and Lu (2008). This article focuses on semiparametric estimation under the Cox PH cure model (e.g., Kuk & Chen, 1992; Sy & Taylor, 2000), a two-component mixture cure model, and its application in disease screening and prediction with missing covariates.

Missing covariates add another layer of complexity to the statistical challenges. Most published approaches in the analysis of event times require the missing mechanism to be missing at random (MAR); see, for example, Beesley et al. (2016) and Chen & Ibrahim (2001). Herring & Ibrahim (2002) propose to account for nonignorable missing covariates under the Cox PH frailty models by assuming that the missing covariates follow parametric distributions. We aim to deal with MNAR under the Cox PH cure model with an unknown distribution for missing covariates by extending the approach proposed by Cook, Hu & Swartz (2011).

To facilitate the development of our proposed estimation procedure, we begin with semiparametric maximum likelihood estimation (SPMLE) assuming that the (conditional) distribution of the missing covariates is known. The SPMLE is then adapted to a more realistic situation following Cook, Hu & Swartz (2011). This yields the proposed estimator, a pseudo-SPMLE. By arguments similar to those of Guo, Hu & Liu (2017), we establish the consistency and weak convergence of the pseudo-SPMLE and identify a consistent variance estimator. An algorithm that is intuitive and easy to implement is developed to compute the proposed estimator. The procedure may be viewed as an application of the EM algorithm (Dempster, Laird & Rubin, 1977). It extends the EM application in Sy & Taylor (2000) to account for MNAR covariates. It can also be viewed as an adaptation of the EM application in Cook, Hu & Swartz (2011) for the Cox PH cure model.

We analyze the TB study data using our proposed approach. This analysis provides new insights that explain some counterintuitive findings that were reported in the previous analyses by Cook, Hu & Swartz (2011) and Guo, Hu & Liu (2017). Using the results of our analysis,

we describe how to screen likely nonsusceptible subjects and how to predict the TB onset time for a potentially susceptible subject. In addition, we report a simulation study that examined the finite-sample performance of our proposed approach. The simulation results are consistent with the interesting findings of our analysis of the TB data.

The rest of this article is organized as follows. Section 2 introduces the framework of our approach. Section 3 presents the SPMLE and then the pseudo-SPMLE using readily available information on the population distribution of the covariate for which the observations are subject to MNAR. We use the EM algorithm to compute the pseudo-SPMLE; the asymptotic properties of the estimator and its variance estimation are derived. Section 4 reports an analysis of the TB study data using our approach and illustrates disease screening and prediction based on the analysis. It also includes the results of a simulation study that we conducted to examine the finite-sample performance of the pseudo-SPMLE. A few concluding remarks may be found in Section 5.

2. NOTATION, MODELLING, AND ASSUMPTIONS

Suppose that an event time T conditional on the covariates (X, Z) follows the semiparametric logistic/PH mixture model (the Cox PH cure model). That is, T can be written

$$T = \eta T^* + (1 - \eta)\infty, \tag{1}$$

where η indicates whether an individual is susceptible and will eventually experience the event, $0 \leq T^* < \infty$ denotes the event time of a susceptible individual, and the distributions of η and T^* conditional on (X, Z) are specified as follows.

The indicator η follows the logistic regression model with the predictor variables X and Z, and

$$P(\eta = 1|X, Z) = \frac{\exp\left(\gamma_0 + \gamma_1' X + \gamma_2' Z\right)}{1 + \exp\left(\gamma_0 + \gamma_1' X + \gamma_2' Z\right)}, \tag{2}$$

which is denoted by $\pi(X, Z; \gamma)$; the event time T^* follows the Cox PH model (Cox, 1972),

$$\lambda(t|X, Z) = \lambda_0(t) \exp\left(\alpha' X + \beta' Z\right). \tag{3}$$

Here α, β and $\gamma = (\gamma_0, \gamma_1', \gamma_2')'$ are the regression parameter vectors, and $\lambda_0(\cdot)$ is the unspecified baseline hazard function. Denote the corresponding conditional survivor and density functions of T^* by $S(t|X, Z) = \exp\left\{-\Lambda(t|X, Z)\right\}$ and $f(t|X, Z) = \lambda(t|X, Z)S(t|X, Z)$, respectively, where $\Lambda(t|X, Z) = \int_0^t \lambda(s|X, Z)ds = \Lambda_0(t) \exp\left(\alpha' X + \beta' Z\right)$ with $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$.

Consider a study involving n independent individuals and with such an event time T as the primary response. The realizations of (T, X, Z) associated with the study individuals are assumed to be independent and identically distributed (i.i.d.); they are denoted by (T_i, X_i, Z_i) for $i = 1, \ldots, n$. Suppose that observations on the event times T_i are subject to right censoring with the associated censoring times C_i. Suppose further that, while all the covariates Z_i are available, some covariates X_i are missing, with the indicator R_i for the availability of X_i. Our primary interest is in estimating the regression parameters α, β and γ together with the cumulative baseline hazard function $\Lambda_0(\cdot)$ under the combination of regression models specified in Equations (2) and (3) with the available data $\left\{(U_i, \Delta_i, R_i X_i, R_i, Z_i) : i = 1, \ldots, n\right\}$, where $U_i = T_i \wedge C_i$ is the minimum of T_i and C_i, and Δ_i indicates whether or not $T_i \leq C_i$. Let $\boldsymbol{\theta} = \left\{\alpha, \beta, \gamma, \Lambda_0(\cdot)\right\}$.

We are particularly interested in situations where the assumption of MAR (e.g., Little & Rubin, 2002), that is, $P(R = 1|U, \Delta, X, Z) = P(R = 1|U, \Delta, Z)$, is not necessarily satisfied. Let

the cumulative distribution of X given R and Z be $G_R(\cdot|Z)$. The distribution of X conditional on Z is

$$G(x|Z) = p(Z)G_1(x|Z) + \{1 - p(Z)\}G_0(x|Z) \tag{4}$$

with $p(z) = P(R = 1|Z = z)$. Here $G_1(\cdot|Z)$ and $G_0(\cdot|Z)$ are the conditional distributions of the observed X and the unobserved X given Z, respectively.

We make the following two assumptions, which are plausible in many practical situations.

A1. The censoring is noninformative conditional on (X, Z): $T \perp C|(X, Z)$; that is, $\big[T, C|X, Z\big] = \big[T|X, Z\big]\big[C|X, Z\big]$.

A2. Given the covariates (X, Z), the missingness of X is independent of (U, Δ): $\big[U, \Delta|R, X, Z\big] = \big[U, \Delta|X, Z\big]$.

Although η, the indicator of susceptibility in the cure model specified in Equation (1) is in general unobservable, it is known that $\eta = 1$ if $\Delta = 1$. The contribution to the observed data likelihood function of $\boldsymbol{\theta}$ from an individual is proportional to $\big[U, \Delta|X, Z\big]\big[X|R = 1, Z\big]\big[R = 1|Z\big]$ when $R = 1$ and $\int \big[U, \Delta|X, Z\big]d\big[X|R = 0, Z\big]\big[R = 0|Z\big]$ when $R = 0$, where $\big[U, \Delta|X, Z\big]$ is proportional to

$$\Big\{\big[T^* = U|X, Z\big]\big[\eta = 1|X, Z\big]\Big\}^{\Delta}\Big\{\big[T^* > U|X, Z\big]\big[\eta = 1|X, Z\big] + \big[\eta = 0|X, Z\big]\Big\}^{1-\Delta}. \tag{5}$$

Under the Cox cure model, Equation (5) becomes

$$\Big\{f(U|X, Z)\pi(X, Z; \gamma)\Big\}^{\Delta}\Big\{S(U|X, Z)\pi(X, Z; \gamma) + \bar{\pi}(X, Z; \gamma)\Big\}^{1-\Delta}, \tag{6}$$

where $\bar{\pi}(X, Z; \gamma) = 1 - \pi(X, Z; \gamma)$. Thus, its value is determined by the PH function identified in Equation (3). For ease of exposition, let $B(U, \Delta, X, Z|\boldsymbol{\theta})$ denote Equation (6).

The likelihood function of $\boldsymbol{\theta}$ based on the observed data, denoted by $L(\boldsymbol{\theta}; G(\cdot))$, is

$$\prod_{i=1}^{n}\Big[\Big\{B(U_i, \Delta_i, X_i, Z_i|\boldsymbol{\theta})g_1(X_i|Z_i)p(Z_i)\Big\}^{R_i} \times\Big\{\int B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta})dG_0(x|Z_i)[1 - p(Z_i)]\Big\}^{1-R_i}\Big] \tag{7}$$

with $g_1(x|z)$ the density/probability mass function corresponding to $G_1(x|z)$ defined in Equation (4). In the situations where none of the quantities identified in Equation (4) is known, it is hard to maximize the likelihood function in Equation (7) to obtain an estimator of $\boldsymbol{\theta}$. This consideration motivated the approach to estimation that we describe in the next section.

3. LIKELIHOOD-BASED SEMIPARAMETRIC ESTIMATION

In this section we outline a likelihood-based procedure for estimating the parameter $\boldsymbol{\theta}$ using the available data. To facilitate the presentation, we begin with SPMLE assuming that $G_0(x|z)$ in Equation (4), the distribution of the missing X_i given $Z_i = z$, is known. We employ the EM algorithm to compute the SPMLE. We then present our proposed estimator for the model parameter $\boldsymbol{\theta}$, a pseudo-SPMLE, using supplementary information on $G(x|z)$, the overall conditional distribution of X. Finally, we establish the consistency and weak convergence of the pseudo-SPMLE and identify a consistent estimator of its variance.

3.1. SPMLE with Known $G_0(x|z)$

The observed data likelihood function identified in Equation (7) is proportional to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{B(U_i, \Delta_i, X_i, Z_i|\boldsymbol{\theta})\right\}^{R_i} \left\{ \int B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta}) dG_0(x|Z_i) \right\}^{1-R_i}. \tag{8}$$

Following the notion of the nonparametric maximum likelihood estimator of Kiefer and Wolfowitz (1956), we focus on right continuous $\Lambda_0(\cdot)$ with jumps only at observed event times, and we maximize the observed data likelihood function given in Equation (8) over the parameter space

$$\mathcal{H} = \Big\{\boldsymbol{\theta} = \{\alpha, \beta, \gamma, \Lambda_0\} \,:\, (\alpha, \beta, \gamma) \in \Theta_0, \Lambda_0(\cdot) \text{ is an increasing step function in } [0, \tau] \text{ with jumps at the observed failure times and } \Lambda_0(0) = 0\Big\},$$

where τ is conventionally chosen in practical settings as a sufficiently large finite number such that $\max_i\{U_i\} < \tau$.

Let the jump size of $\Lambda_0(\cdot)$ at t be $\lambda_0\{t\}$, and let the distinct values of the observed event times be ordered as $0 < U_{(1)} < \cdots < U_{(J)} < \infty$. The subcomponent Λ_0 of $\boldsymbol{\theta}$ is fully determined by the finite-dimensional vector $\underline{\lambda}_0$ with the components $\lambda_0\{U_{(j)}\}, j = 1, \ldots, J$:

$$\Lambda_0(t) = \sum_{j:\, U_{(j)} \le t} \lambda_0\{U_{(j)}\} = \mathbf{1}_t' \underline{\lambda}_0 \text{ for } 0 < t < \tau,$$

where $\mathbf{1}_t$ is the J-dimensional vector such that its first J_t components are 1 and the remaining $J - J_t$ components are 0 with J_t the size of $\{j \,:\, U_{(j)} \le t\}$. This yields the following procedure for estimating $\boldsymbol{\theta}$ by maximizing $L(\boldsymbol{\theta})$ with respect to the finite number of unknown parameters, namely α, β, γ and $\underline{\lambda}_0$.

Note that when the observation indicator $R = 1$, $\left[U, \Delta, \eta|RX, R, Z\right]$ under the Cox cure model is proportional to $\left\{\lambda(U|X,Z)^{\Delta} S(U|X,Z)\pi(X,Z;\gamma)\right\}^{\eta}\left\{\bar{\pi}(X,Z;\gamma)\right\}^{1-\eta}$ and when $R = 0$ it is proportional to $\left\{ \int \lambda(U|x,Z)^{\Delta} S(U|x,Z) dG_0(x|Z)\pi_0(Z;\gamma)\right\}^{\eta}\left\{\bar{\pi}_0(Z;\gamma)\right\}^{1-\eta}$ with $\pi_0(z;\gamma) = \int \pi(x,z;\gamma) dG_0(x|z)$ and $\bar{\pi}_0(Z;\gamma) = 1 - \pi_0(z;\gamma)$, the probability of $\eta = 1$ conditional on $R = 0$ and $Z = z$. Thus, the log-likelihood function of $\boldsymbol{\theta}$ based on the observed data augmented by $\underline{\eta} = \{\eta_i \,:\, i = 1, \ldots, n\}$ is $l_C(\boldsymbol{\theta}|\underline{\eta}) = l_{C1}(\gamma; \underline{\eta}) + l_{C2}(\alpha, \beta, \Lambda_0; \underline{\eta})$, where $l_{C1}(\gamma; \underline{\eta})$ and $l_{C2}(\alpha, \beta, \Lambda_0; \underline{\eta})$ are

$$\sum_{i=1}^{n} \Big[R_i\left\{\eta_i \log[\pi(X_i, Z_i; \gamma)] + (1 - \eta_i) \log[\bar{\pi}(X_i, Z_i; \gamma)]\right\} + (1 - R_i)\left\{\eta_i \log[\pi_0(Z_i; \gamma)] + (1 - \eta_i) \log[\bar{\pi}_0(Z_i; \gamma)]\right\}\Big]$$

and

$$\sum_{i=1}^{n} \eta_i \left\{ R_i \log\left[\lambda(U_i|X_i, Z_i)^{\Delta_i} S(U_i|X_i, Z_i)\right] + (1 - R_i) \log\left[\int \lambda(U_i|x, Z_i)^{\Delta_i} S(U_i|x, Z_i) dG_0(x|Z_i)\right]\right\},$$

respectively. Applying the EM algorithm (Dempster, Laird & Rubin, 1977) yields the following algorithm.

Algorithm 1. *In the kth iteration for $k \geq 1$ with the current estimate of $\boldsymbol{\theta}$ from the previous iteration given by $\boldsymbol{\theta}^{(k-1)} = \{\alpha^{(k-1)}, \beta^{(k-1)}, \Lambda_0^{(k-1)}, \gamma^{(k-1)}\}$ and an initial estimate $\boldsymbol{\theta}^{(0)}$:*

E-Step. *Calculate* $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) = E\left\{ l_C(\boldsymbol{\theta}; \underline{\eta}); \boldsymbol{\theta}^{(k-1)} \right\}$.

M-Step. *Maximize* $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ *with respect to* $\boldsymbol{\theta}$ *to obtain the updated estimate* $\boldsymbol{\theta}^{(k)}$.

Repeat alternating E- and M-Steps until the sequence $\{\boldsymbol{\theta}^{(k)} : k = 1, ...\}$ *converges.*

Under appropriate conditions given in Wu (1983), the limit of the sequence exists and is the SPMLE, denoted by $\widehat{\boldsymbol{\theta}}_n(G_0)$, from the observed data likelihood function $L(\boldsymbol{\theta})$. We now discuss the implementation of these two alternating steps.

Implementation of the E-Step

Note that

$$\left[\eta = 1 | U, \Delta = 0, RX, R, Z\right] = \frac{\left[U, \Delta = 0 | \eta = 1, RX, R, Z\right]\left[\eta = 1 | RX, R, Z\right]}{\left[U, \Delta = 0 | RX, R, Z\right]},$$

and thus the conditional expectation of η, $E\left[\eta | U, \Delta = 0, RX, R, Z; \boldsymbol{\theta}\right]$, is

$$\frac{\left\{S(U|X,Z)\pi(X,Z;\gamma)\right\}^R \left\{ \int S(U|x,Z) dG_0(x|Z) \pi_0(Z;\gamma) \right\}^{(1-R)}}{\left\{S(U|X,Z)\pi(X,Z;\gamma)\right\}^R \left\{ \int S(U|x,Z) dG_0(x|Z) \pi_0(Z;\gamma) \right\}^{(1-R)} + \bar{\pi}(X,Z;\gamma)^R \bar{\pi}_0(Z;\gamma)^{(1-R)}}. \quad (9)$$

Since $l_C(\boldsymbol{\theta}; \underline{\eta})$ depends linearly on η_i, the conditional expectation in the E-Step is $l_C(\boldsymbol{\theta}; \hat{\underline{\eta}}^{(k-1)})$ with the components $\hat{\eta}_i^{(k-1)} = \hat{\eta}_i(\boldsymbol{\theta}^{(k-1)}; G_0)$ given by

$$\hat{\eta}_i(\boldsymbol{\theta}; G_0) = \Delta_i + (1 - \Delta_i) E_i(\eta | \boldsymbol{\theta}; G_0), \quad (10)$$

where $E_i(\eta|\boldsymbol{\theta}; G_0) = E\left[\eta_i | U_i, \Delta_i = 0, R_i X_i, R_i, Z_i; \boldsymbol{\theta}, G_0\right]$ is the realization of Equation (9) using the data from subject i.

Implementation of the M-Step

In many practical situations, the M-Step can be implemented as follows:

M1-Step. Solve the estimating equation $\partial l_{C1}(\gamma; \hat{\underline{\eta}}^{(k-1)}) \big/ \partial \gamma = 0$ to obtain $\gamma^{(k)}$.

M2-Step. Solve the estimating equations $\partial l_{C2}(\alpha, \beta, \Lambda_0; \hat{\underline{\eta}}^{(k-1)}) \big/ \partial(\alpha, \beta, \underline{\lambda}_0) = 0$ to obtain $\alpha^{(k)}$, $\beta^{(k)}$, $\Lambda_0^{(k)}$.

In general, it is straightforward to carry out M1-Step. We adapt the estimation procedure of Cook, Hu & Swartz (2011) to implement M2-Step as outlined below.

Let $Y_i(t) = I(U_i \geq t)$ and $N_i(t) = I(T_i \leq t)\Delta_i$ denote the at-risk indicator and the count of the observed event of individual i at time t, respectively. Note that $dN_i(t) = dt$ when $t = U_i, \Delta_i = 1$ and 0 otherwise. Moreover, let

$$\widehat{X}_i(\boldsymbol{\theta}^\star) = \mathrm{E}\left[X \big| U_i, \Delta_i, R_i = 0, Z_i; \boldsymbol{\theta}^\star\right] = \int x dG\left(x \big| U_i, \Delta_i, R_i, Z_i; \boldsymbol{\theta}^\star\right)\big|_{R_i=0},$$

$$\widehat{e^{\alpha X_i}}(\boldsymbol{\theta}^\star) = \mathrm{E}\Big[e^{\alpha X}\big|U_i, \Delta_i, R_i = 0, Z_i; \boldsymbol{\theta}^\star\Big] = \int e^{\alpha x} dG\big(x\big|U_i, \Delta_i, R_i, Z_i; \boldsymbol{\theta}^\star\big)\big|_{R_i=0},$$

and

$$\widehat{X_i e^{\alpha X_i}}(\boldsymbol{\theta}^\star) = \mathrm{E}\Big[Xe^{\alpha X}\big|U_i, \Delta_i, R_i = 0, Z_i; \boldsymbol{\theta}^\star\Big] = \int xe^{\alpha x} dG\big(x\big|U_i, \Delta_i, R_i, Z_i; \boldsymbol{\theta}^\star\big)\big|_{R_i=0},$$

where $\boldsymbol{\theta}^\star$ is an index, $dG(x\big|U_i, \Delta_i, R_i, Z_i; \boldsymbol{\theta})|_{R_i=0}$ is the distribution of X conditional on the observed data with given $\boldsymbol{\theta}$, that is, the ratio of $B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta})dG_0(x|Z_i)$ and $\int B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta})$ $dG_0(x|Z_i)$. Note that when X is a binary variable, $\int B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta})dG_0(x|Z_i)$ can be calculated using

$$\int B(U_i, Z_i, \Delta, x|\boldsymbol{\theta})dG_0(x|Z_i) = B(U_i, Z_i, \Delta_i, x|\boldsymbol{\theta})|_{x=1}P(X_i = 1|Z_i, R_i = 0)$$
$$+B(U_i, Z_i, \Delta_i, x|\boldsymbol{\theta})|_{x=0}P(X_i = 0|Z_i, R_i = 0).$$

In general, one can use the sample average $\sum_{j=1}^{J} B(U_i, Z_i, \Delta_i, X^{(j)}|\boldsymbol{\theta})\big/J$ with $X^{(1)}, \ldots, X^{(J)}$ generated from $G_0(\cdot|Z_i)$ to calculate $\int B(U_i, \Delta_i, x, Z_i|\boldsymbol{\theta})dG_0(x \mid Z_i)$.

Denote

$$D_n^{(0)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}^\star; G_0)\Big[R_i e^{\alpha X_i} + (1 - R_i)\widehat{e^{\alpha X_i}}(\boldsymbol{\theta}^\star)\Big]e^{\beta Z_i},$$

$$D_n^{(1\alpha)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}^\star; G_0)\Big[R_i X_i e^{\alpha X_i} + (1 - R_i)\widehat{X_i e^{\alpha X_i}}(\boldsymbol{\theta}^\star)\Big]e^{\beta Z_i},$$

and

$$D_n^{(1\beta)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}^\star; G_0)\Big[R_i e^{\alpha X_i} + (1 - R_i)\widehat{e^{\alpha X_i}}(\boldsymbol{\theta}^\star)\Big]Z_i e^{\beta Z_i}.$$

With $\boldsymbol{\theta}^\star = \boldsymbol{\theta}^{(k-1)}$, the estimating equations in the M2-Step can be written as

$$\Lambda_0(t; \alpha, \beta) = \sum_{i=1}^{n}\int_0^t \frac{Y_i(s)dN_i(s)}{nD_n^{(0)}(s; \alpha, \beta|\boldsymbol{\theta}^\star, G_0)}, \quad 0 < t \le \tau, \tag{11}$$

$$\sum_{i=1}^{n}\int_0^\tau Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}^\star; G_0)\left[\big\{R_i X_i + (1 - R_i)\widehat{X}_i(\boldsymbol{\theta}^\star)\big\} - \frac{D_n^{(1\alpha)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0)}{D_n^{(0)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0)}\right]dN_i(t) = 0, \tag{12}$$

and

$$\sum_{i=1}^{n}\int_0^\tau Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}^\star; G_0)\left[Z_i - \frac{D_n^{(1\beta)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0)}{D_n^{(0)}(t; \alpha, \beta|\boldsymbol{\theta}^\star, G_0)}\right]dN_i(t) = 0. \tag{13}$$

The well-established procedure for estimating the baseline hazard function and the regression parameters in the Cox PH model can be applied to compute the updated estimates of α, β and Λ_0, which yields the M2-Step. Specifically, $\alpha^{(k)}$ and $\beta^{(k)}$ are obtained by solving Equations (12)

and (13), which are extended partial likelihood estimating equations, and substituting them into Equation (11) to obtain $\Lambda_0^{(k)}(\cdot)$.

The resulting SPMLE for the cumulative baseline hazard function is an extended Breslow estimator:

$$\hat{\Lambda}_{0n}(t) = \sum_{i=1}^{n} \int_0^t \frac{Y_i(s)dN_i(s)}{nD_n^{(0)}(s; \hat{\alpha}_n, \hat{\beta}_n|\hat{\theta}_n, G_0)}, \quad 0 < t \le \tau. \tag{14}$$

When there is no missing covariate, the estimating equations (11), (12) and (13) reduce to the estimator presented in Sy & Taylor (2000) for the baseline hazard function and the regression parameters in the Cox PH model for the susceptible individuals' event times. Moreover, they lead to the well-established Breslow estimator and partial likelihood estimating equations in the case where the Cox cure model reduces to the Cox PH model.

3.2. Pseudo Semiparametric Maximum Likelihood Estimation

Assuming that $G_0(x|z)$, the distribution of X given $R = 0, Z = z$, is known is likely to be impractical. Thus, we consider semiparametric estimation in the following two practical situations.

Procedure of pseudo-SPMLE

One may choose to specify $G_0(x|z) = G_0(x|z; \phi)$ up to a finite-dimensional parameter ϕ. The SPMLE in Section 3.1 can be adapted straightforwardly with $\theta = \{\alpha, \beta, \Lambda_0, \gamma, \phi\}$, having an additional component ϕ. Specifically, in the kth iteration of the EM algorithm, we use $G_0(x|z; \phi^{(k-1)})$ instead of $G_0(x|z)$, and there is an additional step after the M2-Step, to obtain the updated estimate $\phi^{(k)}$:

M3-Step. When neither $p(z)$ nor $G_1(x|z)$ in Equation (4) contains any information on ϕ, solve $\partial l_C(\theta^{(k)}, \phi; \underline{\hat{\eta}}^{(k-1)})\big/\partial\phi = 0$; otherwise, without loss of generality, assuming all the terms in Equation (4) are functions of ϕ, solve $\partial l_{CF}(\theta^{(k)}, \phi; \underline{\hat{\eta}}^{(k-1)})\big/\partial\phi = 0$. Here $l_{CF}(\alpha, \beta, \Lambda_0, \phi; \underline{\eta}) = l_C(\theta^{(k)}, \phi|\text{Observed-Data}; \underline{\eta}) + \sum_{i=1}^{n} \left[R_i \log\left\{g_1(X_i|Z_i)p(Z_i; \phi)\right\} + (1 - R_i)\log\left\{1 - p(Z_i; \phi)\right\}\right]$, which is based on Equation (7) while $l_C(\cdot)$ is based on Equation (8).

The parametric model specification of $G_0(x|z; \phi)$ concerns the distribution of unobserved X_i. Thus, the model cannot be determined using the observed data. Moreover, the associated estimation procedure includes an additional layer of computing intensity to handle the nuisance parameter ϕ. These considerations motivate the following alternative approach, a pseudo-SPMLE approach.

With supplementary information on $G(x|z)$

We assume that there is a consistent estimator for $G_0(\cdot|z)$, denoted $\tilde{G}_0(\cdot|z)$. Often there could be information that is readily available concerning the overall conditional distribution $G(x|z)$ for the population. In the TB study, for example, there are well-documented population proportions of HIV infection within various subpopulations. Following the approach of Cook, Hu & Swartz (2011), we obtain $\tilde{G}_0(\cdot|z)$ based on Equation (4) if $G(\cdot|z)$ is known:

$$\tilde{G}_0(x|z) = \frac{G(x|z) - \hat{p}(z)\hat{G}_1(x|z)}{1 - \hat{p}(z)},$$

where $\hat{p}(z)$ and $\hat{G}_1(x|z)$ are consistent estimators of $p(z)$ and $G_1(\cdot|z)$, say, the MLE, with the data $\{(R_i, Z_i) : i = 1, \ldots\}$ and $\{(X_i, Z_i) : R_i = 1; i = 1, \ldots\}$ from the observed data, respectively. In the case where $G(\cdot|z)$ is unknown and X is a binary variable taking values 1 and 0, we can estimate $G(x|z)|_{x=1}$ by $P(Z = z|X = 1)P(X = 1)/P(Z = z)$.

We propose to substitute $\tilde{G}_0(\cdot|z)$ into the SPMLE $\widehat{\boldsymbol{\theta}}_n(G_0)$ obtained in Section 3.1. Let the resulting estimator $\widehat{\boldsymbol{\theta}}_n(\tilde{G}_0)$ be $\widetilde{\boldsymbol{\theta}}_n = \{\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\Lambda}_{0n}(\cdot), \tilde{\gamma}_n\}$.

Propostion 1. *Assume conditions C1–C5 listed in the Appendix. The pseudo-SPMLE* $\widetilde{\boldsymbol{\theta}}_n$ *exists almost surely, and satisfies* $\tilde{\alpha}_n \to \alpha$, $\tilde{\beta}_n \to \beta$, $\tilde{\gamma}_n \to \gamma$, *and* $\sup_{t\in[0,\tau]} |\tilde{\Lambda}_{0n}(t) - \Lambda_0(t)| \to 0$ *almost surely as* $n \to \infty$.

Propostion 2. *Assume conditions C1–C5 listed in the Appendix. The pseudo-SPMLE,* $\tilde{\alpha}_n, \tilde{\beta}_n$ *and* $\tilde{\gamma}_n$ *are asymptotically efficient, and* $\sqrt{n}\big(\tilde{\alpha}_n' - \alpha', \tilde{\beta}_n' - \beta', \tilde{\gamma}_n' - \gamma', \tilde{\Lambda}_{0n} - \Lambda_0\big)'$ *converges weakly to a zero-mean Gaussian process in the metric space* $\mathcal{R}^d \times l^\infty[0, \tau]$, *where* $\mathcal{R} = (-\infty, \infty)$, *d is the dimension of* $(\alpha', \beta', \gamma')'$ *and* $l^\infty[0, \tau]$ *is the linear space with all the bounded functions on* $[0, \tau]$ *and equipped with the supremum norm.*

Proofs of these two propositions are outlined in the accompanying Appendix.

Furthermore, let $L_{Oi}(\boldsymbol{\theta}; G_0)$ be the ith term in Equation (8) for $i = 1, \ldots, n$; thus, the log-transformed $L(\boldsymbol{\theta})$ in Equation (8) is $l_O(\boldsymbol{\theta}; G_0) = \sum_{i=1}^n \log L_{Oi}(\boldsymbol{\theta}; G_0)$. Let $\underline{l}_O(\boldsymbol{\theta}; G_0) = (\log L_{O1}(\boldsymbol{\theta}; G_0), \ldots, \log L_{On}(\boldsymbol{\theta}; G_0))^T$. Using Theorem 3.2 in White (1982) as a template, we have the following proposition:

Propostion 3. *The variance matrix of the pseudo-SPMLE* $\big(\tilde{\alpha}_n', \tilde{\beta}_n', \tilde{\gamma}_n'\big)'$ *is estimated consistently by* $\Pi(\widetilde{\boldsymbol{\theta}}_n)$, *where*

$$\Pi(\boldsymbol{\theta}) = \left[-\frac{\partial^2 l_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial(\alpha, \beta, \gamma)^2}\right]^{-1} \left(\frac{\partial \underline{l}_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial(\alpha, \beta, \gamma)}\right) \left(\frac{\partial \underline{l}_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial(\alpha, \beta, \gamma)}\right)' \left[\left(-\frac{\partial^2 l_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial(\alpha, \beta, \gamma)^2}\right)'\right]^{-1}. \quad (15)$$

The variance of $\tilde{\Lambda}_{0n}(t)$, *which is* $\sum_{j:U_{(j)}\le t} \tilde{\lambda}_{0n}\{U_{(j)}\} = \mathbf{1}_t' \widetilde{\underline{\lambda}}_0$ *for* $0 < t \le \tau$, *is estimated consistently by* $\mathbf{1}_t' \Psi(\widetilde{\boldsymbol{\theta}}_n)\mathbf{1}_t$ *with*

$$\Psi(\boldsymbol{\theta}) = \left[-\frac{\partial^2 l_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial \underline{\lambda}_0^2}\right]^{-1} \left(\frac{\partial \underline{l}_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial \underline{\lambda}_0}\right) \left(\frac{\partial \underline{l}_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial \underline{\lambda}_0}\right)' \left[\left(-\frac{\partial^2 l_O(\boldsymbol{\theta}; \tilde{G}_0)}{\partial \underline{\lambda}_0^2}\right)'\right]^{-1}. \quad (16)$$

The variance estimator given in Equation (15) is the robust sandwich variance estimator commonly used for the pseudo-MLE of a finite-dimensional parameter. When $\widetilde{\boldsymbol{\theta}}_n$ reduces to the SPMLE, $\Pi(\boldsymbol{\theta})$ is reduced to the first term on the right-hand side of Equation (15), a consistent estimator of the variance by the inverse of the observed Fisher information matrix. The variance estimator could be computationally intensive because it requires the calculation of inverses of large dimensional matrices. Practitioners may prefer a resampling-based variance estimation procedure.

4. EMPIRICAL STUDIES

In this section we illustrate our proposed approach and its application to disease screening/prediction using the TB study data. We then report a simulation study that we conducted to evaluate the findings from our analysis of the study data and examine the finite-sample performance of the approach.

TABLE 1: Summary information concerning the BCCDC-TB study (Cook et al. 2005)

Part A. List of potential risk factors with their categories in *italics*	
HIV status	*Positive, negative*
Gender	*Male, female*
Age at contact	*Continuous variable in years*
Level of contact	*Casual, nonhousehold, household*
Country of birth	*Canada, foreign countries*
Source type	*Cluster-status of the TB source case or not*

Part B. Summary of the observed times[a]: $U_i = T_i \wedge C_i$					
Percentile	5%	25%	50%	75%	95%
Whole data set	1,778	2,163	2,576	3,003	3,193
Observed TB times (U_i's with $\Delta_i = 1$)	3.1	11	81	354.5	2,061
Censoring times (U_i's with $\Delta_i = 0$)	1,778	2,163	2,592	3,003	3,193

[a]Elapsed days since the subject's physical contact with a TB case.

4.1. Analysis of the TB Contacts Study Data

Study description

TB is an airborne infectious disease. Many risk factors increase a person's chance of being infected or progressing to develop active disease. HIV infection is a very important risk factor. People living with HIV are 19 times (95% confidence interval [CI] (15, 22)) more likely to develop active TB disease than people without HIV (World Health Organization, 2018). The aforementioned study by BCCDC aimed to evaluate the association of the time to TB and latent TB infection with a list of potential risk factors. It focused on subjects in the Greater Vancouver area who had physical contact with active infectious TB patients (Cook et al., 2005). The identified potential risk factors are listed in Table 1. The study investigators expected heavy right censoring with respect to the observation of the study endpoint, the time from physical contact to TB onset. In addition, only 2.2% of the study subject had records of HIV status because of the study's data-collection mechanism. Viewing the data missing as MNAR, Cook, Hu & Swartz (2011) analyzed the study data using the Cox PH model and relevant population information on HIV infection. They provided explanations for some of the puzzling features in the initial analysis. However, their estimated survivor curve featured a long, literally unchanging right tail; see Table 1B and Figure 1 in Section A of the accompanying Supplementary Material. Moreover, one of their reported findings was that younger subjects had a higher risk of TB development, which is rather counterintuitive. In studies of infectious disease, especially those concerned with disease screening such as the TB study, there is often a substantial portion of nonsusceptible subjects in the study who would never experience the disease. The objectives of such a study are usually to study the cure rate, the survivor distribution of the time to disease and covariate effects. Conventional survival analysis which treats all the subjects as susceptible can be inappropriate, which motivated us to consider a two-component mixture cure model. We assumed that the time to TB onset since physical contact among susceptible subjects follows a Cox PH model. Using the method of estimation that we proposed in Section 3, we analyzed the data from the TB study subjects who had values recorded for all the covariates of interest except HIV status, taking them as realizations of the observed data with $n = 7{,}754$ and X_i being the HIV infection status of subject i. There were 63 subjects with observed times to TB onset.

Data analysis using our proposed approach

Following Cook, Hu & Swartz (2011) and Guo, Hu & Liu (2017), we used available HIV prevalence information to estimate the overall conditional distribution $G(\cdot|z)$, which is determined by $P(X = 1|Z = z)$. We assumed that the conditional distribution of unobserved HIV status $G_0(x|Z)$ depended only on *age* and *gender*, which we denote by Z^*, a subcomponent vector of the covariates Z. Specifically, we considered three age groups: young (0–29), middle-aged (30–49) and old (50 or above). According to the posted information in *HIV/AIDS Epi Updates* (2007, Public Health Agency of Canada), the proportions of HIV infection in these groups were $(0.177, 0.487, 0.112)$ for males and $(0.082, 0.124, 0.012)$ for females. With the overall HIV infection rate $P(X = 1)$, we then obtained the population rates $P(X = 1|Z^* = z^*)$ using $P(Z^* = z^*|X = 1)P(X = 1)/P(Z^* = z^*)$. From Equation (4), we estimated $G_0(x|Z^*)$ using

$$\tilde{G}_0(x|z^*) = \{G(x|z^*) - \hat{p}(z^*)\hat{G}_1(x|z^*)\}/[1 - \hat{p}(z^*)],$$

where $\hat{p}(z^*) = n_1(z^*)/n(z^*)$ and $\hat{G}_1(x|z^*) = \sum_i R_i I(X_i \le x, Z_i^* = z^*)/n_1(z^*)$ with $n_1(z^*) = \sum_i R_i I(Z_i^* = z^*)$ and $n(z^*) = \sum_i I(Z_i^* = z^*)$ based on the available data. In our analysis we followed Cook, Hu & Swartz (2011) and used $P(X = 1) = 0.1\%, 0.3\%$ and 1% .

Table 2 summarizes the estimates of the regression parameters (α, β, γ) in the Cox cure model. For comparison, it also shows the estimates of (α, β) in the Cox PH model from the report of Cook, Hu & Swartz (2011), together with regression estimates derived for the two competing models using the data of the complete cases. The estimated standard errors in the table are obtained using the robust sandwich variance estimator identified in Equation (15). The significant effects are presented in boldface. The identified sets of important risk factors for TB time, based on the estimates of α and β in our analysis that assumed the Cox cure model agree with those effects that we identified using the Cox PH model after we adjust for the MNAR HIV status. Moreover, they were also consistent across the different population prevalence rates that we considered.

The estimates of the regression parameter γ in the probability model for susceptibility $P(\eta = 1|X, Z)$ from our approach provided additional insights. For example, they confirmed that HIV-infected individuals were much more susceptible to TB infection. Our findings also suggested that, while the type of source TB was significantly associated with the TB onset time, it was not significant for identifying susceptibility to TB. Moreover, our analysis indicated that older TB contacts were more susceptible to TB infection, and younger TB contacts, if susceptible, developed TB faster. This explains the surprising finding from Cook, Hu & Swartz (2011) about the relationship between TB development and age at contact. An additional finding concerned the nature of the association between susceptibility to TB and the type of disease exposure that was experienced. It appears that nonhousehold contact leads to the highest probability of susceptibility to TB. Household contact resulted in intermediate susceptibility, and casual contact led to the lowest. On the other hand, if susceptible, household contacts led to TB onset faster than those that occurred outside the household surroundings.

Figure 1 shows the estimated cumulative baseline hazard curves under the Cox PH model and the Cox PH cure model. The much lower baseline hazards for the Cox model resulted from the large proportion of nonsusceptible subjects. This suggests that our approach may help to improve the effectiveness of current screening practices and prediction of the time to TB onset.

Risk prediction based on our proposed approach

We now showcase two applications of our proposed approach for risk prediction. We consider the following two procedures for identifying susceptible individuals, with a predetermined cut-off level $c > 0$.

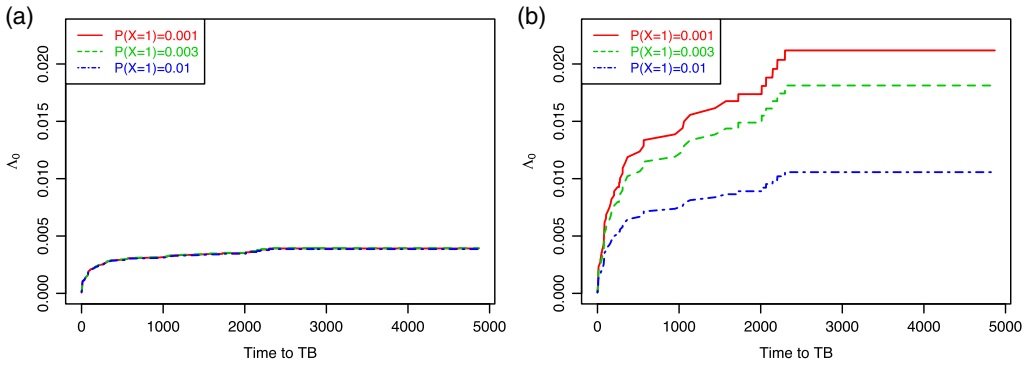

FIGURE 1: Estimated cumulative baseline hazards functions for the BCCDC TB contacts study data using either (a) the Cox PH model or (b) the Cox PH cure model.

- By $\tilde{\pi}_i$: subject i is classified as susceptible to TB if $\tilde{\pi}_i > c$, where

$$\tilde{\pi}_i = \begin{cases} \pi(X_i, Z_i; \tilde{\gamma}_n) = P(\eta = 1 | X_i, Z_i; \tilde{\gamma}_n) & \text{if } R_i = 1, \\ \tilde{\pi}_0(Z_i; \tilde{\gamma}_n) = \int \pi(x, Z_i; \tilde{\gamma}_n) d\tilde{G}_0(x | Z_i) & \text{if } R_i = 0. \end{cases}$$

- By $\tilde{\eta}_i$: subject i is classified as susceptible to TB if $\tilde{\eta}_i > c$, where $\tilde{\eta}_i = \hat{\eta}_i(\tilde{\theta}_n; \tilde{G}_0)$, is derived using Equation (10) from the E-Step of the last loop in the EM algorithm described in Section 3.

For illustration, Table 1 in Section A of the Supplementary Materials shows the predicted counts of susceptible and nonsusceptible individuals in the TB study with $c = 0.25$ and an assumed population HIV infection rate of 0.3%. We chose the cut-off level $c = 0.25$ to illustrate a common practice in disease screening, namely adopting a conservative strategy to avoid classifying susceptible subjects into the nonsusceptible group. Clearly, the screening based on $\tilde{\eta}_i$ is more efficient since it uses not only the baseline covariates X_i and Z_i but also the information (U_i, Δ_i) from subject i. On the other hand, the procedure based on $\tilde{\pi}_i$ is applicable to situations when individuals do not have any follow-up information on T_i. We also found that the procedure based on $\tilde{\pi}_i$ exhibited satisfactory performance; it incorrectly classified only 2 of 63 subjects with $\Delta_i = 1$, that is, with observed $\eta_i = 1$, into the nonsusceptible group. The next section confirms by simulation these observations on the performance of the two prediction methods.

Further, estimated survivor curves $S(\cdot | X, Z; \tilde{\alpha}, \tilde{\beta})$ can be used to predict the time to TB onset for a susceptible individual according to his/her risk factors. See Figure 2 for the estimated survivor curves for different subgroups.

4.2. Simulation Study

We conducted a simulation study to support the findings from our analysis of the TB contacts study. Two simulation settings were considered. In *Simulation A*, the missing covariate X was generated depending on a latent variable M; in *Simulation B*, both Z and M were involved.

Simulation A

We simulated a cohort with $n = 1{,}000$ independent subjects and generated the data for subject i as follows:

(i) Generate $Z_{i1} \sim N(0, 1)$ and $Z_{i2} \sim B(1, 0.6)$ independently.

TABLE 2: Estimated regression coefficients and corresponding estimated standard errors for the BCCDC TB contacts study data; effects that were statistically significant, that is, $p \leq 0.05$, are marked with boldface type.

	Cox PH model				Cox PH cure model			
	Complete	HIV infection rate			Complete	HIV infection rate		
Factor	cases	.1%	.3%	1%	cases	.1%	.3%	1%
Estimates of α, β in $\lambda(\cdot\|X, Z)$								
HIV status	.609	**3.403**	**3.372**	**2.966**	−1.510	**2.750**	**2.711**	**3.086**
(infected vs. not)	(.650)	(.490)	(.495)	(.473)	(1.065)	(.462)	(.502)	(.447)
Gender	.174	.137	.123	.050	−.134	**.496**	**.437**	**.706**
(male vs. female)	(.456)	(.256)	(.256)	(.262)	(.421)	(.259)	(.258)	(.274)
Age at contact	−.017	**−.024**	**−.025**	**−.025**	**−.061**	**−.067**	**−.066**	**−.073**
(in years)	(.018)	(.010)	(.011)	(.011)	(.025)	(.011)	(.011)	(.012)
Birth country	**1.197**	**.493**	**.496**	**.550**	−.349	**.711**	**.727**	**.924**
(foreign vs. Can.)	(.568)	(.290)	(.290)	(.292)	(.668)	(.333)	(.332)	(.343)
Contact1	.610	**.931**	**.929**	**.909**	.603	.312	.331	.229
(nonhousehold vs. casual)	(.609)	(.410)	(.410)	(.403)	(.983)	(.433)	(.433)	(.433)
Contact2	.563	**2.466**	**2.468**	**2.494**	1.067	**2.841**	**2.979**	**2.919**
(household vs. casual)	(.602)	(.358)	(.359)	(.364)	(.806)	(.384)	(.384)	(.378)
Source type	.437	**1.253**	**1.263**	**1.360**	**3.464**	**1.867**	**1.920**	**1.873**
(clustered vs. not)	(.543)	(.299)	(.299)	(.283)	(.865)	(.356)	(.365)	(.365)
Estimates of γ in $P(\eta = 1\|X, Z)$								
Intercept					−1.140	**−3.844**	**−3.705**	**−3.274**
					(.995)	(.534)	(.513)	(.342)
HIV status					.890	**3.301**	**3.437**	**2.465**
(infected vs. not)					(1.310)	(.535)	(.549)	(.563)
Gender					.101	−.180	−.117	−.265
(male vs. female)					(.561)	(.264)	(.242)	(.292)
Age at contact					.006	**.019**	**.017**	**.031**
(in years)					(.035)	(.009)	(.008)	(.005)
Birth country					1.287	1.108	1.094	1.347
(foreign vs. Can.)					(.785)	(.164)	(.165)	(.180)
Contact1					1.015	**1.309**	**1.334**	**1.216**
(nonhousehold vs. casual)					(.941)	(.305)	(.300)	(.277)
Contact2					.249	**.822**	**.688**	**.456**
(household vs. casual)					(.963)	(.309)	(.288)	(.272)
Source type					−.683	.212	.114	.147
(clustered vs. not)					(.795)	(.262)	(.268)	(.278)

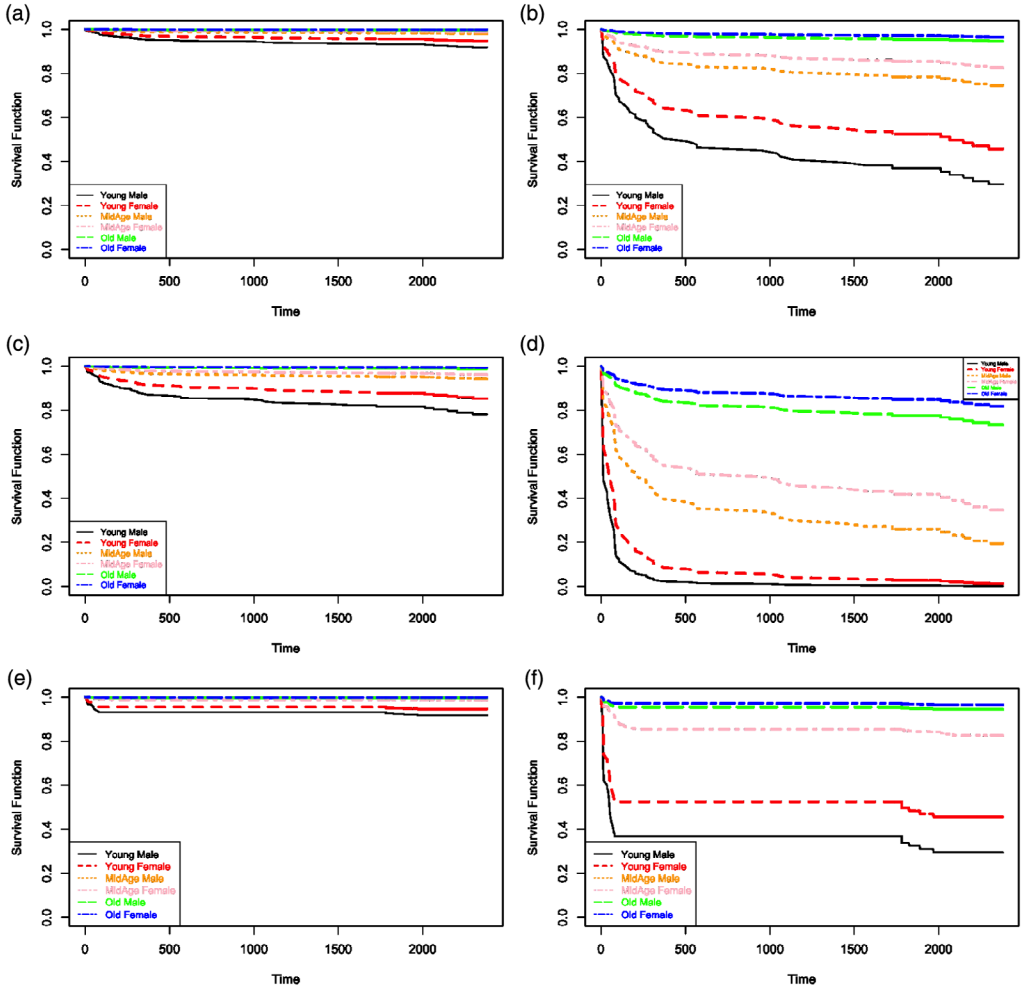

FIGURE 2: Estimated survivor functions for various subgroups of the BCCDC TB contacts study data; (a) HIV negative, nonhousehold contact; (b) HIV negative, household contact; (c) HIV positive, nonhousehold contact; (d) HIV positive, household contact; (e) HIV status missing, nonhousehold contact; (f) HIV status missing, household contact.

(ii) Generate $M_i \sim B(1, 0.1)$, the Bernoulli distribution with probability of success 0.1; generate $X_i \sim B(1, q_X(M_i))$ with $logit\{q_X(m)\} = 2m$; and then generate the missingness indicator $R_i \sim B(1, q_R(X_i, D_i))$ with $logit\{q_R(x, D_i)\} = -3 - 0.1D_i$, $0.2 + 0.1D_i$ for $x = 1, 0$, respectively, where $D_i \sim U(0, 1)$. The intent of this procedure was to simulate an MNAR mechanism with an overall missingness rate of 70%.

(iii) With the generated data (X_i, Z_{1i}, Z_{2i}), generate $\eta_i \sim B(1, q_\eta)$ with $logit(q_\eta) = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_{1i} + \gamma_3 Z_{2i}$, and generate T_i from the Cox PH cure model in Equation (1) with $\lambda(t|X_i, Z_i) = \lambda_0(t) \exp(\alpha X_i + \beta_1 Z_{1i} + \beta_2 Z_{2i})$.

(iv) Generate the censoring times $C_i \sim U(\tau/2, \tau)$ with τ chosen to give a right censoring rate of 90%.

In the simulation settings, $G_0(\cdot)$ is fully determined by $G_0(x|z_1, z_2, m) = G_0(x|m) = P(X = x|R = 0, M = m)$ with $x = 0, 1$ and $m = 0, 1$. Note that $G_0(x|m) = \{G(x|m) - P(R = 1|$

$M = m)G_1(x|m)\}/[1 - P(R = 1|, M = m)]$ with $G(x|m) = P(X = x|M = m)$. We considered three methods of estimating estimate $G(x|m)$ and then obtain an estimate of $G_0(\cdot)$ in Equation (4):

Method A1. By $\tilde{G}_{0A1}(\cdot)$: Use the true $G(x|m)$. This simulates the situation where $G_0(\cdot)$ is unknown but the overall conditional distribution is known. We estimated $P(R = 1|M = m)$ using the sample proportion of $R = 1$ based on the generated R_i's in the group with $M = m$, that is, $\hat{p}_1(m) = \sum_{i:R_i=1} I(M_i = m)\big/ \sum_i I(M_i = m)$, and we estimated $G_1(x|z_1, z_2, m) = G_1(x|m)$ using the sample proportion of $RX = x$ based on the generated R_iX_i's in the group with $M = m$, that is, $\hat{G}_1(x|m) = \sum_{i:R_i=1} I(X_i \leq x, M_i = m)\big/ \sum_{i:R_i=1} I(M_i = m)$.

Method A2. By $\tilde{G}_{0A2}(\cdot)$: Use $\tilde{G}(x) = \sum_m G(x, m)$ to estimate $G(x|m)$. This simulates the situation where $G_0(\cdot)$ is known only marginally.

Method A3. By $\tilde{G}_{0A3}(\cdot)$: Generate an additional dataset and estimate $G(x|m)$ by $\tilde{G}(x|m) = \sum_{i=1}^{n} I(X_i = x, M_i = m)\big/ \sum_{i=1}^{n} I(M_i = m)$ from that dataset. This simulates the situation where $G_0(\cdot)$ can be estimated using additional data.

The simulation used $\alpha = 2.0$, $\beta_1 = -0.5$, $\beta_2 = 1.0$, $\gamma_0 = -2.0$, $\gamma_1 = 0.4$, $\gamma_2 = -1.5$ and $\gamma_3 = 0.8$, based on the regression parameter estimates from the TB study. We evaluated the pseudo-SPMLE $\tilde{\theta}_n$ proposed in Section 3, together with its consistent standard error estimators identified in Equations (15) and (16), respectively, using the generated observed data and each of the three estimators $\tilde{G}_{0A1}(\cdot)$, $\tilde{G}_{0A2}(\cdot)$ and $\tilde{G}_{0A3}(\cdot)$. We also evaluated the SPMLE with the true $G_0(\cdot)$ given in Section 3.1, to provide a point of reference.

Table 3 summarizes the outcomes based on 300 simulation repetitions. The sample means of all three pseudo-SPMLE estimates are close to the true parameter values. This illustrates the consistency of the pseudo-SPMLE estimator. The corresponding sample standard deviations of the estimates are slightly larger than those for the SPMLE estimator using the true $G_0(\cdot)$, confirming the asymptotic efficiency of the pseudo-SPMLE. In addition, the sample means of the estimated standard errors that were derived using the robust estimator are close to the corresponding sample standard deviations, indicating good performance of the variance estimator that we identified in Equation (15).

Figure 3 shows the sample means of the estimated baseline cumulative hazard function under the Cox PH model and the Cox PH cure model together with approximate 95% CIs. All three sets of pseudo-SPMLE estimates are close to the estimated curve for the SPMLE with the true G_0. To attain a censoring rate of 90%, the selected values of τ can be very small and the sample means of the estimated curves remain unchanged when $t \geq \tau$ as no events would be observed after τ. Except for the plateau evident in Figure 3b when $t \geq 0.045$, the resulting estimates were all close to the true $\Lambda_0(\cdot)$ with the Cox PH cure model but noticeably different from the curve estimated from the fitted Cox PH model; see Figure 3a. This is because the simulation assumed the Cox cure PH model and the estimate derived using the Cox PH model which treats all subjects as susceptible would not be expected to perform well. This finding corroborates the remark made by Sy & Taylor (2000): caution is required when interpreting the outcomes from a conventional survival model in a study with a potentially large group of nonsusceptible individuals.

To examine the performance of the two screening procedures that we outlined in Section 4.1, we evaluated $\tilde{\pi}_i$ and $\tilde{\eta}_i$ using (i) the true G_0, (ii) the estimator $\tilde{G}_{0A1}$ and (iii) the full set of generated X_i with different realizations of the simulated *observed data* set. We then carried out the disease screening.

The specificities of the two procedures with $c = 0.25$ were high while their sensitivities achieved reasonable levels. As expected, the outcomes based on $\tilde{\eta}_i$ were better overall than those based on $\tilde{\pi}_i$ in terms of both sensitivity and specificity. This supports the corresponding

TABLE 3: Estimated regression coefficients obtained in *Simulation A* from 300 repetitions with $n = 1{,}000$, 90% right censoring and an overall missingness rate of 70%.

	Cox PH cure model							Cox PH model		
Parameter	α	β_1	β_2	γ_0	γ_1	γ_2	γ_3	α	β_1	β_2
(true value)	2.0	−.5	1.0	−2.0	.4	−1.5	.8	2.0	−.5	1.0
	Using only data with $R = 1$ (the complete-cases data)									
bias	−.182	.042	−.036	−.015	.187	−.033	.006	−.857	.267	.290
sm($\hat{\theta}$)[a]	1.818	−.458	.964	−2.015	.587	−1.533	.806	1.143	−.233	1.290
ssd($\hat{\theta}$)[b]	.713	.351	.545	.361	.453	.188	.289	.742	.356	.645
sm($\widehat{se(\hat{\theta})_F}$)[c]	.764	.349	.546	.358	.455	.183	.285	.708	.351	.643
	With the whole dataset and true G_0									
bias	−.012	−.003	.021	−.017	.008	−.002	.002	.345	−0.401	.344
sm($\hat{\theta}$)	1.988	−.503	1.021	−2.017	.408	−1.502	.802	2.345	−.901	1.344
ssd($\hat{\theta}$)	.257	.133	.231	.353	.352	.176	.275	.412	.135	.300
sm($\widehat{se(\hat{\theta})_F}$)	.250	.128	.233	.351	.350	.171	.271	.386	.131	.276
	With the whole dataset and $\tilde{G}_{0A1}$									
bias	−.013	−.003	.022	−.014	.004	−.011	.000	.444	−.402	.345
sm($\tilde{\theta}$)	1.987	−.503	1.022	−2.014	.404	−1.511	.800	2.444	−.902	1.345
ssd($\tilde{\theta}$)	.270	.138	.253	.353	.348	.175	.274	.415	.137	.300
sm($\widehat{se(\tilde{\theta})_F}$)	.259	.138	.250	.351	.343	.177	.270	.388	.132	.278
sm($\widehat{se(\tilde{\theta})_R}$)[d]	.265	.141	.252	.355	.342	.179	.275	.396	.136	.282
	With the whole dataset and $\tilde{G}_{0A2}$									
bias	−.029	−.014	.028	−.045	.003	−.009	.002	.485	−.596	.341
sm($\tilde{\theta}$)	1.971	−.514	1.028	−2.045	.403	−1.509	.802	2.485	−1.096	1.341
ssd($\tilde{\theta}$)	.275	.135	.258	.353	.345	.176	.267	.411	.135	.298
sm($\widehat{se(\tilde{\theta})_F}$)	.270	.137	.252	.355	.346	.174	.269	.384	.132	.276
sm($\widehat{se(\tilde{\theta})_R}$)	.274	.141	.257	.361	.352	.173	.268	.394	.132	.276
	With the whole dataset and $\tilde{G}_{0A3}$									
bias	−.024	−.002	−.005	.003	.009	.042	−.005	.411	−.593	.339
sm($\tilde{\theta}$)	1.976	−.502	.995	−1.997	.409	−1.458	.795	2.411	−1.093	1.339
ssd($\tilde{\theta}$)	.275	.143	.254	.355	.343	.179	.272	.412	.136	.300
sm($\widehat{se(\tilde{\theta})_F}$)	.270	.137	.252	.352	.342	.175	.270	.385	.132	.274
sm($\widehat{se(\tilde{\theta})_R}$)	.274	.144	.258	.363	.356	.175	.270	.390	.133	.275

[a] sm($\tilde{\theta}$) or sm($\hat{\theta}$): sample mean of the evaluations of $\tilde{\theta}$ or $\hat{\theta}$, the pseudo-SPMLE or the SPMLE.
[b] ssd($\tilde{\theta}$) or ssd($\hat{\theta}$): sample standard deviations of the evaluations of $\tilde{\theta}$ or $\hat{\theta}$.
[c] sm($\widehat{se(\tilde{\theta})_F}$): sample mean of the standard error estimates for $\tilde{\theta}$ by the using the inverse of the observed information matrix.
[d] sm($\widehat{se(\tilde{\theta})_R}$): sample mean of the standard error estimates for $\tilde{\theta}$ obtained using the robust variance estimator identified in (15).

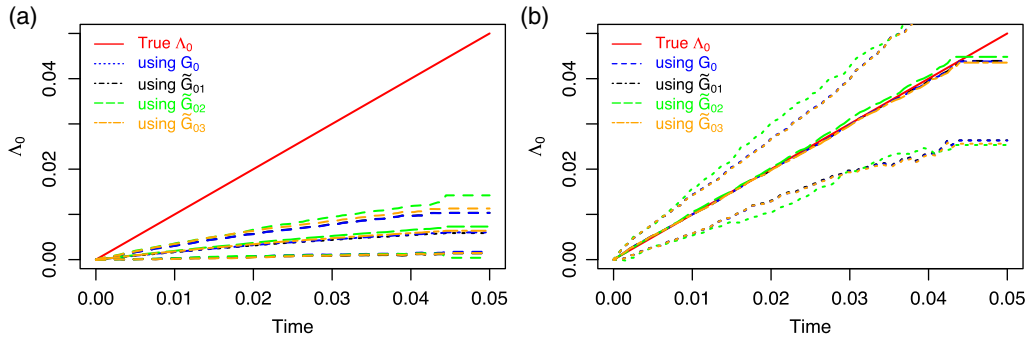

FIGURE 3: Sample means and approximate 95% pointwise confidence limits for $\Lambda_0(\cdot)$ obtained in *Simulation A* for (a) the Cox PH model and (b) the Cox PH cure model.

observation that we identified from our analysis findings for the TB contacts study. On the other hand, the outcomes in Case (iii) were the best overall. Those observed in Case (ii), which mimics certain practical situations, were close to the ones in Case (i). See Table 2 in Section B of the Supplementary Materials for a summary of the outcomes using a single realization of the simulated data set, and for the corresponding sample sensitivities and specificities. Figure 4 shows histograms of the $\tilde{\pi}_i$ and $\tilde{\eta}_i$ estimates in Case (ii) based on three realizations of the simulated data sets; the generated real η_i are included in the histograms, with green indicating $\eta_i = 0$ and pink indicating $\eta_i = 1$.

Simulation B

In this simulation, we generated the realizations of X from the model $X \sim B(1, q_X(M, Z_2))$ with $logit\{q_X(M, Z_2)\} = 0.3M + 0.2Z_2$. This design was intended to mimic a scenario similar to the TB study where supplementary information was available on the overall distribution of X conditional on another covariate Z_2 in the regression model. All the other variables were simulated in the same way as we described for *Simulation A*.

In this particular setting, we assumed $G_0(x|z_1, z_2) = G_0(x|z_2)$, and fixed its values for $m = 0, 1$ by $\sum_m G_0(x|z_2, m)P(M = m)$. We considered two methods of estimating $G(x|z_2)$ and then obtained an estimate of $G_0(x|z_2)$ as outlined in Equation (4):

Method B1. By $\tilde{G}_{0B1}(\cdot)$: Use the true $G(x|z_2)$, which is $\sum_m G(x|z_2, m)P(M = m)$.

Method B2. By $\tilde{G}_{0B2}(\cdot)$: Generate an additional dataset and estimate $G(x|z_2)$ by $\tilde{G}(x|z_2) = \sum_{i=1}^{n} I(X_i = x, Z_{2i} = z_2) \big/ \sum_{i=1}^{n} I(Z_{2i} = z_2)$ from that dataset. This approach corresponded to our actual analysis of the TB study, which relied on population information to estimate G_0.

Table 4 summarizes the outcomes based on 300 simulation repetitions. The pseudo-SPMLE $\tilde{\boldsymbol{\theta}}_n$ with estimators $\tilde{G}_{0B1}(\cdot)$ and $\tilde{G}_{0B2}(\cdot)$ along with their consistent standard error estimators identified in Equations (15) and (16) were evaluated. We also evaluated the SPMLE with the true $G_0(\cdot)$ to provide a reference. The observed outcomes exhibited a pattern similar to the one that we obtained in *Simulation A*: the sample means of both pseudo-SPMLE estimates were close to the true parameter values. Although the sample means of the estimated standard errors that were derived using the robust sandwich estimator were larger than the corresponding values obtained in *Simulation A*, they were nevertheless close to the corresponding sample standard deviations. The sample means of the cumulative baseline hazard estimates under the Cox PH

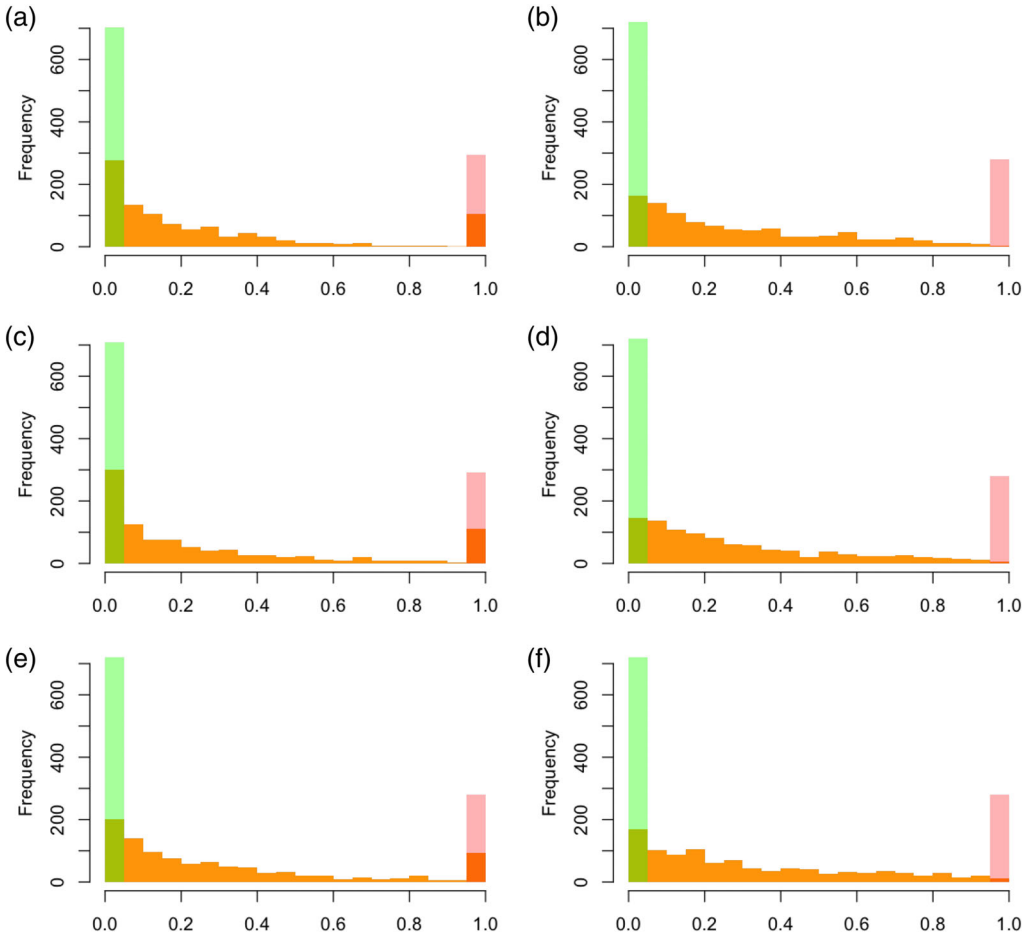

FIGURE 4: Histograms of $\tilde{\eta}_i$ and $\tilde{\pi}_i$ with three simulated datasets using $\tilde{G}_{01}(\cdot)$ in *Simulation A*. (a) Histogram of $\tilde{\eta}_i$ with dataset 1. (b) Histogram of $\tilde{\pi}_i$ with dataset 1. (c) Histogram of $\tilde{\eta}_i$ with dataset 2. (d) Histogram of $\tilde{\pi}_i$ with dataset 2. (e) Histogram of $\tilde{\eta}_i$ with dataset 3. (f) Histogram of $\tilde{\pi}_i$ with dataset 3.

model and the Cox PH cure model together with approximate 95% CIs are displayed in Figure 2 in Supplementary Material. All of the pseudo-SPMLE estimates were close to the estimated curve for the SPMLE with the true G_0. They were all close to the true $\Lambda_0(\cdot)$ with the Cox PH cure model but different from the estimated curve based on the Cox PH model. This is consistent with our findings in the TB study.

We also evaluated $\tilde{\pi}_i$ and $\tilde{\eta}_i$ using (i) the true G_0, (ii) the estimator $\tilde{G}_{0B1}$, and (iii) the full set of generated X_i with different realizations of the simulated *observed data* set. Table 3 in Section B of Supplementary Materials provides a summary of the observed outcomes.

As we expected, overall the outcomes based on $\tilde{\eta}_i$ were superior to those based on $\tilde{\pi}_i$ both with respect to sensitivity and specificity. On the other hand, $\tilde{\pi}_i$ can also represent a reasonable alternative since the corresponding estimated sensitivity and specificity achieved satisfactory levels. This finding agrees with our observations based on the analysis of the TB contacts study that we reported in Section 4.1. We also explored the outcomes with the cut-off level c set to 0.5 or 0.1. The resulting procedures have either lower estimated sensitivity rates or lower

TABLE 4: Estimated regression coefficients obtained in *Simulation B* from 300 repetitions with $n = 1{,}000$, 90% right censoring, and an overall missingness rate of 70%.

	Cox PH cure model							Cox PH Model		
Parameter	α	β_1	β_2	γ_0	γ_1	γ_2	γ_3	α	β_1	β_2
(true value)	2.0	−.5	1.0	−2.0	.4	−1.5	.8	2.0	−.5	1.0
	Using only data with $R = 1$ (the complete-cases data)									
bias	.510	−.368	.365	.093	.027	−.002	.097	−.596	−.866	.970
sm($\hat{\theta}$)[a]	2.510	−.868	1.365	−1.907	0.427	−1.502	0.897	1.404	−1.366	1.970
ssd($\hat{\theta}$)[b]	.824	.498	.738	.482	.488	.199	.293	.837	.378	.865
sm($\widehat{se(\hat{\theta})_F}$)[c]	.831	.456	.803	.361	.480	.182	.283	.840	.369	.858
	With the whole dataset and true G_0									
bias	−.048	.011	.041	−.031	.004	−.003	.013	.335	−.693	.364
sm($\hat{\theta}$)	1.952	−.489	1.041	−2.031	0.404	−1.503	.813	2.335	−1.193	1.364
ssd($\hat{\theta}$)	.269	.139	.272	.388	.382	.186	.286	.568	.129	.327
sm($\widehat{se(\hat{\theta})_F}$)	.262	.137	.302	.391	.385	.175	.282	.456	.120	.307
	With the whole dataset and $\tilde{G}_{0B1}$									
bias	−.049	.015	.048	−.025	.007	−.008	.008	.348	−.694	.348
sm($\tilde{\theta}$)	1.951	−.485	1.048	−2.025	.407	−1.508	.808	2.348	−1.194	1.348
ssd($\tilde{\theta}$)	.325	.147	.283	.402	.391	.195	.291	.562	.128	.330
sm($\widehat{se(\tilde{\theta})_F}$)	.302	.145	.281	.395	.385	.139	.250	.453	.121	.308
sm($\widehat{se(\tilde{\theta})_R}$)[d]	.328	.148	.291	.411	.386	.186	.261	.468	.131	.318
	With the whole dataset and $\tilde{G}_{0B2}$									
bias	−.038	.011	−.002	−.056	.016	−.010	.013	.322	−.691	.295
sm($\tilde{\theta}$)	1.962	−.489	.998	−2.056	.416	−1.510	.813	2.322	−1.191	1.295
ssd($\tilde{\theta}$)	.324	.143	.292	.382	.388	.179	.297	.563	.128	.329
sm($\widehat{se(\tilde{\theta})_F}$)	.302	.144	.284	.350	.371	.174	.281	.457	.120	.307
sm($\widehat{se(\tilde{\theta})_R}$)	.340	.146	.285	.379	.399	.176	.293	.558	.130	.308

[a] sm($\tilde{\theta}$) or sm($\hat{\theta}$): sample mean of the evaluations of $\tilde{\theta}$ or $\hat{\theta}$, the pseudo-SPMLE or the SPMLE.
[b] ssd($\tilde{\theta}$) or ssd($\hat{\theta}$): sample standard deviations of the evaluations of $\tilde{\theta}$ or $\hat{\theta}$
[c] sm($\widehat{se(\tilde{\theta})_F}$): sample mean of the standard error estimates for $\tilde{\theta}$ using the inverse of the observed information matrix.
[d] sm($\widehat{se(\tilde{\theta})_R}$): sample mean of the standard error estimates for $\tilde{\theta}$ obtained using the robust variance estimator identified in Equation (15).

estimated specificity rates. Further investigation leads to a systematic method of ascertaining an appropriate cut-off level.

5. FINAL REMARKS

Motivated by the TB study, we have considered likelihood-based semiparametric estimation under the Cox PH cure model with right-censored event times in the presence of covariate MNAR. Our method can be viewed as an adaptation of the work by Sy & Taylor (2000) using the idea from Cook, Hu & Swartz (2011) to supplement the current data with available population

information. We illustrated our proposed approach and its application in disease screening and prediction by analyzing the TB study data, and we validated our results via a simulation study.

The idea underlying our approach could readily be applied to analysis under a different cure model, such as the semiparametric transformation cure model introduced in Lu & Ying (2004). Some of the subjects in the TB study had physical contact with the same TB patients; this suggests that the subjects were likely clustered according to their source cases. It may be possible to account for the potential within-cluster correlation by introducing a frailty variable to the cure model, following Liu et al. (2019). In addition, the current variance estimation for the SPMLE and pseudo-SPMLE is rather time-consuming. The robust sandwich variance estimator is used for pseudo-SPMLE. We remark that the variance estimator may be biased as it does not fully take into account the variation that is due to using an estimated $\tilde{G}_0$. It would be worth exploring alternative variance estimators. For example, one could consider the variance estimation that involves a resampling procedure.

Several additional investigations would also be interesting. First, branching out from the procedures for identifying susceptible subjects using $\hat{\pi}$ or $\hat{\eta}$ in Section 4.1, it might be of theoretical and practical interest to develop an approach for dynamic screening and prediction. Second, as mentioned in Section 4, we could investigate how to systematically determine a cut-off level c that satisfies a prespecified criterion. Finally, we could extend the current procedures to process disease screening and prediction with a given level of confidence.

APPENDIX

This appendix outlines our derivations for the asymptotic properties of the pseudo-SPMLE $\tilde{\boldsymbol{\theta}}_n$ identified in Section 3.2. Following Lu (2008), in the context of the Cox PH cure model we adapt the arguments of Guo, Hu & Liu (2017) in their asymptotics study, which extends the arguments in Zeng, Lin & Lin (2008) to situations involving a covariate observation MNAR.

Part A. Assumptions

We assume the following regularity conditions.

C1. There exists a constant $\delta_0 > 0$ such that $P(C \geq T^* \geq \tau | X, Z) > \delta_0$ almost surely, where τ is the duration of the study and $0 < \tau < \infty$.
C2. There exists a constant M_0 such that $P(||X|| + ||Z|| \leq M_0) = 1$.
C3. The true baseline cumulative function $\Lambda_0(t)$ is strictly increasing over $[0, \tau]$ and continuously differentiable; in addition, $\Lambda_0(0) = 0$.
C4. The true value of (α, β, γ) belongs to a known compact set $\Theta = \{(\alpha, \beta, \gamma) : ||\gamma|| \leq A_0, ||\beta|| \leq B_0, ||\gamma|| \leq C_0\}$ with finite constants A_0, B_0, C_0.
C5. $\int_{-\infty}^{\infty} d\tilde{G}_0(x|Z) = 1$ for all Z, and $n^{-1/2} \sum_{i=1}^{n} \sup_x \left[\tilde{G}_0(x|Z_i) - G_0(x|Z_i)\right] \to 0$ as $n \to \infty$.

Conditions C1–C4 are conventional in multivariate survival analysis. Specifically, some subjects have experienced the event of interest for the TB study so that condition C1 is satisfied. Furthermore, the covariates considered in the TB study are finite so condition C2 is satisfied. Condition C3 concerns the true value of the baseline hazard function. Condition C4 sets up the boundary for the regression coefficients. In practice, A_0, B_0, C_0 can be chosen to be sufficiently large to cover the true values. Condition C5 is useful in proving the weak convergency of the pseudo-SPMLE. Actually, we need only that $\tilde{G}_0$ is a consistent estimation to derive consistency. The estimator $\tilde{G}_0(\cdot|Z)$ proposed in "With supplementary information on $G(x|z)$" section satisfies Condition C5.

Part B. A Proof of Proposition 1

The observed data likelihood function $L(\boldsymbol{\theta})$ identified in Equation (8) with G_0 replaced by its estimator $\tilde{G}_0$ can be written as $L(\boldsymbol{\theta}|\tilde{G}_0) = \prod_{i=1}^n \lambda_0\{U_i\}^{\Delta_i} \tilde{W}_{ni}(\boldsymbol{\theta})$. Similar to the proof of Lemma 3.1 in Guo Hu & Liu (2017), we can show that $\tilde{W}_{ni}(\boldsymbol{\theta})$ is bounded by $c_0\{1+\Lambda_0(U_i)\}^{-(\Delta_i+1)}$ with probability 1, where c_0 is a constant independent of $\boldsymbol{\theta}$. Thus $L(\boldsymbol{\theta}|\tilde{G}_0)$ is bounded from above by $\prod_{i=1}^n \lambda_0\{U_i\}^{\Delta_i} c_0\{1+\Lambda_0(U_i)\}^{-(\Delta_i+1)}$. We conclude that the jump size of Λ_0 must be finite. In addition, (α, β, γ) belongs to a compact set by Condition C4, and $L(\boldsymbol{\theta}|\tilde{G}_0)$ is continuous with respect to $\boldsymbol{\theta}$. Therefore, the pseudo-SPMLE $\tilde{\boldsymbol{\theta}}_n = \text{argmax} L(\boldsymbol{\theta}|\tilde{G}_0)$ exists. The consistency of $\tilde{\boldsymbol{\theta}}_n$ is proved by the following steps.

Step 1. We prove that the pseudo-MLE

$$\tilde{\Lambda}_{0n}(t) = \sum_{i=1}^n \int_0^t \frac{Y_i(s)dN_i(s)}{nD_n^{(0)}(s; \tilde{\alpha}_n, \tilde{\beta}_n|\tilde{\boldsymbol{\theta}}_n, \tilde{G}_0)}, \quad 0 < t \le \tau,$$

which is obtained from Equation (14) with G_0 replaced by $\tilde{G}_0$, has an upper bound in $[0, \tau]$ with probability one. Based on the construction of $\tilde{\Lambda}_{0n}(t)$ we define two other functions, for $0 < t \le \tau$,

$$\overline{\Lambda}_{0n}(t) = \sum_{i=1}^n \int_0^t \frac{Y_i(s)dN_i(s)}{nD_n^{(0)}(s; \alpha, \beta|\boldsymbol{\theta}, G_0)}, \quad \overline{\overline{\Lambda}}_{0n}(t) = \sum_{i=1}^n \int_0^t \frac{Y_i(s)dN_i(s)}{nD_n^{(0)}(s; \alpha, \beta|\boldsymbol{\theta}, \tilde{G}_0)}.$$

We verify below that $\overline{\Lambda}_{0n}(t)$ and $\overline{\overline{\Lambda}}_{0n}(t)$ both converge to $\Lambda_0(t)$ uniformly in $t \in [0, \tau]$ with probability one.

By the Glivenko–Cantelli property of the class $\mathcal{F} = \Big\{ Y_i(t)\hat{\eta}_i(\boldsymbol{\theta}; G_0)\big[R_i e^{\alpha X_i} + (1-R_i)\widehat{e^{\alpha X_i}}(\boldsymbol{\theta})\big] e^{\beta Z_i} : (\alpha, \beta, \gamma) \in \Theta, t \in [0, \tau], \Lambda_0(0) = 0, \Lambda_0(\tau) \le C \Big\}$, we obtain

$$\sup_{t\in[0,\tau]} \left| D_n^{(0)}(t; \alpha, \beta|\boldsymbol{\theta}, G_0) - \mu(t) \right| \to 0, \tag{A.1}$$

almost surely, where $\mu(t) = E\left\{ S_C(t|X_i, Z_i) f(t|X_i, Z_i)/\Lambda_0'(t) \right\}$ with $S_C(t|X_i, Z_i)$ the survivor function of C_i given X_i and Z_i. Note that $\mu(t)$ is bounded away from zero uniformly, and thus $\overline{\Lambda}_n(t) \overset{a.un.}{\to} E\left[I(U_i \le t)\Delta_i/\mu(U_i) \right]$, which is $\Lambda_0(t)$. As a result, we can obtain that $\sup_{t\in[0,\tau]} \left| D_n^{(0)}(t; \alpha, \beta|\boldsymbol{\theta}, \tilde{G}_0) - \mu(t) \right| \to 0$ almost surely by Conditions C2, C4, C5. This proves $\overline{\overline{\Lambda}}_{0n}(t) \overset{a.un.}{\to} \Lambda_0(t)$.

Clearly, $n^{-1}p\ell(\tilde{\boldsymbol{\theta}}_n) - n^{-1}p\ell(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}) \ge 0$ with $p\ell(\boldsymbol{\theta})$ equal to the log-transformed likelihood function $L(\boldsymbol{\theta}|\tilde{G}_0)$ identified in Equation (8) with $G_0(\cdot)$ replaced by its estimator $\tilde{G}_0(\cdot)$. By Equation (A.1) and since $0 \le \pi(X_i, Z_i; \gamma) \le 1$, $0 \le 1 - \pi(X_i, Z_i; \gamma) + \pi(X_i, Z_i; \gamma) e^{-\overline{\overline{\Lambda}}_{0n}(U_i) e^{\alpha X_i + \beta Z_i}} \le 1$, we have

$$n^{-1}p\ell\left(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}\right) = O(1) + \frac{1}{n}\sum_{i=1}^n \Delta_i \log(n^{-1}).$$

On the other hand, $n^{-1}p\ell(\tilde{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^n \Delta_i \log \tilde{\lambda}_0\{U_i\} + \frac{1}{n}\sum_{i=1}^n \log \tilde{W}_{ni}(\tilde{\boldsymbol{\theta}}_n)$ is bounded from above by $O(1) + \frac{1}{n}\sum_{i=1}^n \Delta_i \log \tilde{\lambda}_0\{U_i\} - \frac{1}{n}\sum_{i=1}^n (\Delta_i + 1)\log\{1 + \tilde{\Lambda}_{0n}(U_i)\}$. It follows that

$$0 \le O(1) + \frac{1}{n}\sum_{i=1}^n \Delta_i \log\{n\tilde{\lambda}_0\{U_i\}\} - \frac{1}{n}\sum_{i=1}^n (\Delta_i + 1)\log\{1 + \tilde{\Lambda}_{0n}(U_i)\}. \tag{A.2}$$

We can show that if $\tilde{\Lambda}_{0n}(\tau) \to \infty$, the right-hand side of Equation (A.2) is eventually negative.

Thus we have shown that $\tilde{\Lambda}_{0n}(\tau)$ has an upper bound with probability one. By the Helly selection theorem, we can assume that $\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$.

Step 2. We now show that $\boldsymbol{\theta}^* = \boldsymbol{\theta}$. Denote $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}|G_0)$. Note that $0 \le n^{-1}p\ell(\tilde{\boldsymbol{\theta}}_n) - n^{-1}p\ell(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n})$ is equal to

$$n^{-1}\left\{ \left[p\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\tilde{\boldsymbol{\theta}}_n)\right] + \left[\ell\left(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}\right) - p\ell\left(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}\right)\right] + \left[\ell(\tilde{\boldsymbol{\theta}}_n) - \ell\left(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}\right)\right]\right\}. \tag{A.3}$$

By Condition C5, the first and second terms in Equation (A.3) diverge to 0 as $n \to \infty$. The third term can be written as

$$\frac{1}{n}\sum_{i=1}^{n} \Delta_i \log\left[\frac{\tilde{\lambda}_0}{\overline{\overline{\lambda}}_0}\right] + \frac{1}{n}\sum_{i=1}^{n} \log \tilde{W}_{ni}(\tilde{\boldsymbol{\theta}}_n) - \frac{1}{n}\sum_{i=1}^{n} \log \tilde{W}_{ni}\left(\alpha, \beta, \gamma, \overline{\overline{\Lambda}}_{0n}\right).$$

We can prove that the third term diverges to the negative Kullback–Leibler information as $n \to \infty$. The identifiability thus implies that $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

Part C. A Proof of Proposition 2

The proof includes the following two steps.

Step 1. By the arguments employed in Guo, Hu & Liu (2017), we can establish the asymptotic distribution of the SPMLE $\widehat{\boldsymbol{\theta}}_n$ that we introduced in Section 3.1.

Step 2. First, note that $\tilde{\Psi}_n(\tilde{\boldsymbol{\theta}}_n) = 0$ and $\Psi(\boldsymbol{\theta}) = 0$. Define $\mathcal{U} = \{(h_1, h_2, h_3, h_4) : (h_1, h_2, h_3) \in \mathcal{R}^d$ and $||h_1|| \le 1, ||h_2|| \le 1, ||h_3|| \le 1; h_4(t)$ with bounded variation on $t \in [0, \tau]$ and $||h_4(\cdot)||_V \le 1\}$, where $||h_4(\cdot)||_V = \sup_{0=t_1<t_2<\cdots<t_m=\tau} \sum_k |h_4(t_k) - h_4(t_{k-1})|$. We construct the following four random maps $\Psi_n, \tilde{\Psi}_n, \Psi$, and $\tilde{\Psi}$: for $\underline{h} = (h_1, h_2, h_3, h_4) \in \mathcal{U}$,

$$\Psi_n(\boldsymbol{\theta})[h_1, h_2, h_3, h_4] \equiv \mathcal{P}_n\left\{ h_1^T l_\alpha(\boldsymbol{\theta}) + h_2^T l_\beta(\boldsymbol{\theta}) + h_3^T l_\gamma(\boldsymbol{\theta}) + l_\Lambda(\boldsymbol{\theta})\left[\int h_4 d\Lambda\right]\right\},$$

$$\tilde{\Psi}_n(\boldsymbol{\theta})[h_1, h_2, h_3, h_4] \equiv \mathcal{P}_n\left\{ h_1^T pl_\alpha(\boldsymbol{\theta}) + h_2^T pl_\beta(\boldsymbol{\theta}) + h_3^T pl_\gamma(\boldsymbol{\theta}) + pl_\Lambda(\boldsymbol{\theta})\left[\int h_4 d\Lambda\right]\right\},$$

$$\Psi(\boldsymbol{\theta})[h_1, h_2, h_3, h_4] \equiv \mathcal{P}\left\{ h_1^T l_\alpha(\boldsymbol{\theta}) + h_2^T l_\beta(\boldsymbol{\theta}) + h_3^T l_\gamma(\boldsymbol{\theta}) + l_\Lambda(\boldsymbol{\theta})\left[\int h_4 d\Lambda\right]\right\}, \text{ and}$$

$$\tilde{\Psi}(\boldsymbol{\theta})[h_1, h_2, h_3, h_4] \equiv \mathcal{P}\left\{ h_1^T pl_\alpha(\boldsymbol{\theta}) + h_2^T pl_\beta(\boldsymbol{\theta}) + h_3^T pl_\gamma(\boldsymbol{\theta}) + pl_\Lambda(\boldsymbol{\theta})\left[\int h_4 d\Lambda\right]\right\}.$$

Here $l(\boldsymbol{\theta})$ is the loglikelihood function for the ith subject, $l_\alpha(\boldsymbol{\theta})$, $l_\beta(\boldsymbol{\theta})$ and $l_\gamma(\boldsymbol{\theta})$ are the scores of α, β and γ, respectively; $l_\Lambda(\boldsymbol{\theta})[\int h_4 d\Lambda]$ is the score for Λ_0 along the submodel $\Lambda + \epsilon \int h_4 d\Lambda$. In addition, $pl_\alpha(\boldsymbol{\theta})$, $pl_\beta(\boldsymbol{\theta})$ and $pl_\gamma(\boldsymbol{\theta})$ are the pseudo-scores of α, β and γ, respectively; $pl_\Lambda(\boldsymbol{\theta})[\int h_4 d\Lambda]$ is the pseudo-score for Λ_0 along the submodel $\Lambda + \epsilon \int h_4 d\Lambda$. We use $\mathcal{P}_n$ and $\mathcal{P}$ to denote the empirical measure and the expectation with the population distribution of n i.i.d observations, respectively.

We can verify the four properties listed in Theorem 3.3.1 of van der Vaart & Wellner (1996) and then achieve the asymptotic distribution of the pseudo-SPMLE $\widetilde{\boldsymbol{\theta}}_n$. The four required properties are listed below in our notation.

P1. $\sqrt{n}(\tilde{\Psi}_n - \Psi)(\tilde{\boldsymbol{\theta}}_n) - \sqrt{n}(\tilde{\Psi}_n - \Psi)(\boldsymbol{\theta}) = o_p(1 + \sqrt{n}||\tilde{\alpha}_n - \alpha|| + \sqrt{n}||\tilde{\beta}_n - \beta|| + \sqrt{n}||\tilde{\gamma}_n - \gamma|| + \sqrt{n}||\tilde{\Lambda}_{0n} - \Lambda_0||_{l^\infty[0,\tau]})$, where $||\cdot||_{l^\infty[0,\tau]}$ is the supremum norm in $[0, \tau]$.

P2. $\Psi(\cdot)$ is Fréchet-differentiable at the true value of $\boldsymbol{\theta}$, namely, $||\Psi(\boldsymbol{\theta}^\star) - \Psi(\boldsymbol{\theta}) - \dot{\Psi}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\star - \boldsymbol{\theta})|| = o(||\alpha^\star - \alpha|| + ||\beta^\star - \beta|| + ||\gamma^\star - \gamma|| + ||\Lambda_0^\star - \Lambda_0||_{l^\infty[0,\tau]})$ as $\alpha^\star \to \alpha, \beta^\star \to \beta, \gamma^\star \to \gamma, \Lambda_0^\star \to \Lambda_0$.

P3. $\sqrt{n}(\tilde{\Psi}_n - \Psi)(\boldsymbol{\theta})$ converges in distribution to a tight random element V.

P4. The derivative of $\Psi(\cdot)$ at $\boldsymbol{\theta}$, denoted by $\dot{\Psi}(\boldsymbol{\theta})$, is continuously invertible.

Following Theorem 3.3.1 of van der Vaart & Wellner (1996), $\sqrt{n}(\tilde{\alpha}_n - \alpha, \tilde{\beta}_n - \beta, \tilde{\gamma}_n - \gamma, \tilde{\Lambda}_{0n} - \Lambda_0)$ converges weakly to $-\dot{\Psi}^{-1}V$ with $\dot{\Psi}(\alpha^\star - \alpha, \beta^\star - \beta, \gamma^\star - \gamma, \Lambda_0^\star - \Lambda_0)[h_1, h_2, h_3, \int h_4 d\Lambda_0]$ equal to

$$(\alpha^\star - \alpha)\mathcal{W}_1(\underline{h}) + (\beta^\star - \beta)\mathcal{W}_2(\underline{h}) + (\gamma^\star - \gamma)\mathcal{W}_3(\underline{h}) + \int_0^\tau \mathcal{W}_4(\underline{h}) d(\Lambda_0^\star - \Lambda_0),$$

where $\underline{h} = (h_1, h_2, h_3, h_4(\cdot))$ and

$$\mathcal{W}_1(\underline{h}) = E[l_{\alpha\alpha}(\boldsymbol{\theta})]h_1 + E[l_{\beta\alpha}(\boldsymbol{\theta})]h_2 + E[l_{\gamma\alpha}(\boldsymbol{\theta})]h_3 - E\Big\{ \int_0^{U_i} h_4(t) d\Lambda_0(t) \frac{\partial b_i(\boldsymbol{\theta})}{\partial \alpha} \Big\},$$

$$\mathcal{W}_2(\underline{h}) = E[l_{\alpha\beta}(\boldsymbol{\theta})]h_1 + E[l_{\beta\beta}(\boldsymbol{\theta})]h_2 + E[l_{\gamma\beta}(\boldsymbol{\theta})]h_3 - E\Big\{ \int_0^{U_i} h_4(t) d\Lambda_0(t) \frac{\partial b_i(\boldsymbol{\theta})}{\partial \beta} \Big\},$$

$$\mathcal{W}_3(\underline{h}) = E[l_{\alpha\gamma}(\boldsymbol{\theta})]h_1 + E[l_{\beta\gamma}(\boldsymbol{\theta})]h_2 + E[l_{\gamma\gamma}(\boldsymbol{\theta})]h_3 - E\Big\{ \int_0^{U_i} h_4(t) d\Lambda_0(t) \frac{\partial b_i(\boldsymbol{\theta})}{\partial \gamma} \Big\},$$

$$\mathcal{W}_4(\underline{h}) = E[l_{\alpha\underline{\lambda}}(\boldsymbol{\theta})I(U_i \geq t)]h_1 + E[l_{\beta\underline{\lambda}}(\boldsymbol{\theta})I(U_i \geq t)]h_2 + E[l_{\gamma\underline{\lambda}}(\boldsymbol{\theta})I(U_i \geq t)]h_3$$
$$- E\big[I(U_i \geq t) b_i(\boldsymbol{\theta}) h_4(t)\big] - E\Big[\int_0^{U_i} h_4(t) d\Lambda_0(t) \frac{\partial b_i(\boldsymbol{\theta})}{\partial \underline{\lambda}} \Big].$$

Here $l_{rs} = \partial l(\boldsymbol{\theta}) \big/ \partial(r, s)$ and $b_i(\boldsymbol{\theta}) = \tilde{\eta}_i(\boldsymbol{\theta}) \big[R_i e^{\alpha X_i} + (1 - R_i) \widetilde{e^{\alpha X_i}}(\boldsymbol{\theta}) \big] e^{\beta Z_i}$.

Moreover, the derivation above shows that for $\underline{h} = (h_1, h_2, h_3, h_4) \in \mathcal{U}$,

$$\sqrt{n}\Big\{ (\tilde{\alpha}_n - \alpha)h_1 + (\tilde{\beta}_n - \beta)h_2 + (\tilde{\gamma}_n - \gamma)h_3 + \int_0^\tau h_4 d(\tilde{\Lambda}_{0n} - \Lambda_0) \Big\}$$
$$= -\sqrt{n}(\mathcal{P}_n - \mathcal{P})\Big\{ l_\alpha(\boldsymbol{\theta})\tilde{h}_1 + l_\beta(\boldsymbol{\theta})\tilde{h}_2 + l_\gamma(\boldsymbol{\theta})\tilde{h}_3 + l_\Lambda(\boldsymbol{\theta})\Big[\int \tilde{h}_4 d\Lambda_0 \Big] \Big\} + o_p(1), \qquad \text{(A.4)}$$

where $(\tilde{h}_1, \tilde{h}_2, \tilde{h}_3, \tilde{h}_4) = (\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3, \mathcal{W}_4)^{-1}\underline{h}$, and $o_p(1)$ denotes a random variable $r_n(\underline{h})$ such that $\sup_{\underline{h} \in \mathcal{U}} |r_n(\underline{h})| \to 0$ in probability. Thus, by choosing $h_4(\cdot) \equiv 0$ in Equation (A.4), we conclude that $\tilde{\alpha}_n, \tilde{\beta}_n$ and $\tilde{\gamma}_n$ are asymptotically linear estimators for the parameter α, β and γ, respectively, and the corresponding influence functions are on the space spanned by the score functions. Therefore, the semiparametric efficiency theory (cf. Gill & van der Vaart, 1993) yields the required result that $\tilde{\alpha}_n, \tilde{\beta}_n$ and $\tilde{\gamma}_n$ are asymptotically efficient estimators.

SUPPLEMENTARY MATERIAL

Additional numerical results referenced in Sections 3 and 4 are reported in the online Supplementary Materials. Our source code for the data analysis and simulation is also included. We would be happy to help with requests to reproduce the numerical results that we reported in this article.

ACKNOWLEDGEMENTS

The research was partially supported by the grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Statistical Sciences Institute (CANSSI), and the Fundamental Research Funds for the Central Universities, South-Central University for Nationalities (CZQ20008). We are grateful to Professor Yanyan Liu for helpful discussions. We thank the editor, a copyeditor, an associate editor and two anonymous reviewers for their constructive comments and suggestions, which led to a much improved version of the article.

BIBLIOGRAPHY

Beesley, L. J., Bartlett, J. W., Wolf, G. T., & Taylor, J. M. (2016). Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine*, 35, 4701–4717.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Chen, M. H. & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, 57, 43–52.

Cook, V. J., Hernandez-Garduno, E., Hu, X. J., Elwood, R. K., & FitzGerald, J. M. (2005). The influence of cluster-status of source cases on contact evaluation and the development of secondary active tuberculosis. *Proceedings of the American Thoracic Society (PATS) 2005; 2 Abstract Issue*.

Cook, V. J., Hu, X. J., & Swartz, T. B. (2011). Cox regression with covariates missing not at random. *Statistics in Biosciences*, 3, 208–222.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Fang, H. B., Li, G., & Sun, J. G. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, 32, 59–75.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14, 257–262.

Gill, R. D. & van der Vaart, A. W. (1993). Non- and semi-parametric maximum likelihood estimators and the von Mises method: II *Scandinavian Journal of Statistics*, 20, 271–288.

Guo, L., Hu, X. J., & Liu, Y. (2017). Estimation under Cox proportional hazards model with covariates missing not at random *Communications in Statistics – Theory and Methods*, 46, 8952–8972.

Herring, A. H. & Ibrahim, J. G. (2002). Maximum likelihood estimation in random effects cure rate models with nonignorable missing covariates *Biostatistics*, 3, 387–405.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters *Journal of the American Statistical Association*, 27, 887–906.

Kuk, A. Y. C. & Chen, C. H. (1992). A mixture model combining logistic regression with proportional hazards regression *Biometrika*, 79, 531–541.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, New York.

Liu, L., Liu, Y., Xiong, Y., & Hu, X. J. (2019). Cox regression of clustered event times with covariates missing not at random *Scandinavian Journal of Statistics*, 46, 1315–1346.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model *Annals of the Institute of Statistical Mathematics*, 60, 545–574.

Lu, W. & Ying, Z. (2004). On semiparametric transformation cure models *Biometrika*, 91, 331–343.

Maller, R. A. & Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Peng, Y. W. & Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation *Biometrics*, 56, 237–243.

Public Health Agency of Canada. (2007). *HIV/AIDS Epi Updates*.

Sy, J. P. & Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model *Biometrics*, 56, 227–236.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models *Biometrics*, 51, 899–907.
van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, Berlin.
White, H. (1982). Maximum likelihood estimation of misspecified models *Econometrica*, 50, 1–25.
World Health Organization. (2018). *Tuberculosis*. [Online]. https://www.who.int/en/news-room/fact-sheets/detail/tuberculosis.
Wu, C. F. J. (1983). On the convergence properties of the EM algorithm *The Annals of Statistics*, 11, 95–103.
Zeng, D. L., Lin, D. Y., & Lin, X. H. (2008). Semiparametric transformation models with random effects for clustered failure time data *Statistica Sinica*, 18, 355–377.

Received 01 November 2018
Revised 31 December 2019
Accepted 12 January 2020