Published in final edited form as:
Commun Stat Simul Comput. 2018 ; 47(6): 1677–1695. doi:10.1080/03610918.2017.1322699.

Analysis of Dependently Truncated Data in Cox Framework

Yang Liu,
Division of Analysis, Research, and Practice Integration, National Center for Injury Prevention and Control, U.S. Centers for Disease Control and Prevention, Atlanta, GA 30341, USA

Ji Li, and
Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, USA

Xu Zhang
Division of Clinical and Translational Sciences, Department of Internal Medicine, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Abstract

Truncation is a known feature of bone marrow transplant (BMT) registry data, for which the survival time of a leukemia patient is left truncated by the waiting time to transplant. It was recently noted that a longer waiting time was linked to poorer survival. A straightforward solution is a Cox model on the survival time with the waiting time as both truncation variable and covariate. The Cox model should also include other recognized risk factors as covariates. In this paper we focus on estimating the distribution function of waiting time and the probability of selection under the aforementioned Cox model.

Keywords

Dependent truncation; Cox model; inverse probability weighting

1 Introduction

The Center for International Blood and Marrow Transplant Research (CIBMTR) is a network of clinicians and basic science researchers who confidentially share data on blood and bone marrow transplant (BMT) patients. The CIBMTR Data Collection Center, located at the Medical College of Wisconsin, is a registry of patient data contributed from more than 450 transplant centers worldwide. The registry does not collect data of patients who died waiting for matched donors. Therefore, a patient cohort from the registry is a truncated sample in which the time-to-failure is left truncated by the waiting time to transplant.

It is of great importance to assess the effects of prognostic factors on survival or leukemia-free survival. With the registry data, it is necessary to deal with the truncation issue because diagnosis of leukemia has to be the time original for survival so that effec-tiveness of

Disclaimer: The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

transplantation can be compared to that of an alternative treatment such as chemotherapy. Therefore, the left-truncated version of the Cox model is used to associate the covariates to the survival outcome. In studies using the registry data, researchers assumed that the survival time and waiting time to transplant were independent (Barrett et al., 1994). In recent years, there was clinical evidence that a longer waiting time for transplant was a poor prognosis. Balduzzi group was among the first to study the effect of waiting time to transplant on survival. They brought up the point that a higher level of toxicity from chemotherapy could be cumulated during an extended waiting period, and the patient could be subsequently associated with a poor prognosis. In addition, the effect of waiting time causes the problem of dependent truncation in analysis of the registry data. To solve the problem, the registry data should be analyzed by a left-truncated version of the Cox model with the waiting time as a covariate.

The BMT registry data can be explored for disparity studies. Two disparity-related questions can be investigated. The first question relates to the probability of being in the truncated sample. We may estimate this probability in racial or social-economic subgroups and examine if any characteristic is associated with obviously lower chance of receiving transplants. Second, we may estimate the distribution function of waiting time to transplant in racial or social-economic subgroups, so that we can learn if patients in any subgroup have to wait substantially longer time to receive transplants. As an illustrative example we consider a CIBMTR cohort consisting of 376 children receiving transplants in their second complete remission in 1990–1999 (Barrett et al., 1994). We use the Cox model with waiting time to transplant as both truncation variable and covariate to deal with the dependence between survival time and waiting time to transplant. Based on this model, we estimate the probability of selection and the distribution function of waiting time to transplant in some subsets. We wish to investigate whether the waiting time to transplant varies between the subsets.

The general concept of truncation in statistical sampling means unobservability of a continuous variable when the value is above or below certain threshold. In the field of lifetime data analysis, truncation refers to the scenario that a pair of continuous variables T and L are only observable if $L < T$. For the CIBMTR registry data, T is the survival time from diagnosis for a patient reported to the registry, and L is patient’s waiting time to transplant. Two types of truncation coexist in that T is left truncated by L and L is right truncated by T. The majority of statistical inferences for truncated data were developed assuming independence between L and T in the observable quadrant $L < T$, which is denoted as quasi-independence (Tsai, 1990). The survival function of T can be estimated by the left-truncated version of the Kaplan-Meier estimator (1958), though truncation issue was discussed as late entrance in their paper. The asymptotic properties of this estimator was studied by Woodroofe (1985), Wang, Jewell and Tsai (1986), Keiding and Gill (1990) among others. Right truncation has routinely been dealt with by transforming L to $\tau - L$ where τ is a very large constant. The transformed variable $\tau - L$ is left truncated by $\tau - T$, and then the methodologies developed for left truncation would become applicable. According to this principle, the distribution function of L is estimated by the right-truncated version of the Kaplan-Meier estimator (Keiding and Gill, 1990). Inverse probability weighting was found out to be an alternative solution to truncation. The nonparametric inverse-probability-weighted (IPW) estimators were shown to be identical to the truncated-version Kaplan-Meier estimators (Shen, 2003). The semi-parametric IPW estimator was proposed by Wang (1989) when the distribution of the truncation variable can be parametrisized. Such an estimator is more efficient than its nonparametric counterpart. Some statisticians noted the usefulness of one truncation parameter, the probability of selection $P(L < T)$. The inferences about this parameter under random truncation have been studied by Woodroofe (1985) and Keiding and Gill (1990).

When the truncated data contain covariates associated with T, the left-truncated version of the Cox model is the commonly used analytical method. One important application of this model is to analyze BMT registry data, to evaluate prognostic factors and compare transplant versus other treatment option (Klein and Zhang, 1996). When the association between covariates and L is of study interest, the standard analytical methods include the right-truncated version of the Cox model targeting on the retro-hazard function (Kalbfleisch and Lawless, 1991; Gross and Huber-Carol, 1992) and the full-likelihood-based Cox model on the hazard function (Finkelstein et al., 1993).

Several tests have been proposed to test the quasi-independence with truncated data, including the Kendall's tau and the weighted rank statistics by Tsai (1990) and Efron and Petrosian (1992), respectively. Another test was suggested by Jones and Crowley (1992) by taking L as a covariate of T in the left-truncated version of the Cox model. Let $\lambda_T(t; z)$, $\lambda_0(t)$ and α be respectively the conditional hazard function of T, the unspecified nonnegative function and the regression coefficient. Jones and Crowley proposed to use the score test based on the model, $\lambda_T(t; L) = \lambda_0(t)e^{\alpha L}$, to test the quasi-independence.

The Cox model with L as both truncation variable and covariate was later found out to be a convenient method to address the dependence in truncated data. Mackenzie (2012) used the Cox model with age as both truncation variable and covariate to analyze the survival data of users of VA health system. In that paper, age was the only covariate so the assumed model was exactly $\lambda_T(t; L) = \lambda_0(t)e^{\alpha L}$. Mackenzie proposed the estimators for the distribution functions of truncation and lifetime variables using the inverse-probability-weighting technique. Zhang et al. (2015) illustrated an application of this model in analyzing BMT registry data. The waiting time to transplant is the truncation variable and also associated with the survival. Under the Cox model that the waiting time to transplant is the only covariate, Zhang et al. studied the inference for the probability of selection. It is natural to believe that for leukemia patients receiving bone marrow transplants the demographic factors such as age, race and clinical factors such as T-cell phenotype are associated with the survival. In this study we analyze the BMT registry data using a left-truncated version Cox model with covariate L and other prognostic factors. The methodological development targets at estimating the probability of selection and the distribution function of truncation variable under such a Cox model.

The remainder of this paper is organized as follows. Section 2 presents the point and variance estimators for the probability of selection and the distribution function of truncation variable with truncated data. The results of a simulation study are provided in Section 3. In Section 4, the BMT registry data set of 376 children is analyzed to illustrate the proposed methods. The final discussion is given in Section 5. The appendix of this paper sketches the derivation of the asymptotic distribution of the proposed estimator.

2 The methods

2.1 To estimate the probability of selection

The sample is summarized as $\{L_i^*, T_i^*, \boldsymbol{Z}_i\}$, $i = 1, \cdots, n$, $L_i^* < T_i^*$. Distributions of L^* and T^* are the conditional distributions of L and T given $L < T$. $\boldsymbol{Z}_i$ is the vector of p covariates. Let (a_K, b_K) is the interior of the support of a distribution function K, $a_K = \inf\{x: K(x) > 0\}$ and $b_K = \sup\{x: K(x) < 1\}$. Let G be the distribution function of L and $F_{\boldsymbol{z}}$ be the conditional distribution of T given $\boldsymbol{z}$. Similar to the conditions used in the nonparametric setting (Woodroofe, 1985), we assume that $a_G < a_{F_{\boldsymbol{z}}}$ and $b_G < b_{F_{\boldsymbol{z}}}$. In addition, only conditional distributions of $F_{\boldsymbol{z}}$ and G given respectively $T|\boldsymbol{z} \geq a_G$ and $L \leq b_{F_{\boldsymbol{z}}}$ are estimable.

The distribution function of T is determined through the underlying model $\lambda_T(t; L, \boldsymbol{z}) = \lambda_0(t)\exp\{g(\boldsymbol{\alpha}, L) + \boldsymbol{\gamma}^T\boldsymbol{z}\}$ for $t \geq L$. In the regressor, g is a known function, $\boldsymbol{\gamma}$ is the vector of p regression coefficients, and $\boldsymbol{\alpha}$ is the vector of regression coefficients associated with polynomial or other terms of L. For example, if $g()$ is a linear combination of $\{L, L^2, \cdots, L^k\}$, the dimension of $\boldsymbol{\alpha}$ is k. In order to have simple presentation, we let $g(\boldsymbol{\alpha}, L) = \alpha L$, which pertains to a constant hazard ratio between any two levels in L. This simplification does not reduce the generality of the method. The complete specification of the working model is given by

$$\lambda_T(t; L, z) = \begin{cases} \lambda_0(t)\ \exp\{\boldsymbol{\gamma}^T z\} & t < L \\ \lambda_0(t)\ \exp\{\alpha L + \boldsymbol{\gamma}^T z\} & t \geq L \end{cases}. \quad (1)$$

It is known that in the setting without covariates independence of L and T cannot be tested in the region $T < L$ (Tsai, 1990). In this region, because no data are observed, the relation of L and T cannot be recognized. In the above model, the specification for $t < L$ is untestable because data are not observed in this region. If effect of a covariate is believed to change over time, the model and methods discussed in the paper are not applicable to handle such a covariate.

We define the notations $\tilde{\boldsymbol{Z}}_i^T = \{L_i^* \boldsymbol{Z}_i^T\}$, $\boldsymbol{\beta}^T = \{\alpha\ \boldsymbol{\gamma}^T\}$, as well as the counting processes, $N_{T,i}(x) = I(T_i^* \leq x)$ and $Y_i(x) = I(L_i^* \leq x \leq T_i^*)$. Define

$$\boldsymbol{S}^{(p)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t)\tilde{\boldsymbol{Z}}_i^{\otimes p} \exp(\boldsymbol{\beta}^T\tilde{\boldsymbol{Z}}_i), \quad p = 0, 1, 2,$$

where $a^{\otimes 2} = aa^T$. The partial likelihoods (Cox, 1972; Andersen et al., 1993) can be constructed for Model (1), yielding the following score estimation equation,

$$\mathcal{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^\infty\left[\widetilde{\boldsymbol{Z}}_i - \frac{\sum_{j=1}^{n} Y_j(t)\widetilde{\boldsymbol{Z}}_j \exp\left(\boldsymbol{\beta}^T\widetilde{\boldsymbol{Z}}_j\right)}{\sum_{j=1}^{n} Y_j(t)\exp\left(\boldsymbol{\beta}^T\widetilde{\boldsymbol{Z}}_j\right)}\right] dN_{T,i}(t).$$

Let $\hat{\boldsymbol{\beta}}^T = \{\hat{\alpha}\ \hat{\boldsymbol{\gamma}}^T\}$ be the maximum partial likelihood estimate (MLE). It is the solution to $\mathcal{U}(\boldsymbol{\beta}) = \mathbf{0}$. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated by the inverse of

$$\widehat{\mathcal{I}}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\int_0^\infty\left[\frac{\boldsymbol{S}^{(2)}(\hat{\boldsymbol{\beta}},t)}{S^{(0)}(\hat{\boldsymbol{\beta}},t)} - \left(\frac{\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\beta}},t)}{S^{(0)}(\hat{\boldsymbol{\beta}},t)}\right)^{\otimes 2}\right] dN_{T,i}(t).$$

The Breslow estimator of $\Lambda_0(x) = \int_0^x \lambda_0(u)du$ is given by

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_{T,i}(u)}{\sum_{j=1}^{n} Y_j(u)\exp\left(\hat{\boldsymbol{\beta}}^T\widetilde{\boldsymbol{Z}}_j\right)}. \quad (2)$$

Similar to Zhang et al. (2015), we introduce a latent variable $T_0^{\boldsymbol{z}}$ which is the failure time variable given $\boldsymbol{z}$ and is associated with the hazard function $\lambda_0(t)e^{\gamma^T\boldsymbol{z}}$. We assume that the truncation variable L is independent of $T_0^{\boldsymbol{z}}$. Being selected in the sample will alter the hazard function of $T_0^{\boldsymbol{z}}$ by Model (1). Based on this model, given $L = 0$, the covariate-specific survival probability at t is denoted by $S_0(t;\boldsymbol{z})$, where $S_0(t;\boldsymbol{z}) = P(T > t | L = 0, \boldsymbol{z}) = \exp\{-\Lambda_0(t)e^{\boldsymbol{\gamma}^T\boldsymbol{z}}\}$. We also let $\boldsymbol{Z}^t$ to be the covariate vector in the sample associated with an observed truncation time t. Please note that $\boldsymbol{Z}^t$ is not time dependent and we consider fixed covariates $\boldsymbol{Z}$ only in this study. In addition, we assume that there are no ties among the times. The estimators provided in this paper can be easily extended to the data with ties.

Let $G^*(t)$ be the distribution of L given that L is truncated, $G^*(t) = \int P(L \le t \mid L \le T_{\boldsymbol{z}}^0)dV(z)$, where $V(z)$ is the distribution of $\boldsymbol{z}$ in the truncated sample. Let $P(\boldsymbol{\beta})$ be the probability that the truncated sample is selected from the underlying population,

$$P(\boldsymbol{\beta}) = \left[\int_0^\infty \frac{1}{S_0(t;\boldsymbol{Z}^t)}dG^*(t)\right]^{-1}.$$

$G(t)$ has the expression

$$G(t) = P(\boldsymbol{\beta})\int_0^t \frac{1}{S_0(u;\boldsymbol{Z}^t)}dG^*(u).$$

Note that both $P(\boldsymbol{\beta})$ and $G(t)$ are expressed in the inverse-probability-weighting form. Inverse probability weighting is a commonly used technique in analysis of truncated samples.

$S_0(t; z)$ can by estimated by $\hat{S}_0(t; z) = \exp\{-\hat{\Lambda}_0(t)e^{\hat{\gamma}^T z}\}$. G^* can be naturally estimated by the empirical estimator of the truncated sample, $\hat{G}^*(t) = n^{-1}\sum_{i=1}^{n} I(L_i^* \le t)$. In the following context $\boldsymbol{Z}^t = \boldsymbol{Z}_i I(L_i^* = t; i = 1, \cdots, n)$. We can estimate $P(\boldsymbol{\beta})$ by

$$\hat{P}(\hat{\boldsymbol{\beta}}) = \left[\int_0^\infty \frac{1}{\hat{S}_0(t; \boldsymbol{Z}^t)} d\hat{G}^*(t)\right]^{-1} = \left[n^{-1}\sum_{i=1}^{n} \frac{1}{\hat{S}_0(L_i^*; \mathbf{Z}_i)}\right]^{-1}.$$

To estimate the distribution function of truncation variable L, the IPW estimator can be employed.

$$\hat{G}(t) = \hat{P}(\hat{\boldsymbol{\beta}}) \int_0^t \frac{1}{\hat{S}_0(u; \boldsymbol{Z}^t)} d\hat{G}^*(u) = \left[n^{-1}\sum_{i=1}^{n} \frac{1}{\hat{S}_0(L_i^*; \mathbf{Z}_i)}\right]^{-1} \times n^{-1}\sum_{i=1}^{n} \frac{I(L_i^* \le t)}{\hat{S}_0(L_i^*; \mathbf{Z}_i)}.$$

We assume the standard regularity conditions for Cox model and selection bias model (Andersen and Gill, 1982; Vardi, 1985):

1. There exists $s^{(0)}$, $\boldsymbol{s}^{(1)}$, $\boldsymbol{s}^{(2)}$ such that

   $$\sup_{\boldsymbol{\beta}, t \in \,|\, 0,\ \infty\,)} \|\boldsymbol{S}^{(p)}(\boldsymbol{\beta}, t) - \boldsymbol{s}^{(p)}(\boldsymbol{\beta}, t)\| \rightarrow_{\mathscr{P}} 0.$$

   $s^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from zero and

   $$\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}} s^{(0)}(\boldsymbol{\beta}, t), \quad \boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial^2}{\partial \boldsymbol{\beta}^2} s^{(0)}(\boldsymbol{\beta}, t).$$

   Define $\boldsymbol{e} = \boldsymbol{s}^{(1)} = s^{(0)}$, $\boldsymbol{v} = \boldsymbol{s}^{(2)} = s^{(0)} - \boldsymbol{e}^2$. The matrix

   $$\boldsymbol{\Sigma} = \int_0^\infty \boldsymbol{v}(\beta, t) s^{(0)}(\boldsymbol{\beta}, t)\lambda_0(t) dt$$

   is positive definite.

2. Suppose that $S_0(t; \boldsymbol{Z}^t)$ and $G(t)$ are the continuous functions defined on $[0, \infty)$. The following condition is satisfied,

   $$\int_0^\infty [S_0(t; \boldsymbol{Z}^t)]^{-1} dG(t) < \infty.$$

> Vardi (1985, §8) and Keiding and Gill (1990, §6) required the similar condition for selection bias model and random truncation model, respectively. The above condition is the counterpart for the context with covariates.

We first study the asymptotic distribution of $\sqrt{n}\left(\hat{P}(\hat{\boldsymbol{\beta}})^{-1} - P(\boldsymbol{\beta})^{-1}\right)$. In the following context ≈ denotes asymptotic equivalence.

$$\sqrt{n}\left(\hat{P}(\hat{\boldsymbol{\beta}})^{-1} - P(\boldsymbol{\beta})^{-1}\right) \approx \sqrt{n}\int_0^\infty [S_0(t;\mathbf{Z}^t)]^{-1} d[\hat{G}^*(t) - G^*(t)] + \sqrt{n}\int_0^\infty \left[\frac{1}{\hat{S}_0(t;\mathbf{Z}^t)} - \frac{1}{S_0(t;\mathbf{Z}^t)}\right] dG^*(t) .$$

Let $W_1 = \sqrt{n}\int_0^\infty [S_0(t;\mathbf{Z}^t]^{-1} d[\hat{G}^*(t) - G^*(t)]$ and $W_2 = \sqrt{n}\int_0^\infty \left[\hat{S}_0(t;\mathbf{Z}^t)^{-1} - S_0(t;\mathbf{Z}^t)^{-1}\right] dG^*(t)$. By central limit theorem, given t, when $n \to \infty$, $\sqrt{n}\left(\hat{P}(\hat{\boldsymbol{\beta}})^{-1} - P(\boldsymbol{\beta})^{-1}\right)$ converges in distribution to a mean-zero normal random variable. Since W_1 and W_2 are asymptotically independent (Keiding and Gill, 1990), the variance of the limiting distribution is the sum of the asymptotic variances of W_1 and W_2. Let $\sigma^2_{P,1}$ and $\sigma^2_{P,2}$ be the asymptotic variance of W_1 and W_2, respectively. Since $\hat{G}^*$ is an empirical estimator of the truncated sample, it is straightforward to obtain

$$\begin{aligned}\sigma^2_{P,1} &\approx \int_0^\infty [S_0(t,\mathbf{Z}^t)]^{-2} dG^*(t) - \left[\int_0^\infty [S_0(t;\mathbf{Z}^t)]^{-1} dG^*(t)\right]^2 \\ &= P(\boldsymbol{\beta})^{-1}\int_0^\infty [S_0(t;\mathbf{Z}^t)]^{-1} dG(t) - P(\boldsymbol{\beta})^{-2} .\end{aligned}$$

Following the standard result of Cox model, we obtain the explicit expression of $\sigma^2_{P,2}$ but show the derivation in the appendix,

$$\sigma^2_{P,2} = \int_0^\infty \phi(u)^2 \frac{\lambda_0(u)du}{s^{(0)}(\boldsymbol{\beta},u)} + \boldsymbol{\kappa}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\kappa},$$

where

$$\phi(u) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} S_0(L_i^*;\mathbf{Z}_i)^{-1} I(u \le L_i^*) e^{\boldsymbol{\gamma}^T \mathbf{Z}_i}$$

and

$$\boldsymbol{\kappa} = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} S_0(L_i^*; \boldsymbol{Z}_i)^{-1} \boldsymbol{h}(L_i^*; \boldsymbol{Z}_i).$$

Using the generalized delta method, given t, the limiting distribution of $\sqrt{n}\left(\hat{P}(\hat{\boldsymbol{\beta}}) - P(\boldsymbol{\beta})\right)$ is normal with mean zero and the asymptotic variance $P(\boldsymbol{\beta})^4(\sigma_{P,1}^2 + \sigma_{P,2}^2)$ with the explicit expression

$$P(\boldsymbol{\beta})^3 \int_0^\infty [S_0(t; \boldsymbol{Z}^t)]^{-1} dG(t) - P(\boldsymbol{\beta})^2 + P(\boldsymbol{\beta})^4 \left(\int_0^\infty \phi(u)^2 \frac{\lambda_0(u)du}{s^{(0)}(\boldsymbol{\beta}, u)} + \boldsymbol{\kappa}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\kappa} \right).$$

Based on this asymptotic result, we estimate the variance of $\hat{P}(\hat{\boldsymbol{\beta}})$ by

$$\hat{\sigma}_P^2 = n^{-1} \hat{P}(\hat{\boldsymbol{\beta}})^3 \int_0^\infty [\hat{S}_0(t; \boldsymbol{Z}^t)]^{-1} d\hat{G}(t) - n^{-1} \hat{P}(\hat{\boldsymbol{\beta}})^2 + \hat{P}(\hat{\boldsymbol{\beta}}, \mathscr{Z})^4 \left(\int_0^\infty \hat{\phi}(u)^2 \frac{d\widehat{\Lambda}_0(u)}{nS^{(0)}(\hat{\boldsymbol{\beta}}, u)} + n^{-1} \hat{\boldsymbol{\kappa}}^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\kappa}} \right),$$

where

$$\hat{\phi}(u) = n^{-1} \sum_{i=1}^{n} \hat{S}_0(L_i^*, \boldsymbol{Z}_i)^{-1} I(u \le L_i^*) e^{\hat{\boldsymbol{\gamma}}^T \boldsymbol{Z}_i},$$

$$\hat{\boldsymbol{\kappa}} = n^{-1} \sum_{i=1}^{n} \hat{S}_0(L_i^*; \boldsymbol{Z}_i)^{-1} \hat{\boldsymbol{h}}(L_i^*; \boldsymbol{Z}_i),$$

$$\hat{\boldsymbol{h}}(t; z) = \int_0^t e^{\hat{\boldsymbol{\gamma}}^T z} \left(\begin{bmatrix} 0 \\ z \end{bmatrix} - \frac{\boldsymbol{S}^{(1)}(\hat{\boldsymbol{\beta}}, u)}{S^{(0)}(\hat{\boldsymbol{\beta}}, u)} \right) d\widehat{\Lambda}_0(u)$$

and $\widehat{\boldsymbol{\Sigma}} = n^{-1} \widehat{\mathscr{I}}(\hat{\boldsymbol{\beta}})$.

The $(1 - q)$100% linear confidence interval is given by

$$\hat{P}(\hat{\boldsymbol{\beta}}) \pm z_{1-q/2} \hat{\sigma}_P,$$

where z_q is the $(100q)$th percentile of the standard normal distribution. It is known that a linear confidence interval for a probability may not be constraint in the interval [0, 1]. A few transformed confidence intervals were demonstrated to have better performance. A commonly used one is the log-log transformed confidence interval (Borgan and Liestøl, 1990) with the formula

$$\exp\left\{\frac{\pm z_{1-q/2}\hat{\sigma}_P}{\hat{P}(\hat{\boldsymbol{\beta}})\log\left[\hat{P}(\hat{\boldsymbol{\beta}})\right]}\right\}.$$

In the remaining part of Section 2 we present a few probability estimators under different contexts. The linear and log-log transformed confidence intervals have similar forms as the above two equations. For simplicity the formulas of other confidence intervals are omitted.

2.2 To estimate distribution function of truncation variable

The estimator of L, $\hat{G}(t)$, is an IPW estimator. This type of estimator is widely used in various contexts. It is known that in survey statistics a selected observation should be inversely weighted by its probability of selection. The estimator was also studied by Vardi (1985) in the context that observations in a sample are subject to various known sampling probabilities. In the truncated sample the ith observation is selected to the sample by the probability $S_0(L_i^*; \boldsymbol{Z}_i)$. Therefore, the ith observation is inversely weighted by the estimated probability, $\hat{S}_0(L_i^*; \boldsymbol{Z}_i)$. The first term of the estimator is the normalizing term to produce proper probability estimate. In survey statistics and in the context studied by Vardi, the probability of selection is known. The above estimator differs from the traditional IPW estimator in that the probability of selection needs to be estimated.

Here we provide a brief derivation of the asymptotic distribution of $\hat{G}(t)$. The variation of our IPW estimator can be explained by two sources, the variation of an IPW estimator using known weight, and the variation due to weight estimation. We define an interim term

$$\hat{G}(t; \boldsymbol{\beta}) = \hat{P}(\boldsymbol{\beta})n^{-1}\sum_{i=1}^{n}\frac{I(L_i^* \le t)}{S_0(L_i^*; \boldsymbol{Z}_i)},$$

$$\hat{P}(\boldsymbol{\beta}) = \left[n^{-1}\sum_{i=1}^{n}\frac{1}{S_0(L_i^*; \boldsymbol{Z}_i)}\right]^{-1}.$$

Essentially, $\hat{G}(t; \boldsymbol{\beta})$ is an IPW estimator using known weights. Then,

$$\sqrt{n}\left[\hat{G}(t) - G(t)\right] = \sqrt{n}\left[\hat{G}(t; \boldsymbol{\beta}) - G(t)\right] + \sqrt{n}\left[\hat{G}(t) - \hat{G}(t; \boldsymbol{\beta})\right].$$

Let $K_1(t) = \sqrt{n}\left[\hat{G}(t; \boldsymbol{\beta}) - G(t)\right]$ and $K_2(t) = \sqrt{n}\left[\hat{G}(t) - \hat{G}(t; \boldsymbol{\beta})\right]$. First, we consider weak convergence of $K_1(t)$. Vardi (1985) studied the problem of estimating a distribution function when sampling weights are known. The proposed weighted estimator was proved to be maximum likelihood estimate, and the weak convergence result was sketched in the paper. Wang (1989) studied the semiparametric IPW estimator with an independently truncated sample, when the parametric distribution of the truncation variable is known. She explicitly decomposed the variation of the IPW estimator into two sources, the variation of the estimator using known weights, which agrees with Vardi's result, and the variation due to weight estimation. There is a high level of similarity between our IPW estimator and Wang's IPW estimator. According to Vardi (1985, Section 8), Wang (1989, Lemma 3.3) as well as derivation provided in Zhang (2012), we have the following convergence result. $\sqrt{n}\left[\widehat{G}(t;\boldsymbol{\beta}) - G(t)\right]$ converges in distribution to a normal variate with mean zero and variance

$$\sigma^2_{G,1}(t) = P(\boldsymbol{\beta})\int_0^t [S_0(u;\mathbf{Z}^u)]^{-1} dG(u) + P(\boldsymbol{\beta})G(t)^2\int_0^\infty [S_0(u;\mathbf{Z}^u)]^{-1} dG(u) - 2P(\boldsymbol{\beta})G(t$$
$$)\int_0^t [S_0(u;\mathbf{Z}^u)]^{-1} dG(u)$$

Here we sketch the derivation of weak convergence of $K_2(t) = \sqrt{n}\left[\widehat{G}(t) - \widehat{G}(t;\boldsymbol{\beta})\right]$.

$$\begin{aligned}
&\sqrt{n}\left[\widehat{G}(t) - \widehat{G}(t;\boldsymbol{\beta})\right] \\
&= n^{-1/2}\widehat{P}(\widehat{\boldsymbol{\beta}})\left[\sum_{i=1}^n \frac{I(L_i^* \le t)}{\widehat{S}_0(L_i^*,\mathbf{Z}_i)} - \sum_{i=1}^n \frac{I(L_i^* \le t)}{S_0(L_i^*;\mathbf{Z}_i)}\right] + n^{-1/2}\left[\widehat{P}(\widehat{\boldsymbol{\beta}}) - \widehat{P}(\boldsymbol{\beta})\right]\sum_{i=1}^n \frac{I(L_i^* \le t)}{S_0(L_i^*;\mathbf{Z}_i)} \\
&\approx n^{-1}P(\boldsymbol{\beta})\sum_{i=1}^n S_0(L_i^*;\mathbf{Z}_i)^{-1} I(L_i^* \le t)\sqrt{n}\left[\widehat{\Lambda}_0(L_i^*;\mathbf{Z}_i) - \Lambda_0(L_i^*;\mathbf{Z}_i)\right] \\
&-n^{-1}P(\boldsymbol{\beta})G(t)\sum_{i=1}^n S_0(L_i^*;\mathbf{Z}_i)^{-1}\sqrt{n}\left[\widehat{\Lambda}_0(L_i^*;\mathbf{Z}_i) - \Lambda_0(L_i^*;\mathbf{Z}_i)\right]
\end{aligned}$$

It can be further expressed as

$$\begin{aligned}
\sqrt{n}\left[\widehat{G}(t) - \widehat{G}(t;\boldsymbol{\beta})\right] &\approx n^{-1/2}P(\boldsymbol{\beta})\sum_{j=1}^n \int_0^\infty \{\eta(u,t) - G(t)\phi(u)\}\frac{dM_j(u)}{s^{(0)}(\beta,u)} \\
&+n^{-1/2}P(\boldsymbol{\beta})\,\{\boldsymbol{\rho}(t) - G(t)\times\boldsymbol{\kappa}\}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^n \int_0^\infty\left(\widetilde{\mathbf{Z}}_j - \frac{s^{(1)}(\beta,u)}{s^{(0)}(\beta,u)}\right)dM_j(u)\,.
\end{aligned}$$

where

$$\begin{aligned}
\eta(u,t) &= \lim_{n\to\infty} n^{-1}\sum_{i=1}^n S_0(L_i^*;\mathbf{Z}_i)^{-1} I(u \le L_i^* \le t)e^{\boldsymbol{\gamma}^T\mathbf{Z}_i}, \\
\boldsymbol{\rho}(t) &= \lim_{n\to\infty} n^{-1}\sum_{i=1}^n S_0(L_i^*;\mathbf{Z}_i)^{-1} I(L_i^* \le t)\boldsymbol{h}(L_i^*;\mathbf{Z}_i)\,.
\end{aligned}$$

Using the martingale central limit theorem, $\sqrt{n}\left[\widehat{G}(t) - \widehat{G}(t;\beta)\right]$ converges in distribution to a zero-mean normal variate with variance

$$\sigma_{G,2}^2(t) = P(\boldsymbol{\beta})^2 \int_0^\infty \{\eta(u,t) - G(t)\phi(u)\}^2 \frac{\lambda_0(u)du}{s^{(0)}(\beta,u)} + P(\boldsymbol{\beta})^2 \{\boldsymbol{\rho}(t) - G(t) \times \boldsymbol{\kappa}\}^T \boldsymbol{\Sigma}^{-1} \{\boldsymbol{\rho}(t) - G(t) \times \boldsymbol{\kappa}\}\,.$$

Based on the arguments used in Wang's derivation, we have the independence between $\sqrt{n}\left[\widehat{G}(t) - \widehat{G}(t;\boldsymbol{\beta})\right]$ and $\sqrt{n}\left[\widehat{G}(t;\boldsymbol{\beta}) - G(t)\right]$. Therefore, given t, $\sqrt{n}\left[\widehat{G}(t) - G(t)\right]$ converges in distribution to a zero-mean normal random variable, with the variance $\sigma_{G,1}^2(t) + \sigma_{G,2}^2(t)$. Based on this asymptotic result, we propose the following variance estimator for $\hat{G}(t)$,

$$\begin{aligned}
&\hat{\sigma}_G(t) = n^{-1}\widehat{P}(\widehat{\boldsymbol{\beta}}) \int_0^t \widehat{S}_0(u;\mathbf{Z}^u)^{-1} d\widehat{G}(u) + n^{-1}\widehat{P}(\widehat{\boldsymbol{\beta}})\widehat{G}(t)^2 \int_0^\infty \widehat{S}_0(u;\mathbf{Z}^u)^{-1} d\widehat{G}(u) \\
&-2n^{-1}\widehat{P}(\widehat{\boldsymbol{\beta}})\widehat{G}(t) \int_0^t \widehat{S}_0(u;\mathbf{Z}^u)^{-1} d\widehat{G}(u) \\
&+\widehat{P}(\widehat{\boldsymbol{\beta}})^2 \int_0^\infty \left\{\widehat{\eta}(u,t) - \widehat{G}(t)\widehat{\phi}(u)\right\}^2 \frac{d\widehat{\Lambda}_0(u)}{nS^{(0)}(\widehat{\boldsymbol{\beta}},u)} \\
&+n^{-1}\widehat{P}(\widehat{\boldsymbol{\beta}})^2 \left\{\widehat{\boldsymbol{\rho}}(t) - \widehat{G}(t) \times \widehat{\boldsymbol{\kappa}}\right\}^T \widehat{\boldsymbol{\Sigma}}^{-1} \left\{\widehat{\boldsymbol{\rho}}(t) - \widehat{G}(t) \times \widehat{\boldsymbol{\kappa}}\right\},
\end{aligned}$$

where

$$\widehat{\eta}(u,t) = n^{-1} \sum_{i=1}^n \widehat{S}_0(L_i^*;\mathbf{Z}_i)^{-1} I(u \le L_i^* \le t) e^{\widehat{\boldsymbol{\gamma}}^T \mathbf{Z}_i},$$

$$\widehat{\boldsymbol{\rho}}(t) = n^{-1} \sum_{i=1}^n \widehat{S}_0(L_i^*;\mathbf{Z}_i)^{-1} I(L_i^* \le t) \widehat{\boldsymbol{h}}(L_i^*;\mathbf{Z}_i)\,.$$

2.3 The estimators for right censored and left truncated sample

Here we focus on the scenario that T is also subject to right censoring. Only trivial extension should be made to the methods introduced in Section 2.2. Therefore, we directly provide relevant estimators. Let C be the censoring time variable. The truncated and censored sample is described as $\{ L_i^*, X_i^*, \Delta_i, \mathbf{Z}_i\}$, $i = 1, \cdots, n$, where $L_i^* < X_i^*, X_i^* = \min(T_i^*, C_i^*), \Delta_i = I(T_i^* \le C_i^*)$ and it remains the same $\widetilde{\mathbf{Z}}_i^T = \{L_i^* \mathbf{Z}_i^T\}$. Following the routine requirements for the context of left truncation and right censoring, we assume that C_i^* is independent of T_i^* given L_i^* and $P(L_i^* < C_i^*) = 1$. C_i^* and L_i^* do not necessarily need to be independent. The counting process notations should be revised, $N_{T,i}^C(x) = \Delta_i I(X_i^* \le x)$, $Y_i^C(x) = I(L_i^* \le x \le X_i^*)$,

$$\mathbf{S}_C^{(p)}(\boldsymbol{\beta},t) = n^{-1} \sum_{i=1}^n Y_i^C(t) \widetilde{\mathbf{Z}}_i^{\otimes p} \exp(\boldsymbol{\beta}^T \widetilde{\mathbf{Z}}_i), \quad p = 0, 1, 2.$$

The estimating equation becomes

$$\mathcal{U}_C(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\infty \left[\widetilde{\mathbf{Z}}_i - \frac{\sum_{j=1}^{n} Y_j^C(t) \widetilde{\mathbf{Z}}_j^* \exp(\boldsymbol{\beta}^T \widetilde{\mathbf{Z}}_j)}{\sum_{j=1}^{n} Y_j^C(t) \exp(\boldsymbol{\beta}^T \widetilde{\mathbf{Z}}_j)} \right] dN_{T,i}^C(t).$$

Let $\hat{\boldsymbol{\beta}}_C^T = \{\hat{\alpha}_C \hat{\boldsymbol{\gamma}}_C^T\}$ be the MLE that solves $\mathcal{U}_C(\boldsymbol{\beta}) = 0$. The estimated information matrix is expressed as

$$\widehat{\mathcal{I}}_C(\hat{\boldsymbol{\beta}}_C) = \sum_{i=1}^{n} \int_0^\infty \left[\frac{\boldsymbol{S}_C^{(2)}(\hat{\boldsymbol{\beta}}_C, t)}{S_C^{(0)}(\hat{\boldsymbol{\beta}}_C, t)} - \left(\frac{\boldsymbol{S}_C^{(1)}(\hat{\boldsymbol{\beta}}_C, t)}{S_C^{(0)}(\hat{\boldsymbol{\beta}}_C, t)} \right)^2 \right] dN_{T,i}^C(t).$$

The Breslow estimator for the cumulative baseline hazard function, $\Lambda_0(t)$, is given by

$$\widehat{\Lambda}_0^C(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_{T,i}^C(u)}{\sum_{j=1}^{n} Y_j^C(u) \exp(\hat{\boldsymbol{\beta}}_C^T \widetilde{\mathbf{Z}}_j)}.$$

Let $\hat{S}_0^C(L_i^*; \mathbf{Z}_i) = \exp\left\{ -\widehat{\Lambda}_0^C(L_i^*) e^{\hat{\boldsymbol{\gamma}}_C^T \mathbf{Z}_i} \right\}$. The probability of selection and the distribution function of L can be respectively estimated by

$$\hat{P}_C(\hat{\boldsymbol{\beta}}_C) = \left[n^{-1} \sum_{i=1}^{n} \frac{1}{\hat{S}_0^C(L_i^*; \mathbf{Z}_i)} \right]^{-1}$$

and

$$\hat{G}_C(x) = \left[n^{-1} \sum_{i=1}^{n} \frac{1}{\hat{S}_0^C(L_i^*; \mathbf{Z}_i)} \right]^{-1} \times n^{-1} \sum_{i=1}^{n} \frac{I(L_i^* \le x)}{\hat{S}_0^C(L_i^*; \mathbf{Z}_i)}.$$

The variance of estimated probability of selection can be estimated by

$$\hat{\sigma}_{P,C}^2 = n^{-1} \hat{P}_C(\hat{\boldsymbol{\beta}}_C)^3 \int_0^\infty [\hat{S}_0(t; \mathbf{Z}^t)]^{-1} d\hat{G}_C(t) - n^{-1} \hat{P}_C(\hat{\boldsymbol{\beta}}_C)^2 + \hat{P}_C(\hat{\boldsymbol{\beta}}_C)^4$$
$$\left(\int_0^\infty \hat{\phi}_C(u)^2 \frac{d\widehat{\Lambda}_0^C(u)}{n S_C^{(0)}(\hat{\boldsymbol{\beta}}_C, u)} + n^{-1} \hat{\boldsymbol{\kappa}}_C^T \widehat{\boldsymbol{\Sigma}}_C^{-1} \hat{\boldsymbol{\kappa}}_C \right),$$

where

$$\hat{\phi}_C(u) = n^{-1}\sum_{i=1}^{n}\hat{S}_0^C(L_i^*;\mathbf{Z}_i)^{-1} I(u \le L_i^*)e^{\hat{\boldsymbol{\gamma}}_C^T\mathbf{Z}_i},$$

$$\hat{\boldsymbol{\kappa}}_C = n^{-1}\sum_{i=1}^{n}\hat{S}_0^C(L_i^*;\mathbf{Z}_i)^{-1}\hat{\boldsymbol{h}}_C(L_i^*;\mathbf{Z}_i),$$

$$\hat{\boldsymbol{h}}_C(t;z) = \int_0^t e^{\hat{\boldsymbol{\gamma}}_C^T z}\left(\begin{bmatrix}0\\ z\end{bmatrix} - \frac{\boldsymbol{S}_C^{(1)}(\hat{\boldsymbol{\beta}}_C,u)}{S_C^{(0)}(\hat{\boldsymbol{\beta}}_C,u)}\right)d\widehat{\Lambda}_0^C(u)$$

and $\widehat{\boldsymbol{\Sigma}}_C = n^{-1}\widehat{\mathscr{I}}_C(\hat{\boldsymbol{\beta}}_C)$.

The variance estimator for $\hat{G}_C(t)$ has the explicit express as follows,

$$\begin{aligned}
&\hat{\sigma}_{G,C}^2(t) = n^{-1}\hat{P}_C(\hat{\boldsymbol{\beta}}_C)\int_0^t\left[\hat{S}_0^C(u;\mathbf{Z}^u)\right]^{-1}d\hat{G}_C(u)\\
&+n^{-1}\hat{P}_C(\hat{\boldsymbol{\beta}}_C)\hat{G}_C(t)^2\int_0^\infty\left[\hat{S}_0^C(u;\mathbf{Z}^u)\right]^{-1}d\hat{G}_C(u)\\
&-2n^{-1}\hat{P}_C(\hat{\boldsymbol{\beta}}_C)\hat{G}_C(t)\int_0^t\left[\hat{S}_0^C(u;\mathbf{Z}^u)\right]^{-1}d\hat{G}_C(u)\\
&+\hat{P}_C(\hat{\boldsymbol{\beta}}_C)^2\int_0^\infty\left\{\hat{\eta}_C(u,t) - \hat{G}_C(t)\hat{\phi}_C(u)\right\}^2\frac{d\widehat{\Lambda}_0^C(u)}{nS_C^{(0)}(\hat{\boldsymbol{\beta}}_C,u)}\\
&+n^{-1}\hat{P}_C(\hat{\boldsymbol{\beta}}_C)^2\left\{\hat{\boldsymbol{\rho}}_C(t) - \hat{G}_C(t)\times\hat{\boldsymbol{\kappa}}_C\right\}^T\widehat{\boldsymbol{\Sigma}}_C^{-1}\left\{\hat{\boldsymbol{\rho}}_C(t) - \hat{G}_C(t)\times\hat{\boldsymbol{\kappa}}_C\right\},
\end{aligned}$$

where

$$\hat{\eta}_C(u,t) = n^{-1}\sum_{i=1}^{n}\hat{S}_0^C(L_i^*;Z_i)^{-1}I(u \le L_i^* \le t)e^{\hat{\boldsymbol{\gamma}}_C^T\mathbf{Z}_i},$$

$$\hat{\boldsymbol{\rho}}_C(t) = n^{-1}\sum_{i=1}^{n}\hat{S}_0^C(L_i^*;Z_i)^{-1}I(L_i^* \le t)\hat{\boldsymbol{h}}_C(L_i^*;\mathbf{Z}_i).$$

2.4 The estimators with stratified censored and truncated sample

Here we provide the estimation methods with a stratified sample for which the distribution function of the truncation variable varies between the strata. The stratified censored and truncated sample is summarized as $\{L_{ki}^*, X_{ki}^*, \Delta_{ki}, \mathbf{Z}_{ki}\}$, $k = 1, \cdots, K$; $i = 1, \cdots, n$, $L_{ki}^* < X_{ki}^*$. We assume the same relationship between L and T as specified in Model (1). The Cox model related estimators remain the same as if no strata appear in the sample. We still use the notations $\hat{\boldsymbol{\beta}}_C$, $\widehat{\Lambda}_0^C$, $\hat{\mathfrak{Q}}_C(\hat{\boldsymbol{\beta}}_C)$, $\hat{\boldsymbol{\Sigma}}_C$, $\hat{S}_0^C$ and $\hat{\boldsymbol{h}}_C$ for the stratified sample.

Let G_k be the distribution function of the truncation variable in the kth stratum. The probability of selection may vary by stratum. The probability of selection of the kth stratum is denoted as $P^k(\boldsymbol{\beta})$.

The estimators for $P^k(\boldsymbol{\beta})$ and G_k with the stratified sample are given by

$$\hat{P}_C^k(\hat{\boldsymbol{\beta}}_C) = \left[n_k^{-1} \sum_{i=1}^{n_k} \frac{1}{\hat{S}_0^C(L_{ki}^*; \mathbf{Z}_{ki})} \right]^{-1}, \quad (3)$$

and

$$\hat{G}_{C,k}(t) = \left[n_k^{-1} \sum_{i=1}^{n_k} \frac{1}{\hat{S}_0^C(L_{ki}^*; \mathbf{Z}_{ki})} \right]^{-1} \times n_k^{-1} \sum_{i=1}^{n_k} \frac{I(L_{ki}^* \le t)}{\hat{S}_0^C(L_{ki}^*; \mathbf{Z}_{ki})}. \quad (4)$$

The variance of estimated probability of selection is estimated by

$$\hat{\sigma}_{P,C,k}^2 = n_k^{-1} \hat{P}_C^k(\hat{\boldsymbol{\beta}}_C)^3 \int_0^\infty [\hat{S}_0(t; \mathbf{Z}^t)]^{-1} d\hat{G}_{C,k}(t) - n_k^{-1} \hat{P}_C^k(\hat{\boldsymbol{\beta}}_C)^2$$
$$+ n^{-1} \hat{P}_C^k(\hat{\boldsymbol{\beta}}_C)^4 \left(\int_0^\infty \hat{\phi}_{C,k}(u)^2 \frac{d\widehat{\Lambda}_0^C(u)}{S_C^{(0)}(\hat{\boldsymbol{\beta}}_C, u)} + \hat{\boldsymbol{\kappa}}_{C,k}^T \widehat{\boldsymbol{\Sigma}}_C^{-1} \hat{\boldsymbol{\kappa}}_{C,k} \right),$$

where

$$\hat{\phi}_{C,k}(u) = n_k^{-1} \sum_{i=1}^{n_k} \hat{S}_0^C(L_{ki}^*; \mathbf{Z}_{ki})^{-1} I(u \le L_{ki}^*) e^{\hat{\boldsymbol{\gamma}}_C^T \mathbf{Z}_{ki}},$$
$$\hat{\boldsymbol{\kappa}}_{C,k} = n_k^{-1} \sum_{i=1}^{n_k} \hat{S}_0^C(L_{ki}^*; \mathbf{Z}_{ki})^{-1} \hat{\boldsymbol{h}}_C(L_{ki}^*; \mathbf{Z}_{ki}),$$

The variance estimator for $\hat{G}_C(t)$ is given by

$$\begin{aligned}
&\hat{\sigma}^2_{G,C,k}(t) = n_k^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)\int_0^t \hat{S}^C_0(u;\mathbf{Z}^u)^{-1} d\hat{G}_{C,k}(u) \\
&+n_k^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)\hat{G}_{C,k}(t)^2\int_0^\infty \hat{S}^C_0(u;\mathbf{Z}^u)^{-1} d\hat{G}_{C,k}(u) \\
&-2n_k^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)\hat{G}_{C,k}(t)\int_0^t \hat{S}^C_0(u;\mathbf{Z}^u)^{-1} d\hat{G}_{C,k}(u) \\
&+n^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)^2\int_0^\infty \left\{\hat{\eta}_{C,k}(u,t) - \hat{G}_{C,k}(t)\hat{\phi}_{C,k}(u)\right\}^2 \frac{d\widehat{\Lambda}^C_0(u)}{S^{(0)}_C(\hat{\boldsymbol{\beta}}_C,u)} \\
&+n^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)^2\left\{\hat{\boldsymbol{\rho}}_{C,k}(t) - \hat{G}_{C,k}(t)\times\hat{\boldsymbol{\kappa}}_{C,k}\right\}^T \widehat{\boldsymbol{\Sigma}}_C^{-1}\left\{\hat{\boldsymbol{\rho}}_{C,k}(t) - \hat{G}_{C,k}(t)\times\hat{\boldsymbol{\kappa}}_{C,k}\right\},
\end{aligned}$$

where

$$\hat{\eta}_{C,k}(u,t) = n_k^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)\sum_{i=1}^{n_k}\hat{S}^C_0(L^*_{ki};\mathbf{Z}_{ki})^{-1} I(u \le L^*_{ki} \le t)e^{\hat{\boldsymbol{\gamma}}^T_C \mathbf{Z}_{ki}},$$

$$\hat{\boldsymbol{\rho}}_{C,k}(t) = n_k^{-1}\hat{P}^k_C(\hat{\boldsymbol{\beta}}_C)\sum_{i=1}^{n_k}\hat{S}^C_0(L^*_{ki};\mathbf{Z}_{ki})^{-1} I(L^*_{ki} \le t)\hat{\boldsymbol{h}}_C(L^*_{ki};\mathbf{Z}_{ki}).$$

3 The simulation study

We wished to evaluate the practical performance of the proposed IPW estimator of $G(t)$ and the variance estimator. We considered the scenario that L is a predictor of T and a fixed covariate z is also associated with T. The following model was assumed

$$\lambda_T(t;L,z) = \begin{cases} \lambda_0(t)\exp(\gamma z) & \text{if } t < L \\ \lambda_0(t)\exp(\alpha L + \gamma z) & \text{if } t \ge L \end{cases}. \quad (5)$$

The truncation variable L was generated from a uniform distribution in the interval [0,1]. The baseline hazard function in the above model was set to a constant and we searched different values to control the censoring and truncation rates. We considered both positive and negative regression coefficients (α = 0.02 and −0.05) for L. When α is positive, increment in L escalates the risk of failure. When α is negative, increment in L prevents occurrence of failure. Settings with continuous covariate or discrete covariate were both generated. The continuous covariate was generated from a truncated standard normal distribution, restraining in the interval [−3, 3]. Discrete covariate was generated from a Bernoulli distribution with parameter value 0.5. The regression coefficient associated with the covariate was set to 0.5. We also considered setting of two fixed covariates with the underlying model

$$\lambda_T(t; L, z) = \begin{cases} \lambda_0(t)\ \exp\left(\gamma_1 z_1 + \gamma_2 z_2\right) & \text{if } t < L \\ \lambda_0(t)\ \exp\left(\alpha L + \gamma_1 z_1 + \gamma_2 z_2\right) & \text{if } t \geq L \end{cases}. \qquad (6)$$

The first covariate was generated from a standard normal distribution truncated in the interval [−3, 3]. The second covariate is binary with probability 0.5 for taking value 1. We set $(\gamma_1,\ \gamma_2) = (0.5, -0.3)$.

We considered two levels for the truncation rate (25%, 50%) and the censoring rate (25%, 50%). The censoring time was generated from the uniform distribution in the interval [0, *a*] and we only kept the value if it was greater than the value of the truncation variable. We adjusted the values of *a* to control the censoring rate. In simulated settings value of *a* varied from 1.45 to 8.6 to generate censoring rates (25%, 50%). In each setting, there were 1000 replicates with fixed sample size 200. In this simulation study, $\hat{\alpha}$, $\hat{\beta}$ and $\widehat{\Lambda}_0^C$ were evaluated and were shown to have satisfactory performance. Since they are the standard estimators for a truncated version of the Cox model, we decided not to show the simulation results for these estimators. We focused on the estimators $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ and report the simulation results at three time points 0.25, 0.5 and 0.75, leading to 0.25, 0.5 and 0.75 in $G(t)$. We calculated $\hat{G}_C$, $\hat{\sigma}_{G,G}$, as well as 95% linear and log-log transformed confidence intervals for each replicate, and then evaluated the following terms,

$$\overline{G}_C(t) = \frac{1}{1000}\sum_{i=1}^{1000} \widehat{G}_C^{(i)}(t),$$

$$\text{Bias} = \overline{G}_C(t) - G(t),$$

$$\text{var}\left[\widehat{G}_C(t)\right] = \frac{1}{1000-1}\sum_{i=1}^{n}\left[\widehat{G}_C^{(i)}(t) - \overline{G}_C(t)\right]^2,$$

$$\widehat{\text{var}}\left[\widehat{G}_C(t)\right] = \frac{1}{1000}\sum_{i=1}^{1000}\left[\hat{\sigma}_{G,C}^{(i)}(t)\right]^2,$$

where $\widehat{G}_C^{(i)}$ and $\hat{\sigma}_{G,C}^{(i)}$ are the point and precision estimates for the *i*th replicate using the estimators given in Section 2.3. We also calculated the actual proportion of replicates that each type of confidence interval covered $G(t)$.

The simulation results for the setting with continuous covariate and discrete covariate are given in Tables 1–2 and Tables 3–4, respectively. Tables 5 and 6 depict the simulation results for the settings with both continuous and discrete covariates. From the tables, we can see that the bias is very small. The estimated variances are close to the sample variances. The sample variance increases when the truncation rate or censoring rate becomes higher. Compared to the linear confidence interval, the log-log transformed confidence interval has clearly better performance and its coverage is close to the confi-dence level. However, undercoverage is observed when both the censoring and truncation rates are high. When analyzing a read data set, one can estimate the truncation rate $1 - \hat{P}_C(\hat{\boldsymbol{\beta}}_C)$. If heavy censoring and truncation is present, one may consider to use a bootstrap confidence interval.

4 The Bone Marrow Transplant Example

4.1 Data Description

In this section we analyze the transplant outcome data set from The Center for International Blood and Marrow Transplant Research (CIBMTR). The CIBMTR is comprised of clinical and basic scientists who confidentially share data on their blood and bone marrow transplant patients with CIBMTR Data Collection Center located at the Medical College of Wisconsin. The CIBMTR is a repository of information about results of transplants at more than 450 transplant centers worldwide. In this example 376 children receiving transplantation in second complete remission are selected. Since only the patients who received transplants are included in the registry but not patients who died while waiting for transplantation, the BMT sample is a truncated sample. Among 376 children 159 were alive in remission at the cutoff date of study, leading to a censoring rate of 42%. For a disease-free survivor the follow-up time till study cutoff date was the censoring time.

The BMT sample, jointly with a sample of 529 children receiving chemotherapy, was analyzed by Barrett et al. (1994) to evaluate the treatment efficacy on the leukemia-free survival. Cox analysis was performed on each sample to identify the significant risk factors at 0.10 levels. The following factors were identified to be associated with leukemia-free survival using the BMT sample: age ($>$10 yr, ≤ 10 yr), the T-cell phenotype (no, yes), duration of the first remission (≤ 18 months; $>$18 months). We present Barrett's Cox analysis result in Table 5, which will be later compared to the result of our new Cox model including the waiting time to transplant as covariate.

4.2 Effects of waiting time to transplant and other covariates

We added $\alpha g(L)$ into the regressor of Cox model, where L is the waiting time to transplant p and $g(L)$ is a function. We considered the following functional forms, L, L^2, e^L and $\sqrt{L}$. It turned out the quadratic form L^2 yielded the highest level of significance. Therefore, we chose to include the quadratic waiting time in the regressor. A model-building procedure was performed to search for other significant risk factors with p-value 0.05 as the threshold, and age, duration of first remission, T-cell phenotype were selected. The estimated regression coefficients are shown in Table 5, together with those in Barrett's study. Time-dependent variable was created for each covariate and temporarily included in the Cox model to test the proportional hazards assumption. Proportionality approximately held for the variables included in the final Cox model. Age greater than 10 years, T-cell phenotype, and duration of the first remission ≤ 18 months are associated with higher risks of disease relapse or death. Regarding the waiting time, the regression coefficient associated with L^2 is 0.0021 (RR=1.002, P=0.030), indicating that a patient with a longer waiting time for transplantation is more likely to experience relapse or death. The finding that a longer waiting time is a poor prognosis of leukemia-free survival agrees well with the recent clinical observation (Balduzzi, 2008).

4.3 The distribution function of waiting time to transplant

There have been a lot of articles discussing disparities in organ and bone marrow transplantations. Public health researchers are interested in discovering racial, geographic and social economic disparities in receiving transplant and waiting time for transplant. In this example we examined all the covariates and found out that the distribution of the waiting time may differ by the duration of the first remission. The cohort consisted of 124 children staying in the first remission 18 months or less and 252 children with longer than 18 months in the first remission. We estimated the distribution function of the waiting time in these two subgroups using the estimator $\hat{G}_{C,k}$ (Eq (4)) for the stratified sample (Section 2.4). The estimated curves are depicted in Figure 1 showing that patients with shorter duration of first remission waited less for their transplants. The median waiting time for these two subgroups were 2.1 and 3.2 months, respectively. By 6 months in the second remission, 88% of children no more than 18 months in first remission underwent transplants (95% log-log CI 78%–94%) while the proportion was reduced to 77% for children in the other subgroup (95% log-log CI 70%–83%).

We also wished to discover whether the proportion of getting transplant differed by duration of the first remission. We estimated the probability of selection in each subgroup using the estimators $\hat{P}_C^k(\hat{\boldsymbol{\beta}}_C)$(Eq (3)) for the stratified sample (Section 2.4). It turned out that the estimated probabilities yielded from these two subgroups were very close. 83% of children with no more than 18 months in first remission had transplants (95% log-log CI 73%–90%). In children with relatively long remission time, 82% had transplants (95% log-log CI 77%–87%).

We used this example to illustrate the inferences for the stratified sample. It would be interesting to explore the reason how duration of the first remission influenced the waiting time to transplant. Because there are only limited number of covariates, we could barely find meaningful explanations for this study. The proposed methods can be applied in other transplant registry data. The methods are useful in evaluating the waiting time in racial, geographic and social economic subgroups.

5 Final discussion

It is challenging to deal with dependently truncated sample because the dependence pattern is versatile. It is a simple and clever solution to model dependence that the truncation variable is used as a covariate in a Cox model. This idea was suggested a long time ago by Jones and Crowley (1992) and Shen (2003). It was only recently that the applications in registry data were formally discussed and inferences were developed (MacKenzie, 2012; Zhang et al., 2015). MacKenzie and Zhang et al. studied the inferences for a Cox model with the truncation variable as the only covariate. Inclusion of other covariates in the Cox model is practically more meaningful and this motivated us to study the estimation methods under such a Cox model.

A series of estimators have been introduced in this paper. We first presented the estimators for the left truncated only context, and then extended the estimation methods to the right censored and left truncated context, which is more commonly seen in survival data. We also provided the estimators with the stratified data, when the distribution of the truncation variable varies between the strata. The methods can be used to evaluate whether the chance of entrance and the time of entry are homogeneous across the strata, which is a usual topic in health disparities research. In this paper application of the proposed methods has been illustrated by the BMT registry data. There are other real-life applications of the developed inferences. It is known that, when subjects enter a study at random age, age-specific mortality is left truncated by age at entry (Klein and Moeschberger, 2003). In addition, age has important influence on survival. In this sense, the Cox model with age as both truncation variable and covariate has been well accepted by researchers. The methods developed in this paper are useful for the studies in which survival outcomes are present and cohorts consist of subjects entering at random age.

References

1. Andersen, PK., Borgan, Ø., Gill, RD., Keiding, N. Statistical Models Based on Counting Processes. Springer; New York: 1993.
2. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The Annals of Statistics. 1982; 10:1100–1120.
3. Balduzzi A, De Lorenzo P, Schrauder A, Conter V, Uderzo C, Peters C, Klingebiel T, Stary J, Felice MS, Magyarosy E, et al. Eligibility for allogeneic transplantation in very high risk childhood acute lymphoblastic leukemia: the impact of the waiting time. Haematologica. 2008; 93:925–929. [PubMed: 18413892]
4. Barrett AJ, Horowitz MM, Pollock BH, Zhang MJ, Bortin MM, Buchanan GR, Camitta BM, Ochs J, Graham-Pole J, Rowling PA, Rimm AA, Klein JP, Shuster JJ, Sobocinski KA, Gale RP. HLA-identical Sibling Bone Marrow Transplants versus Chemotherapy for Children with Acute Lymphoblas-tic Leukemia in Second Re-mission. The New England Journal of Medicine. 1994; 331:1253–1258. [PubMed: 7935682]
5. Borgan Ø, Liestøl K. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. Scandinavian Journal of Statistics. 1990; 17:35–41.
6. Cox DR. Partial likelihood. Biometrika. 1975; 62:269–276.
7. Efron B, Petrosian V. A simple test of independence for truncated data with applications to redshift surveys. The Astrophysical Journal. 1992; 399:345–352.
8. Finkelstein DM, Moore DF, Schoenfeld DA. A propotional hazards model for truncated AIDS data. Biometrics. 1993:731–740. [PubMed: 8241369]
9. Gross ST, Huber-Carol C. Regression models for truncated survival data. Scandinavian Journal of Statistics. 1992; 19:192–213.
10. Jones MP, Crowley J. Nonparametric tests of the Markov model for survival data. Biometrika. 1992; 79:513–522.
11. Kalbfleisch JD, Lawless JF. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. Statistica Sinica. 1991; 1:19–32.
12. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Statist Assoc. 1958; 84:360–372.
13. Keiding N, Gill RD. Random truncation models and Markov process. The Annals of Statistics. 1990; 18:582–602.
14. Klein, JP., Moeschberger, ML. Survival analysis techniques for censored and truncated data. New York: Springer-Verlag; 2003.
15. Klein JP, Zhang MJ. Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia. Life data: models in reliability and survival analysis. 1996:175.

16. Mackenzie T. Survival curve estimation with dependent left truncated data using Cox's model. The International Journal of Biostatistics. 2012; 8(1)
17. Shen PS. The product-limit estimate as an inverse-probability-weighted average. Communications in Statistics - Theory and Methods. 2003; 32:1119–1133.
18. Tsai WY. Testing the assumption of the independence of truncation time and failure time. Biometrika. 1990; 77:169–177.
19. Wang MC, Jewell NP, Tsai WY. Asymptotic properties of the product limit estimate under random truncation. The Annals of Statistics. 1986; 14:1597–1605.
20. Wang MC. a semiparametric model for randomly truncated data. J Am Statist Assoc. 1989; 84:742–748.
21. Woodroofe M. Estimating a distribution function with truncated data. The Annals of Statistics. 1985; 13:163–177.
22. Zhang X. Nonparametric inference for inverse probability weighted estimators with a randomly truncated sample. Journal of Data Science. 2012; 10:673–691.
23. Zhang X, Li J, Liu Y. Inference for probability of selection with dependently truncated data using a Cox model. Communications in Statistics - Simulation and Computation. 2015; doi: 10.1080/03610918.2015.1024856
24. Vardi Y. Empirical distributions in selection bias models. The Annals of Statistics. 1985; 13:178–203.

Appendix

This appendix presents the derivation of the asymptotic distribution of W_2 (Section 2.1). Under the Cox model $\lambda_T(t; L, \mathbf{z}) = \lambda_0(t)e^{\alpha L + \boldsymbol{\gamma}^T \mathbf{z}}$,

$$M_i(t) = N_{T,i}(t) - \int_0^t Y_i(s) e^{\alpha L_i^* + \boldsymbol{\gamma}^T \mathbf{Z}_i} \lambda_0(s) ds$$

is a martingale. Let $\Lambda_0(t; z) = \Lambda(t; L = 0, z) = \int_0^t \lambda_T(u; L = 0, z) du$. W_2 can be expressed as

$$n^{-1/2} \sum_{i=1}^{n} \left[\frac{1}{\hat{S}_0(L_i^*; Z_i)} - \frac{1}{S_0(L_i^*; Z_i)} \right].$$

Applying the generalized delta method, we have

$$W_2 \approx n^{-1} \sum_{i=1}^{n} S_0(L_i^*; \mathbf{Z}_i)^{-1} \sqrt{n} \left[\hat{\Lambda}_0(L_i^*; \mathbf{Z}_i) - \Lambda_0(L_i^*; \mathbf{Z}_i) \right].$$

Using the standard result of a Cox model (Andersen and Gill, 1982),

$$\sqrt{n}\left[\widehat{\Lambda}_0(L_i^*;\boldsymbol{Z}_i)-\Lambda_0(L_i^*;\boldsymbol{Z}_i)\right]=\sqrt{n}\left[\widehat{\Lambda}(L_i^*;L_i^*=0,\boldsymbol{Z}_i)-\Lambda_0(L_i^*;L_i^*=0,\boldsymbol{Z}_i)\right]$$
$$\approx n^{-1/2}\left[\sum_{j=1}^{n}\int_0^{L_i^*}e^{\boldsymbol{\gamma}^T\boldsymbol{Z}_i}\frac{dM_j(u)}{S^{(0)}(\beta,u)}+\boldsymbol{h}(L_i^*;\boldsymbol{Z}_i)\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\int_0^{\infty}\left(\widetilde{\boldsymbol{Z}}_j-\frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta},u)}{S^{(0)}(\boldsymbol{\beta},u)}\right)dM_j(u)\right],$$

where

$$\boldsymbol{h}(t;z)=\int_0^t e^{\boldsymbol{\gamma}^T z}\left(\begin{bmatrix}0\\z\end{bmatrix}-\frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta},u)}{S^{(0)}(\boldsymbol{\beta},u)}\right)d\Lambda_0(u)\,.$$

Based on the result for $\sqrt{n}\left[\widehat{\Lambda}_0(L_i^*;\boldsymbol{Z}_i)-\Lambda_0(L_i^*;\boldsymbol{Z}_i)\right]$, W_2 can be further expressed as

$$W_2\approx n^{-1/2}\sum_{j=1}^{n}\int_0^{\infty}\phi(u)\frac{dM_j(u)}{S^{(0)}(\boldsymbol{\beta},u)}+n^{-1/2}\boldsymbol{\kappa}^T\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\int_0^{\infty}\left(\widetilde{\boldsymbol{Z}}_j-\frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta},u)}{S^{(0)}(\boldsymbol{\beta},u)}\right)dM_j(u)\,.$$

Using martingale central limit theorem, the limiting distribution of W_2 is normal with variance

$$\sigma_{P,2}^2=\int_0^{\infty}\phi(u)^2\frac{\lambda_0(u)du}{s^{(0)}(\boldsymbol{\beta},u)}+\boldsymbol{\kappa}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}\,.$$

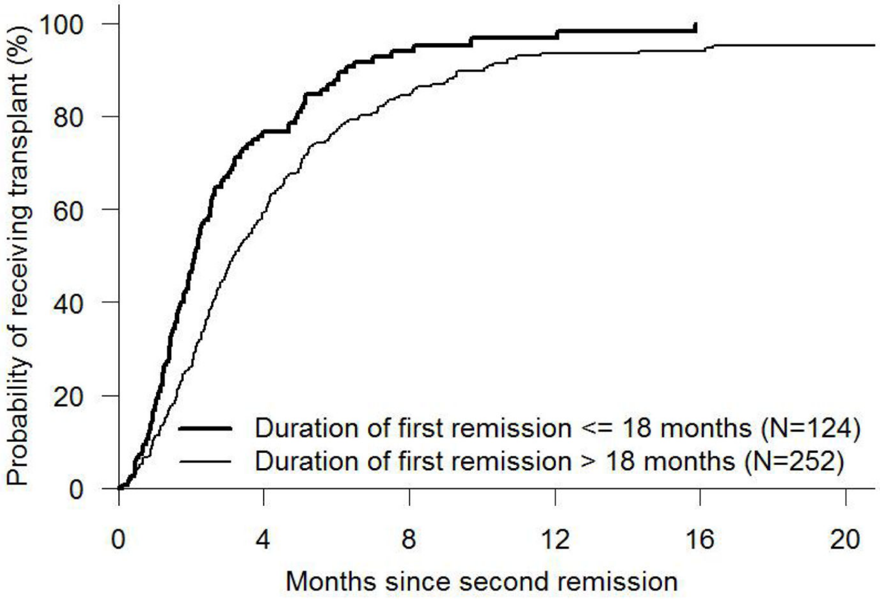

Figure 1.

Estimated distribution functions of transplant waiting time for duration of first remission <= 18 months and > 18 months, respectively

Table 1

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model includes two covariates, L and continuous z, with regression coefficients 0.02 and 0.5, respectively.

(L%, C%)	**$G(t)$**	**Bias**	**var[$\hat{G}_C(t)$]**	$\widehat{\text{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	−0.001	0.0010	0.0010	0.941	0.946
	0.50	−0.002	0.0017	0.0016	0.951	0.950
	0.75	−0.002	0.0015	0.0014	0.961	0.942
(25, 50)	0.25	−0.002	0.0010	0.0010	0.940	0.947
	0.50	−0.001	0.0017	0.0017	0.943	0.946
	0.75	−0.001	0.0015	0.0015	0.950	0.951
(50, 25)	0.25	0.000	0.0016	0.0019	0.940	0.941
	0.50	−0.001	0.0035	0.0036	0.922	0.927
	0.75	0.002	0.0033	0.0037	0.929	0.937
(50, 50)	0.25	−0.002	0.0018	0.0017	0.920	0.925
	0.50	−0.004	0.0044	0.0042	0.917	0.920
	0.75	−0.001	0.0043	0.0043	0.909	0.922

Table 2

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model includes two covariates, L and continuous z, with regression coefficients −0.05 and 0.5, respectively.

(L%, C%)	**$G(t)$**	**Bias**	**var[$\hat{G}_C$ (t)]**	$\widehat{\text{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	0.000	0.0009	0.0010	0.964	0.966
	0.50	0.003	0.0015	0.0016	0.944	0.947
	0.75	0.000	0.0014	0.0014	0.951	0.941
(25, 50)	0.25	0.000	0.0010	0.0011	0.949	0.955
	0.50	−0.001	0.0017	0.0019	0.949	0.953
	0.75	−0.002	0.0016	0.0019	0.967	0.953
(50, 25)	0.25	0.000	0.0017	0.0015	0.933	0.941
	0.50	0.000	0.0038	0.0034	0.923	0.932
	0.75	0.000	0.0036	0.0036	0.942	0.940
(50, 50)	0.25	−0.002	0.0018	0.0017	0.932	0.940
	0.50	−0.004	0.0040	0.0040	0.928	0.939
	0.75	−0.004	0.0038	0.0043	0.929	0.940

Table 3

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model includes two covariates, L and discrete z, with regression coefficients 0.02 and 0.5, respectively.

(L%, C%)	**$G(t)$**	**Bias**	**var[$\hat{G}_C(t)$]**	$\widehat{\text{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	0.001	0.0009	0.0010	0.950	0.958
	0.50	−0.001	0.0015	0.0015	0.958	0.958
	0.75	−0.001	0.0013	0.0013	0.964	0.954
(25, 50)	0.25	0.001	0.0010	0.0010	0.962	0.966
	0.50	0.001	0.0016	0.0016	0.940	0.950
	0.75	0.000	0.0013	0.0013	0.950	0.941
(50, 25)	0.25	−0.002	0.0013	0.0014	0.953	0.959
	0.50	−0.002	0.0028	0.0027	0.941	0.940
	0.75	−0.005	0.0028	0.0026	0.948	0.940
(50, 50)	0.25	−0.002	0.0015	0.0016	0.952	0.961
	0.50	−0.003	0.0034	0.0031	0.934	0.939
	0.75	−0.004	0.0034	0.0030	0.937	0.937

Table 4

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model includes two covariates, L and discrete z, with regression coefficients −0.05 and 0.5, respectively.

(L%, C%)	**$G(t)$**	**Bias**	**var[$\hat{G}_C(t)$]**	$\widehat{\text{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	0.000	0.0010	0.0010	0.940	0.948
	0.50	0.000	0.0016	0.0015	0.952	0.957
	0.75	−0.001	0.0013	0.0013	0.968	0.947
(25, 50)	0.25	−0.001	0.0010	0.0010	0.946	0.952
	0.50	−0.001	0.0015	0.0016	0.953	0.959
	0.75	−0.001	0.0013	0.0013	0.962	0.962
(50, 25)	0.25	−0.002	0.0013	0.0014	0.948	0.956
	0.50	−0.005	0.0028	0.0027	0.943	0.948
	0.75	−0.005	0.0027	0.0026	0.952	0.942
(50, 50)	0.25	−0.003	0.0015	0.0016	0.947	0.959
	0.50	−0.005	0.0034	0.0032	0.945	0.943
	0.75	−0.006	0.0032	0.0031	0.941	0.940

Table 5

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model ncludes three covariates, L, continuous z_1 and discrete z_2, with regression coefficients 0.02, 0.5 and −0.3, respectively.

(L%, C%)	**$G(t)$**	**Bias**	**var[$\hat{G}_C(t)$]**	$\widehat{\text{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	−0.001	0.0011	0.0011	0.936	0.937
	0.50	−0.002	0.0018	0.0018	0.959	0.956
	0.75	−0.003	0.0017	0.0017	0.964	0.955
(25, 50)	0.25	−0.002	0.0011	0.0011	0.949	0.958
	0.50	−0.004	0.0020	0.0020	0.950	0.949
	0.75	−0.003	0.0021	0.0020	0.959	0.959
(50, 25)	0.25	−0.004	0.0018	0.0020	0.965	0.977
	0.50	−0.005	0.0045	0.0046	0.955	0.963
	0.75	−0.005	0.0057	0.0050	0.928	0.941
(50, 50)	0.25	−0.009	0.0025	0.0027	0.955	0.970
	0.50	−0.013	0.0071	0.0064	0.940	0.951
	0.75	−0.009	0.0077	0.0067	0.927	0.944

Table 6

Simulation evaluation of $\hat{G}_C$, $\hat{\sigma}^2_{G,C}$ based on 1000 replicates, together with coverage of 95% confidence intervals. The regressor of the underlying Cox model includes three covariates, L, continuous z_1 and discrete z_2, with regression coefficients −0.05, 0.5 and −0.3, respectively.

(*L*%, *C*%)	***G*(*t*)**	**Bias**	**var[$\hat{G}_C(t)$]**	$\widehat{\mathrm{var}}\,[\hat{G}_C(t)]$	**Linear CI cov.**	**Log-log CI cov.**
(25, 25)	0.25	−0.002	0.0010	0.0010	0.948	0.955
	0.50	−0.002	0.0018	0.0017	0.952	0.953
	0.75	−0.001	0.0017	0.0016	0.965	0.953
(25, 50)	0.25	−0.001	0.0011	0.0011	0.948	0.949
	0.50	−0.000	0.0018	0.0020	0.952	0.954
	0.75	−0.000	0.0016	0.0020	0.952	0.949
(50, 25)	0.25	−0.004	0.0019	0.0020	0.961	0.968
	0.50	−0.005	0.0047	0.0048	0.944	0.962
	0.75	−0.006	0.0052	0.0055	0.929	0.953
(50, 50)	0.25	−0.008	0.0031	0.0029	0.941	0.959
	0.50	−0.011	0.0071	0.0066	0.929	0.940
	0.75	−0.008	0.0081	0.0072	0.917	0.929

Table 7

Estimated hazard ratios for the Cox models based on the BMT sample.

Parameter	Barrett's Study		New analysis	
	Relative risk	P-value	Relative risk	P-value
Transplant time (L^2)	-	-	1.002	0.030
Age >10	1.51	0.003	1.374	0.021
T-cell phenotype	2.16	< 0.001	2.025	< 0.001
Duration of 1st remission ≤ 18 months	2.02	< 0.001	1.504	0.004