WILEY Statistics in Medicine

Analyzing left-truncated and right-censored infectious disease cohort data with interval-censored infection onset

Journal:	*Statistics in Medicine*
Manuscript ID	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Pak, Daewoo; University of Texas MD Anderson Cancer Center, Biostatistics Liu, Jun; University of Texas MD Anderson Cancer Center, Department of Plastic Surgery Ning, Jing; University of Texas MD Anderson Cancer Center, Department of Biostatistics Gomez, Guadalupe; Universitat Politecnica de Catalunya, Estadistica i Investigacio operativa Shen, Yu; University of Texas MD Anderson Cancer Center, Department of Biostatistics
Keywords:	generalized odd rate class of models, incubation period of an infectious disease, interval censoring, left truncation, uncertain initiating event

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Analyzing left-truncated and right-censored infectious disease cohort data with interval-censored infection onset

Daewoo Pak[1] | Jun Liu[2] | Jing Ning[1] | Guadalupe Gómez[3] | Yu Shen*[1]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas
[2]Department of Plastic Surgery, The University of Texas MD Anderson Cancer Center, Houston, Texas
[3]Departament d'Estadística i Investigació Operativa and Barcelona Graduate School of Mathematics BGSMath, Universitat Politècnica de Catalunya, Barcelona, Spain

Correspondence
*Yu Shen, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA Email: yshen@mdanderson.org

Abstract

In an infectious disease cohort study, individuals who have been infected with a pathogen are often recruited for follow up. The period between infection and the onset of symptomatic disease, referred to as the incubation period, is of interest because of its importance on disease surveillance and control. However, the incubation period is often difficult to ascertain due to the uncertainty associated with asymptomatic infection onset time. An additional complication is that the observed infected subjects are likely to have longer incubation periods due to the prevalent sampling. In this paper, we demonstrate how to estimate the distribution of the incubation period with the uncertain infection onset, subject to left-truncation and right-censoring. We employ a family of sufficiently general parametric models, the generalized odds-rate class of regression models, for the underlying incubation period and its correlation with covariates. In simulation studies, we assess the finite sample performance of the model fitting and hazard function estimation. The proposed method is illustrated on data from the HIV/AIDS study on injection drug users admitted to a detoxification program in Badalona, Spain.

KEYWORDS:
generalized odd rate class of models; incubation period of an infectious disease; interval censoring; left truncation; uncertain initiating event

1 | INTRODUCTION

Infectious diseases are caused by a variety of pathogenic microorganisms (e.g., viruses and bacteria) and pose a dangerous threat to human health throughout the world. For instance, a recent outbreak of coronavirus disease 2019 (COVID-19) has spread rapidly to more than 140 countries, causing increasing numbers of associated deaths worldwide.[1] Similarly, acquired immunodeficiency syndrome (AIDS) continues to be a persistent public health issue, causing tens of millions deaths, although the first HIV cases were reported in early 1980.[2] COVID-19 and AIDS are only two examples of infectious diseases affecting societal economic stability , as well as human health. The sudden appearance of a novel infectious disease could trigger a severe pandemic. Understanding the natural history of such infectious diseases will help health care systems to design and implement proper proactive management strategies.

After a patient has been exposed to a new pathogen, pathological changes then occur without any noticeable signs or symptoms during a particular period. This phase of the disease, expanding from infection to onset of symptoms, is referred to as an incubation period. The incubation period varies depending upon the pathogen, which is usually shorter for acute disease (e.g., flu) and longer for chronic disease (e.g., HIV/AIDS), though factors influencing the length of the period are diverse. In clinical practice, knowing the duration and variability of the incubation period is useful in estimating prevalent cases and implementing surveillance systems to manage the development of symptomatic disease.[3,4,5] It can also provide essential information for public health officials to more effectively detect prevalent infection cases within limited resources.[6] While, importantly, the transmission of the infection and the pertinent risk factors of infectious disease prevention have been widely reported, there is a notable and imperative lag in the study of the incubation period of infectious disease, which remains relatively understudied because the incubation period generally cannot be directly observed.

Data on an infectious disease are usually obtained from retrospective studies and those where the date of the infection was only observed within an interval. Donnelly et al[7] reviewed the incubation periods of severe acute respiratory syndrome (SARS) for various populations of patients from sixteen published studies, in which most of them had reported and analyzed interval censored incubation period data due to uncertain infection times. Another example is HIV/AIDS cohort data from the detoxication program for injection drug users (IDUs) in Badalona, Spain. Langohr et al[8] and Gómez et al[9] presumed that the IDUs' unknown dates of HIV infection were between the date of first potential HIV exposure and the date of the first HIV positive test. The analysis of such data becomes more complex when a study recruits only subjects who are infected but disease-free (at risk of progression to symptomatic disease) at the beginning of the program. First, the patients with shorter incubation periods are most likely to be excluded from the cohort, by experiencing the symptomatic disease prior to their first eligibility assessment, thereby causing the left-truncation problem. Secondly, the exact starting date of the infection cannot be ascertained from the data. This may necessitate the use of external data or a reasonable assumption for the time of the infection onset that occurred prior to study recruitment.[10,11] Lastly, the incubation period varies greatly even within a single infectious disease, based on the mechanisms of disease development, infection routes, sources of infection and other underlying factors. A flexible modeling framework is necessary to account for the variability of the incubation period distribution. Failure to address these complexities may result in seriously biased inference on the incubation period and its association with risk factors.

With exact onset dates of infection, extensive literature on left-truncated data can be applied to analyze the incubation period data. Some studies proposed methods using conditional probabilities for the observed truncation times,[12,13,14] whereas others dealt with unconditional approaches.[15,16,17,18,19] Especially for acute infectious diseases, a lognormal model is frequently assumed for the incubation period distribution due to acute course of illness.[20] Although useful, the simple parametric model assumption may be violated in practice, especially for slow progression to the symptomatic disease such as HIV/AIDS.[21] On the other hand, there is limited work on the study of estimation of the incubation period subject to left-truncation and the uncertain infection time. Some studies have handled the absence of information on infection time with external data. For instance, Bacchetti and Jewell[11] obtained information on infection patterns prior to the first recruitment date, fully based on antibody tests on stored sera. This requires that the external information be collected from the population from which the cohorts were drawn. The proper external information may not be available in most studies of infectious disease, because various population types can arise within a social or geographic group as being manifested by many compounding factors, such as health care and co-infection levels. An alternative approach, not relying on external sources, is to assume a parametric form for the truncation variable. Brookmeyer and Goedert[22] parameterized the incubation period and the time of seroconversion which is measured from the first possible date of infection, and allowed for a subject-specific probability density of the infection prior to study entry. Although important, the left-truncation mechanism was not fully considered in Brookmeyer and Goedert's likelihood. Thereafter, Kuo et al[23] adapted the parametric scheme used in Brookmeyer and Goedert[22] to account for left-truncation by grouping the individuals into several categories schematically. Nonetheless, their methods may not always be applicable to a broad range of infectious diseases because of less flexible model assumptions on truncation and incubation periods. Moreover, no existing work has shown rigorous modeling and numerical simulations to support the heuristic arguments.

In this paper, we propose an approach for estimating the distribution of the incubation period using observed left-truncated and right-censored data when the onset of infection is interval-censored. With the consideration of both flexibility and identifiability, we employ a family of sufficiently general parametric models, the generalized odds-rate class of regression models, for the underlying incubation period and its dependence on covariates. Based on this model framework, we derive the distribution of the observable residual incubation period, defined as the time from study enrollment to the manifestation of disease, in a general truncation structure that includes length-biased sampling. We present the likelihood-based estimation procedure in Section 2, followed by a detailed description of the model estimation and associated inference. In Section 3, simulation studies are conducted to evaluate the finite-sample performance of the proposed method in the aspects of the accuracy of the parameter estimators and the coverage rates of confidence intervals for the hazard function. The proposed method is illustrated with the AIDS study with HIV-infected injection drug users in Section 4. We give some concluding remarks in Section 5.

2 | STATISTICAL METHODOLOGIES

2.1 | Notation

Assume T_0 to be the time measured from infection to the manifestation of disease and A_0 to be the time measured from infection to enrollment into the study, within the population of interest. A cohort study is often formed by all recruited individuals for whom $T_0 > A_0$ holds, which leads to left-truncated incubation period data. Let T denote the incubation period of a subject recruited in the cohort study. For the subjects in the cohort study, $T = A + V$, where A is the time from the infection to the study recruitment and V is the time from the study recruitment to the manifestation of disease measured at follow-up. Note that A is dependent of V due to their relationship in T, and both A and T cannot be observed exactly because of the unknown starting date of infection. Nevertheless, the onset of infection can be postulated to be in a defined period $[I_l, I_u)$. As an example of slowly progressing diseases, in the cohort of HIV-infected IDUs whose infections were most likely due to their contaminated intravenous (IV) drug use, the period can be set as the time interval between the date of first IV drug use and the date of first seropositive observation. The upper bound of the interval $[I_l, I_u)$ must be less than or equal to the date of study recruitment E, given the eligibility criteria for inclusion of study, where the equality holds when subjects are followed to monitor for the symptomatic disease immediately after confirming infection. Here, we assume the right half-open interval because it usually takes several hours or weeks after the infection to observe seropositive, i.e., the window period. With the definitions of I_l, I_u and E, a set of all values that A_0 can take is the interval $(l, u] \equiv (E - I_u, E - I_l]$. Denote C as a residual censoring time from the study recruitment to loss to follow-up. See the diagram in Figure 1 visualizing the definition of notations for subjects in the cohort study.

2.2 | Model and likelihood

Consider a random sample of n independent subjects. The observed data for the i-th subject consist of $Y_i = \min(V_i, C_i)$, $\Delta_i = I(V_i \leq C_i)$, $l_i = E_i - I_{ui}$, $u_i = E_i - I_{li}$, and a vector of covariates $\boldsymbol{x}_i$, where $i = 1, \cdots, n$. Under the assumption that A_0 and T_0 are independent given the covariate $\boldsymbol{X} = \boldsymbol{x}$, the truncated joint distribution of (A, V) given $\boldsymbol{x}$ follows,

$$
\begin{aligned}
f_{A,V}(a, v|\boldsymbol{x}) &= P(T_0 = a + v, A_0 = a | T_0 > A_0, \boldsymbol{X} = \boldsymbol{x}) \\
&= \frac{g(a|\boldsymbol{x}) f_{T_0}(a + v|\boldsymbol{x}) I(l < a \leq u)}{\int_0^\infty g(s|\boldsymbol{x}) S_{T_0}(s|\boldsymbol{x}) ds}
\end{aligned}
$$

where f_{T_0} and S_{T_0} are the unbiased density function and the survival function for T_0 given $\boldsymbol{X}$, respectively, and g is the probability density function for A_0 given $\boldsymbol{X}$. Thus, the probability density function of V given $\boldsymbol{X} = \boldsymbol{x}$ is,

$$
f_V(v|\boldsymbol{x}) = \frac{\int_l^u g(a|\boldsymbol{x}) f_{T_0}(a + v|\boldsymbol{x}) da}{\int_0^\infty g(s|\boldsymbol{x}) S_{T_0}(s|\boldsymbol{x}) ds}. \tag{1}
$$

We assume that C is independent of (A_0, V, T_0) given the covariate X. Then, the probability for the i-th subject to have a failure at $Y = y_i$ given $\boldsymbol{x}_i$ is written as,

$$
\begin{aligned}
P\left\{Y \in (y_i, y_i + h), \Delta = 1 | \boldsymbol{X} = \boldsymbol{x}_i\right\} &= P\left\{V \in (y_i, y_i + h), C \geq V | \boldsymbol{X} = \boldsymbol{x}_i\right\} \\
&= f_V(y_i|\boldsymbol{x}_i) h \times S_C(y_i),
\end{aligned} \tag{2}
$$

and, for the subject to be right-censored at $Y = y_i$ given $\boldsymbol{x}_i$, the probability is,

$$
\begin{aligned}
P(Y \in (y_i, y_i + h), \Delta = 0 | \boldsymbol{X} = \boldsymbol{x}_i) &= P\left\{C \in (y_i, y_i + h), V > C | \boldsymbol{X} = \boldsymbol{x}_i\right\} \\
&= \int_{y_i}^\infty f_V(s|\boldsymbol{x}_i) ds \times f_C(y_i) h,
\end{aligned} \tag{3}
$$

where f_C and S_C are the density function and the survival function for the residual censoring time C, respectively. From (2) and (3), the density function of (Y, Δ) for the i-th subject is defined as,

$$
\begin{aligned}
f_{Y,\Delta}(y_i, \delta_i|\boldsymbol{x}_i) &= \lim_{h \to 0^+} \frac{P\{Y \in (y_i, y_i + h), \Delta = \delta_i | \boldsymbol{X} = \boldsymbol{x}_i\}}{h} \\
&= f_V(y_i|\boldsymbol{x}_i)^{\delta_i} S_C(y_i)^{\delta_i} \left\{\int_{y_i}^\infty f_V(s|\boldsymbol{x}_i) ds\right\}^{1-\delta_i} f_C(y_i)^{1-\delta_i}
\end{aligned} \tag{4}
$$

The likelihood function is constructed with (4) after replacing f_V with the formula in (1). The main purpose of constructing the likelihood is to estimate the parameters of interest associated with T_0 via its maximization. Removing the nuisance parameters, f_C and S_C, and considering the observed intervals, (l, u), as being fixed in advance,[24,25] the likelihood function is then proportional to,

$$\begin{aligned} L &= \prod_{i=1}^{n} f_{Y,\Delta}(y_i, \delta_i|\boldsymbol{x}_i) \\ &\propto \prod_{i=1}^{n} f_V(y_i|\boldsymbol{x}_i)^{\delta_i} \left\{ \int_{y_i}^{\infty} f_V(s|\boldsymbol{x}_i)ds \right\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \frac{\left\{ \int_{l_i}^{u_i} g(a|\boldsymbol{x}_i) f_{T_0}(y_i + a|\boldsymbol{x}_i)da \right\}^{\delta_i} \left\{ \int_{l_i}^{u_i} g(a|\boldsymbol{x}_i) S_{T_0}(y_i + a|\boldsymbol{x}_i)da \right\}^{1-\delta_i}}{\int_0^{\infty} g(s|\boldsymbol{x}_i) S_{T_0}(s|\boldsymbol{x}_i)ds}. \end{aligned} \tag{5}$$

To model the association of X and T_0, we employ the generalized odds-rate class of regression models. Within the class of models, we assume that a regression model follows,

$$q(S_{T_0}(t|\boldsymbol{x}_i)) = \phi \log(t/\lambda) + \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{6}$$

where $q(r) = \log(\rho^{-1}(r^{-\rho} - 1))$, $\rho > 0$, $\lambda > 0$, $\phi > 0$, and $\boldsymbol{\beta}$ is a vector of regression parameters for X. The corresponding S_{T_0} and f_{T_0} are, respectively,

$$\begin{aligned} S_{T_0}(t|\boldsymbol{x}_i) &= \left\{ 1 + \rho\, (t/\lambda)^{\phi}\, e^{\boldsymbol{x}_i^T \beta} \right\}^{-1/\rho} \\ f_{T_0}(t|\boldsymbol{x}_i) &= \phi \lambda^{-\phi} t^{\phi - 1} e^{\boldsymbol{x}_i^T \beta} \left\{ 1 + \rho\, (t/\lambda)^{\phi}\, e^{\boldsymbol{x}_i^T \beta} \right\}^{-(1+\rho)/\rho}. \end{aligned}$$

Note that, when $\rho = 1$, the model is equivalent to the (log-logistic) proportional odds model, and, as $\rho \to 0$, it becomes the Weibull family of the proportional hazards model. In the latter case, we use $q(r) = \log(-\log(r))$ at $\rho = 0$. The corresponding hazard function is,

$$h_{T_0}(t|\boldsymbol{x}_i) = \frac{e^{\alpha(t)} \alpha'(t) e^{\boldsymbol{x}_i^T \beta}}{1 + e^{\alpha(t)}},$$

where $\alpha(t) = \phi \log(t/\lambda)$ and $\alpha'(t)$ is the first derivative of $\alpha(t)$ with respect to t.

2.3 | Parameter estimation and inference

We maximize the logarithm of the likelihood (5) to obtain the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ for a parameter vector of interest $\boldsymbol{\theta}$ that consists of the regression model parameters for T_0, i.e., $(\rho, \lambda, \phi, \boldsymbol{\beta})$, and the parameters of $g(a|\boldsymbol{x})$. The integrals over finite or infinite intervals in (5) may not have a closed form. In such case, we use Gauss-Jacobi quadrature with 50 quadrature points after transforming integration over the unit interval. The inference about $\boldsymbol{\theta}$ is performed based on the asymptotic normal distribution of $\widehat{\boldsymbol{\theta}}$, i.e., $N(\widehat{\boldsymbol{\theta}}, I_{obs}^{-1}(\widehat{\boldsymbol{\theta}}))$, where $I_{obs}^{-1}(\widehat{\boldsymbol{\theta}})$ is the observed information matrix for $\boldsymbol{\theta}$ obtained from the summation of the cross-products of the individual score function. In our implementation, the score function is obtained using Richardson's extrapolation that is,

$$(\nabla l(\widehat{\boldsymbol{\theta}}))_j = \frac{-l(\boldsymbol{\theta} + 2h\boldsymbol{e}_j) + 8l(\boldsymbol{\theta} + h\boldsymbol{e}_j) - 8l(\boldsymbol{\theta} - h\boldsymbol{e}_j) + l(\boldsymbol{\theta} - 2h\boldsymbol{e}_j)}{12h} + O(h^4),$$

where $(\nabla l(\widehat{\boldsymbol{\theta}}))_j$ is the j-th element of $\nabla l(\widehat{\boldsymbol{\theta}}) \equiv \partial l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$, $h = 0.0001$, $\boldsymbol{e}_j$ is the j-th unit vector in $\mathbb{R}^p$, and p is the size of the vector $\boldsymbol{\theta}$.

3 | SIMULATIONS

3.1 | Simulation setting

We performed a series of simulation studies to assess the finite sample performance of the likelihood-based inferential approaches for $\boldsymbol{\theta}$. We also compared the performance of the proposed method with several naive approaches using a middle-point imputation for A and/or ignoring left-truncation. In data generation for left-truncated and right-censored data, two scenarios were considered, in which distributions for A_0 are different: (i) a uniform distribution $g_1(a) = 1/\tau$ over $(0, \tau)$, where τ is a predetermined constant, and (ii) a covariate-specific Weibull distribution $g_2(a|x_1) = (\gamma/\eta)(a/\eta)^{\gamma-1}e^{\alpha x_1}e^{-(a/\eta)^{\gamma}e^{\alpha x_1}}$. The former scenario is known as length-biased sampling and τ is chosen large enough to make probability mass $(T_0 > \tau)$ negligible. Specifically, with $g_1(a)$ and $\tau \to \infty$, the denominator of the logarithm of the likelihood of (5) becomes the conditional expectation of T_0 given covariates $\boldsymbol{x}$, i.e., $E(T_0|\boldsymbol{x})$, which is $(\lambda^{\phi}\rho e^{\boldsymbol{x}^T\boldsymbol{\beta}})^{1/\phi}B(1 + 1/\phi, 1/\rho - 1/\phi)$ for $\phi > \rho$, where B is a beta function. In our simulation, we set $\tau = 200$ in $g_1(a)$, and $\eta = 6$, $\gamma = 1.2$ and $\alpha = 0.5$ in $g_2(a|x_1)$. Two independent covariates were generated from $x_1 \sim$ Bernoulli(0.5) and $x_2 \sim$ Uniform(0, 1), i.e., $\boldsymbol{x} = (x_1, x_2)^T$. A vector of parameters in the regression model were set to $(\lambda, \phi, \rho, \beta_1, \beta_2) = (5, 6, 2, 0.5, 0.5)$, where β_1 and β_2 are coefficients for x_1 and x_2, respectively.

To obtain the left-truncated and right-censored data, we first generate a pair of samples of (A_0, T_0) from $A_0 \sim g(a)$ or $g(a|x_1)$ and $T_0 \sim f_{T_0}(t|\boldsymbol{x})$, and only the pairs with $A_0 < T_0$ are taken as samples. The residual censoring time C is generated from uniform distributions on $(0, d)$, where d is set to achieve desired censoring proportions of each simulation scenario. For each pair of samples, an observable pair of (Y, Δ) is obtained by $V \wedge C$ and $I(V \leq C)$, respectively, where $V = T_0 - A_0$. The planned detection times for the initiating event (i.e., infection) is set to be an integer in $[0, Y_i]$ for the i-th subject, but the subject is allowed to miss each planned detection time with a 20% chance, so that the actual number of detection times could vary across subjects. Thus, A_i is bracketed in the interval $(l_i, u_i]$ of which boundaries are two actual detection times near it.

In Monte Carlo simulations, two sample sizes ($n = 400$ and $n = 800$) and two censoring percentages (15% and 30%) were used, and one thousand replicates were generated in each simulation.

3.2 | Simulation results

Table 1 and Table 2 show the simulation results for estimating $\boldsymbol{\theta}$ under two different truncation densities, $g_1(a)$ and $g_2(a|x_1)$, respectively. The right-censoring rates for both tables were 30%. We used the logarithmic transformation for the positive parameters when maximizing the log-likelihood. The proposed method (*Proposed*) was compared to the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The approaches adjusting for left-truncation, *Proposed* and *MP*, have different sets of parameters to estimate according to the density for A_0, which is $\boldsymbol{\theta} = (\log(\rho), \log(\lambda), \log(\phi), \beta_1, \beta_2)^T$ for $g_1(a)$ and $\boldsymbol{\theta} = (\log(\gamma), \log(\eta), \alpha, \log(\rho), \log(\lambda), \log(\phi), \beta_1, \beta_2)^T$ for $g_2(a|x_{1i})$. The two approaches ignoring left-truncation, *ICNT* and *MPNT*, were fitted after reformulating the simulated data to interval-censored data with $T \in (V + l, V + u)$ and right-censored data with $T = V + (l + u)/2$, respectively.

Given the four methods, only the proposed method has the unbiased estimators and good empirical coverage probabilities (CPs) of the 95% confidence intervals for all of the tried scenarios. The empirical standard error of every estimator is close to the mean of the estimated asymptotic standard error, and when the sample size doubled, they decreased by $\sqrt{2}$, as expected given the asymptotic theory. In the naive approaches, by contrast, the estimates could be seriously biased. Specifically, the approaches using the middle-point imputation (*MP* and *MPNT*) could lead to more severe biases for the model parameters (λ, ϕ and ρ) than the approach handling the interval-censored of T (*ICNT*), although *MP* showed relatively better performance on the CPs for those parameters than the other naive approaches. Whereas, the covariate coefficients (β_1, β_2) for the naive approaches tend to be biased and have relatively lower CPs, especially when the density for A_0 was associated with the covariate. The simulations for the 15% censoring rate were also performed with two truncation scenarios, but they are relegated to Appendix A.1 as they show very similar performance with that of Table 1 and Table 2.

Figure 2 shows the average estimates and the empirical coverage probabilities of the 95% pointwise confidence intervals of the hazard functions, calculated at $x_1 = 1$ and $x_2 = 0.5$, under two types of the truncation densities, when $n = 400$ with 30% censoring rate. The 95% pointwise confidence intervals were constructed using the asymptotic normal distribution of the log-hazard function derived from the delta method. The average hazard function estimates for the proposed method are closest to the true hazard function, and its empirical coverage probabilities outperform others, as the probabilities being around the nominal level for every time t. The *MP* estimates for the hazard function is close to the truth; however, the empirical coverage probabilities of *MP* deviate from the truth in the early time period. In contrast, the approach ignoring left-truncation (*MPNT* and *ICNT*) lead to serious biased results in the hazard function estimation, as their estimates deviating from the truth with low empirical coverage probabilities overall t. Thus, adjusting left-truncation is essential for the precise estimation of the hazard function of T, compared to using middle-point imputation. Nevertheless, using middle-point imputation would lead to serious biased estimates of left-truncation distribution (see the biases for η and γ of *MP* in Table 2).

The mean of integrated squared error $\int_0^\infty \{\hat{h}(t|x_1, x_2) - h(t|x_1, x_2)\}^2 dt$, referred to as MISE, is calculated at $x_1 = 1$ and $x_2 = 0.5$ for each approach. The proposed method has the lowest value in MISE, and it decreases with the sample size n. All of the figures corresponding to the simulation results with the 15% censoring rate and with $n = 800$ can be found in Appendix A.2. The estimators showed similar trends with the ones presented here, except for worse empirical coverage probabilities when the sample sizes doubled.

4 | APPLICATION

We illustrated the proposed method using the HIV/AIDS cohort study from the detoxification program of 361 injection drug users who were admitted to the detoxification unit of the Hospital Universitari Germans Trias i Pujol in Badalona from 1987 to 2000.[26] We also compared the performance of the proposed method with three naive approaches. In the cohort, a total of 266 individuals were found to be infected with HIV in screening tests during the program. Among them, 10 individuals already developed AIDS, and 256 were HIV-positive but AIDS-free. Thus, 256 individuals were recruited to the study and followed for AIDS development. Of these subjects, only 219 were considered in the analysis, as the other 39 had missing essential information. During follow-up, 84 subjects were diagnosed with AIDS up to the year 2000. In the study, the time interval for possible HIV infections of the subjects was recorded from dates of HIV exposure by first injection drug use until dates of the screening test results showing seropositive observation. For this reason, the target population within the study was in fact HIV-seropositive IDUs (AIDS-free). We assume that the results of this study can be generalized to the population of HIV-infected IDUs who are AIDS-free. The dates of study recruitment are set to the dates of the first seropositive observation during program participant follow-up.

We analyzed the above left-truncated and right-censored HIV/AIDS data using the generalized odd-rate class of regression models in (6). In the regression model, two variables, age at the enrollment and gender, are considered as the potential factors being associated with the incubation period of AIDS. For the approaches adjusting for left-truncation (*Proposed* and *MP*), the period from HIV infection to the study enrollment (a left-truncation variable) were assumed to follow a covariate-specific Weibull distribution with gender as the covariate.

Neither age nor gender have shown statistically significant association with the incubation period of HIV/AIDS. The likelihood ratio test also does not reject the differences between the model with the covariates and the model without the covariates (p-value $= 0.068$). On the other hand, the estimated effect of age on the incubation period tends to be overestimated in the naive approaches compared with the proposed method, and gender is not statistically significant associated with the incubation period in all models. The detailed results can be found in Appendix A.3. Figure 3 shows the estimated survival functions for the approaches and the 95% pointwise confidence interval for the proposed method. All of the naive approaches tend to over-estimate the survival probabilities. This over-estimation is particularly serious when ignoring left-truncation. The retrospectively identified intervals for HIV infection are usually quite large, thus using the middle-point imputation for the interval-censored HIV infection could yield bias in survival estimation in real-life applications. The estimated median incubation period of the proposed method was 9.19 years ($95\%\text{CI} : 7.25 - 11.13$), which is one or two years less than those of the naive approaches (medial incubation period = 10.80 for *MP*, 11.85 for *ICNT*, and 12.03 for *MPNT*).

5 | DISCUSSION

We have proposed a likelihood-based approach to estimate the distribution of the incubation period and evaluate the related covariate effects under left-truncated and right-censored infectious disease data, when the time of infection is uncertain. The proposed method is designed to reflect the data-generating mechanism from a cohort itself by incorporating flexible modeling for both left-truncation and the uncertain incubation period, which can be easily modified with other classes of the parametric family for various purposes. It is also shown to be feasible with general truncation densities or under the stationary assumption, which assumes a stationary Poisson process on the onset of infection.

The incubation periods tend to be right-skewed for acute infectious diseases,[27,28] but because of the large variability, the choice of a certain distribution for the incubation period was often challenging. Several parametric models, such as Weibull, log-normal, and log-logistic distributions, were often compared to identify the best-fit models for the incubation period of infectious disease.[29,30] The validity and critique of the log-normal assumption was also highlighted.[3] To circumvent this issue, we used a flexible class of parametric regression models, referred to as the generalized odds-rate class, on the likelihood to model regression estimators. The interpretation of the covariate effects on survival time is straightforward (Dabrowska and Doksum[31] and Scharfstein et al[32]).

We illustrated the proposed method using the study of HIV-infected IDUs from the detoxication program in Badalona, Spain. In this specific cohort, we do not see gender and age differences in the incubation period of AIDS. Whereas, Hera et al[33], which also studied the incubation period for IDUs, made different conclusions: female IDUs had lower HIV progression to AIDS, but the age at seroconversion did not affect the progression. In other geographic settings with different health care and co-infection levels, other ways of transmission may exist, possibly leading to different conclusions. In the area with high herpes simplex virus infection, for instance, women could experience a faster progression to AIDS, compared to men. Therefore, identifying the data generating mechanism correctly is vital when drawing statistical inferences from the cohort itself.

In the early 1980's, HIV/AIDS was an acute fatal disease; however, it has become a chronic, manageable condition due to the use of medicines to treat HIV infection such as antiretroviral therapy. Because the patients in the data analysis were adolescent and young adults at the study enrollment and only 3.9% of them died before the onset of AIDS, we did not consider the problem of competing risk due to other causes. This competing risk situation could be adopted within our modeling framework, but requires a more complicated likelihood function.

The proposed method, which enjoys model flexibility, is readily applicable to other studies. One example is with studies on emerging infectious diseases, such as COVID-19, for which the incubation period has not been well determined due to the challenge of recalling the exact date of infection. The method can also be applicable to (prevalent) cohort studies, such as a dementia study in which subjects assess the onset date of disease retrospectively and are followed for the occurrence of a subsequent event, such as death.

ACKNOWLEDGEMENTS

This work was supported in part by grants CA016672 and CA193878 from the National Cancer Institute (NCI); by grants MTM2015-64465-C2-1-R and MDM-2014-0445 (MINECO/FEDER) from the Ministerio de Economía y Competitividad (Spain); and by grant 2017 SGR 622 (GRBIO) from the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain).

REFERENCES

1. World Health Organization . Coronavirus disease 2019 (COVID-19): situation report, 45. 2020.
2. UNAIDS . Global AIDS update 2019-communities at the centre. *Geneva, Switzerland: UNAIDS* 2019.
3. Nishiura H. Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerging themes in epidemiology* 2007; 4(1): 2.
4. Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE, Cummings DA. Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet infectious diseases* 2009; 9(5): 291–300.
5. Chan M, Johansson MA. The incubation periods of dengue viruses. *PloS one* 2012; 7(11).
6. Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine* 2020.
7. Donnelly CA, Fisher MC, Fraser C, et al. Epidemiological and genetic analysis of severe acute respiratory syndrome. *The Lancet infectious diseases* 2004; 4(11): 672–683.
8. Gómez G, Luz Calle M, Egea JM, Muga R. Risk of HIV infection as a function of the duration of intravenous drug use: a non-parametric Bayesian approach. *Statistics in Medicine* 2000; 19(19): 2641-2656.
9. Langohr K, Gómez G, Muga R. A parametric survival model with an interval-censored covariate. *Statistics in Medicine* 2004; 23(20): 3159-3175.

10. Jewell NP. Some statistical issues in studies of the epidemiology of AIDS. *Statistics in medicine* 1990; 9(12): 1387–1416.

11. Bacchetti P, Jewell NP. Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics* 1991: 947–960.

12. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)* 1976; 38(3): 290–295.

13. Lai TL, Ying Z. Estimating a distribution function with truncated and censored data. *The Annals of Statistics* 1991: 417–442.

14. Wang MC. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* 1991; 86(413): 130–143.

15. Vardi Y. Nonparametric Estimation in the Presence of Length Bias. *The Annals of Statistics* 1982; 10(2): 616–620.

16. Gill RD, Vardi Y, Wellner JA. Large Sample Theory of Empirical Distributions in Biased Sampling Models. *The Annals of Statistics* 1988; 16(3): 1069–1112.

17. Asgharian M, Wolfson DB, others . Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics* 2005; 33(5): 2109–2131.

18. Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association* 2009; 104(487): 1192–1202.

19. Liu H, Ning J, Qin J, Shen Y. Semiparametric maximum likelihood inference for truncated or biased-sampling data. *Statistica Sinica* 2016: 1087–1115.

20. Brookmeyer R, Blades N, Hugh-Jones M, Henderson DA. The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak. *Biostatistics* 2001; 2(2): 233–247.

21. Nishiura H, Eichner M. Infectiousness of smallpox relative to disease age: estimates based on transmission network and incubation period. *Epidemiology & Infection* 2007; 135(7): 1145–1150.

22. Brookmeyer R, Goedert JJ. Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics* 1989: 325–335.

23. Kuo J, Taylor J, Detels R. Estimating the AIDS incubation period from a prevalent cohort. *American journal of epidemiology* 1991; 133(10): 1050–1057.

24. Oller R, Gómez G, Calle ML. Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics* 2004; 32(3): 315–326.

25. Oller R, Gómez G, Calle ML. Interval censoring: identifiability and the constant-sum property. *Biometrika* 2007; 94(1): 61–70.

26. Langohr K, Gómez G, Muga R. A parametric survival model with an interval-censored covariate. *Statistics in medicine* 2004; 23(20): 3159–3175.

27. Sartwell PE, others . The Distribution of Incubation Periods of Infectious Diseases.. *American Journal of Hygiene* 1950; 51(3): 310–18.

28. Virlogeux V, Li M, Tsang TK, et al. Estimating the distribution of the incubation periods of human avian influenza A (H7N9) virus infections. *American journal of epidemiology* 2015; 182(8): 723–729.

29. Munoz A, Xu J. Models for the incubation of AIDS and variations according to age and period. *Statistics in medicine* 1996; 15(22): 2459–2473.

30. Reich NG, Lessler J, Cummings DAT, Brookmeyer R. Estimating incubation period distributions with coarse data. *Statistics in Medicine* 2009; 28(22): 2769-2784.

31. Dabrowska DM, Doksum KA. Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association* 1988; 83(403): 744–749.

32. Scharfstein DO, Tsiatis AA, Gilbert PB. Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime data analysis* 1998; 4(4): 355–391.

33. Hera d. lMG, Ferreros I, Amo dJ, et al. Gender differences in progression to AIDS and death from HIV seroconversion in a cohort of injecting dug users from 1986 to 2001. *Journal of Epidemiology & Community Health* 2004; 58(11): 944–950.

TABLE 1 Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_1(a)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 30%.

	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.609	1.792	0.693	0.500	0.500	1.609	1.792	0.693	0.500	0.500
($n = 400$ & $cen = 30\%$)										
			Proposed					*ICNT*		
Bias	0.010	0.011	0.017	0.008	0.007	0.064	0.187	0.822	0.107	0.124
ESE	0.039	0.143	0.181	0.290	0.156	0.048	0.138	0.199	0.365	0.193
SE	0.040	0.145	0.184	0.295	0.162	0.052	0.146	0.200	0.371	0.204
CP	0.948	0.952	0.949	0.958	0.974	0.767	0.819	0.013	0.969	0.959
			MP					*MPNT*		
Bias	0.004	0.103	0.125	0.056	0.042	0.089	0.065	0.682	0.040	0.056
ESE	0.037	0.122	0.161	0.260	0.137	0.046	0.114	0.178	0.320	0.167
SE	0.040	0.123	0.163	0.266	0.143	0.050	0.120	0.176	0.329	0.176
CP	0.958	0.860	0.901	0.947	0.946	0.575	0.938	0.042	0.968	0.969
($n = 800$ & $cen = 30\%$)										
			Proposed					*ICNT*		
Bias	0.010	0.023	0.025	0.005	0.002	0.066	0.171	0.811	0.110	0.114
ESE	0.029	0.093	0.119	0.209	0.110	0.035	0.090	0.133	0.260	0.136
SE	0.028	0.100	0.127	0.205	0.112	0.036	0.099	0.137	0.256	0.139
CP	0.943	0.957	0.956	0.950	0.949	0.555	0.618	0.000	0.936	0.908
			MP					*MPNT*		
Bias	0.005	0.113	0.131	0.051	0.048	0.091	0.052	0.675	0.045	0.049
ESE	0.028	0.080	0.106	0.187	0.097	0.034	0.075	0.120	0.230	0.118
SE	0.028	0.085	0.113	0.186	0.099	0.035	0.082	0.122	0.229	0.122
CP	0.945	0.730	0.795	0.936	0.923	0.249	0.935	0.000	0.952	0.945

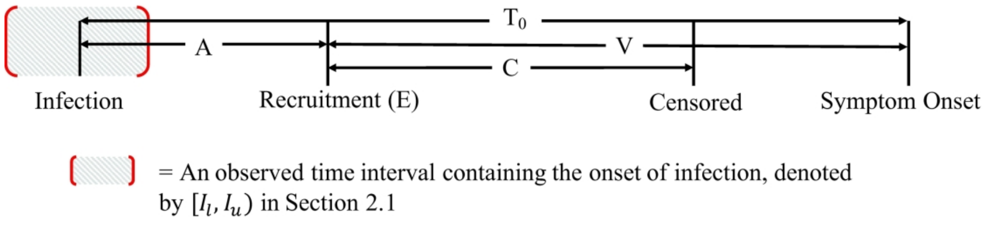
Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

TABLE 2 Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_2(a|x_1)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 30%.

	$\log(\eta)$	$\log(\gamma)$	α	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.792	0.182	0.500	1.609	1.792	0.693	0.500	0.500	1.609	1.792	0.693	0.500	0.500
($n = 400$ & $cen = 30\%$)													
	Proposed								*ICNT*				
Bias	0.012	0.003	0.014	0.004	0.021	0.023	0.002	0.008	0.085	0.067	0.214	0.207	0.012
ESE	0.213	0.069	0.260	0.048	0.149	0.237	0.307	0.163	0.048	0.125	0.216	0.293	0.159
SE	0.213	0.067	0.259	0.047	0.146	0.236	0.308	0.159	0.045	0.125	0.217	0.292	0.154
CP	0.922	0.935	0.955	0.941	0.941	0.960	0.958	0.945	0.523	0.944	0.830	0.913	0.935
	MP								*MPNT*				
Bias	0.120	0.099	0.003	0.010	0.134	0.144	0.081	0.039	0.106	0.040	0.077	0.139	0.039
ESE	0.157	0.045	0.242	0.047	0.130	0.219	0.278	0.142	0.045	0.107	0.201	0.261	0.139
SE	0.143	0.060	0.212	0.048	0.126	0.222	0.282	0.142	0.044	0.106	0.202	0.261	0.136
CP	0.754	0.646	0.915	0.942	0.778	0.915	0.942	0.928	0.332	0.925	0.929	0.934	0.926
($n = 800$ & $cen = 30\%$)													
	Proposed								*ICNT*				
Bias	0.020	0.001	0.005	0.008	0.024	0.018	0.018	0.004	0.082	0.063	0.218	0.187	0.008
ESE	0.137	0.045	0.176	0.033	0.103	0.163	0.215	0.113	0.034	0.087	0.151	0.205	0.110
SE	0.142	0.047	0.177	0.033	0.101	0.162	0.214	0.110	0.032	0.087	0.150	0.203	0.107
CP	0.931	0.948	0.955	0.942	0.931	0.948	0.946	0.943	0.291	0.899	0.693	0.871	0.943
	MP								*MPNT*				
Bias	0.120	0.097	0.008	0.007	0.136	0.137	0.098	0.041	0.102	0.043	0.082	0.122	0.041
ESE	0.106	0.029	0.166	0.033	0.090	0.152	0.196	0.100	0.033	0.074	0.141	0.184	0.097
SE	0.098	0.041	0.146	0.033	0.087	0.153	0.197	0.098	0.031	0.074	0.140	0.183	0.095
CP	0.692	0.307	0.921	0.948	0.630	0.874	0.927	0.910	0.098	0.898	0.889	0.911	0.911

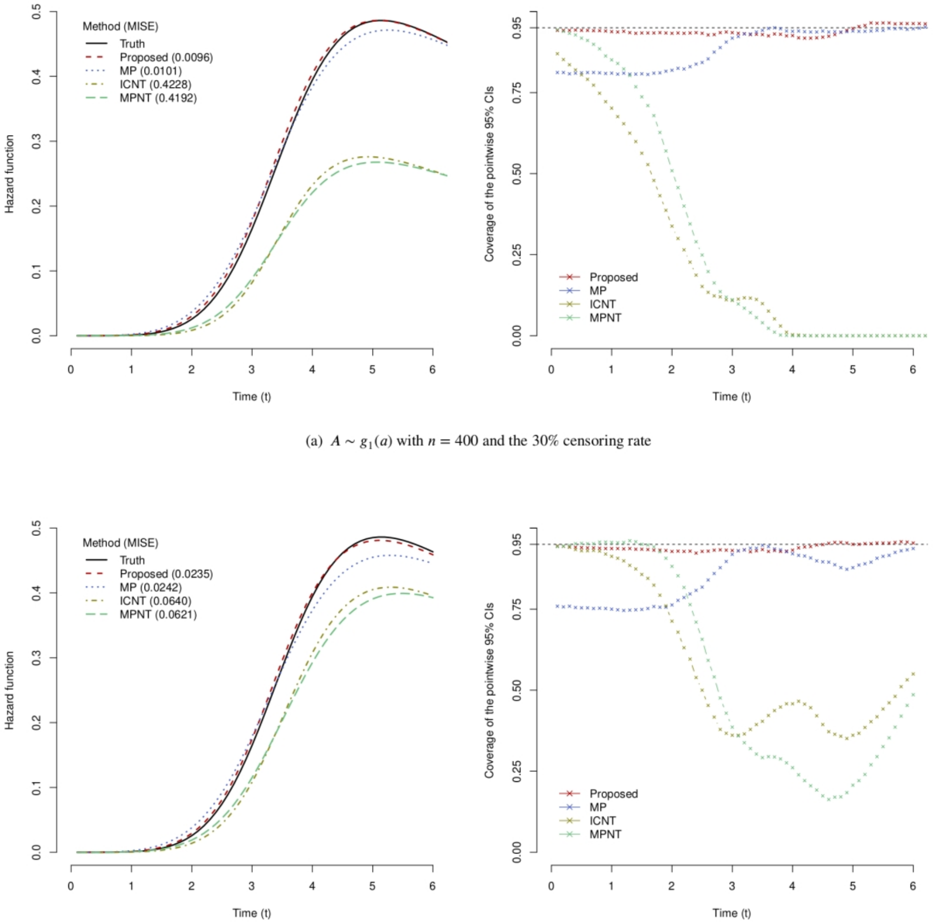
Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

□


T_0
A
V
C
Infection
Recruitment (E)
Censored
Symptom Onset
▨ = An observed time interval containing the onset of infection, denoted by $[I_l, I_u)$ in Section 2.1

The diagram describing the notations

257x61mm (300 x 300 DPI)

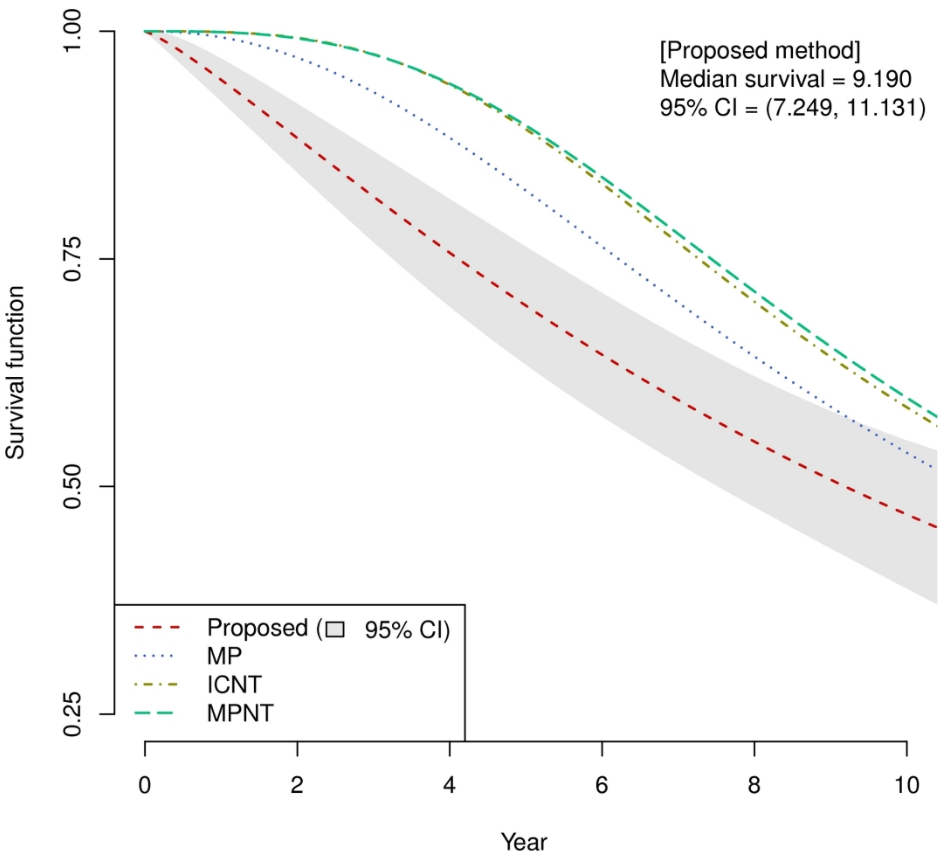


(a) $A \sim g_1(a)$ with $n = 400$ and the 30% censoring rate

(b) $A \sim g_2(a|x_1)$ with $n = 400$ and the 30% censoring rate

The average estimates and the empirical coverage probability of the 95% pointwise confidence intervals of the hazard functions at $x_1 = 1$ and $x_2 = 0.5$

168x171mm (300 x 300 DPI)



Estimates of the survival function with their 95% credible intervals

177x177mm (300 x 300 DPI)

Web-based Supplementary Material: Analyzing Left-Truncated and Right-Censored Infectious Disease Cohort Data with Interval-Censored Infection Onset

Daewoo Pak, Jun Liu, Jing Ning, Guadalupe Gómez, Yu Shen

In this web-based supplementary material, the simulation results for other important scenarios are provided by comparing the proposed method (*Proposed*) to three naive approaches: the middle-point imputation approach adjusting for left-truncation (MP), the approach considering interval censoring but ignoring left-truncation ($ICNT$), and the middle-point imputation approach ignoring left-truncation ($MPNT$). The results illustrate the problems when fail to explore the data generating mechanisms fully. Two censoring rates (15% and 30%) and two sample sizes ($n = 400$ and $n = 800$) are considered in the simulation. The data analysis results are also provided.

A1 The simulation results of model estimation

Table A1: Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_1(a)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 15%.

	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.609	1.792	0.593	0.500	0.500	1.609	1.792	0.593	0.500	0.500
($n = 400$ & $cen = 15\%$)										
			Proposed					*ICNT*		
Bias	0.009	0.016	0.02	0.003	0.004	0.077	0.142	0.66	0.082	0.089
ESE	0.039	0.133	0.169	0.283	0.162	0.048	0.126	0.185	0.331	0.188
SE	0.039	0.139	0.176	0.286	0.157	0.047	0.133	0.184	0.334	0.183
CP	0.946	0.957	0.960	0.957	0.949	0.610	0.877	0.047	0.955	0.948
			MP					*MPNT*		
Bias	0.005	0.107	0.127	0.05	0.044	0.1	0.029	0.529	0.022	0.029
ESE	0.038	0.113	0.151	0.255	0.141	0.046	0.105	0.167	0.296	0.163
SE	0.038	0.118	0.156	0.258	0.139	0.045	0.111	0.164	0.299	0.160
CP	0.953	0.841	0.879	0.943	0.923	0.404	0.954	0.100	0.955	0.955
($n = 800$ & $cen = 15\%$)										
			Proposed					*ICNT*		
Bias	0.007	0.025	0.029	0.005	0.001	0.080	0.130	0.647	0.089	0.082
ESE	0.027	0.097	0.121	0.199	0.109	0.033	0.091	0.129	0.231	0.127
SE	0.027	0.096	0.121	0.198	0.108	0.033	0.091	0.127	0.231	0.125
CP	0.941	0.936	0.944	0.951	0.944	0.307	0.738	0.000	0.941	0.921
			MP					*MPNT*		
Bias	0.007	0.113	0.132	0.041	0.047	0.103	0.022	0.521	0.032	0.025
ESE	0.026	0.083	0.108	0.179	0.097	0.032	0.077	0.117	0.207	0.112
SE	0.027	0.082	0.108	0.180	0.096	0.032	0.077	0.114	0.209	0.111
CP	0.939	0.695	0.757	0.940	0.903	0.096	0.952	0.003	0.951	0.942

Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

Table A2: Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_1(a)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 30%.

	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.609	1.792	0.693	0.500	0.500	1.609	1.792	0.693	0.500	0.500
($n = 400$ & $cen = 30\%$)										
			Proposed					*ICNT*		
Bias	0.010	0.011	0.017	0.008	0.007	0.064	0.187	0.822	0.107	0.124
ESE	0.039	0.143	0.181	0.290	0.156	0.048	0.138	0.199	0.365	0.193
SE	0.040	0.145	0.184	0.295	0.162	0.052	0.146	0.200	0.371	0.204
CP	0.948	0.952	0.949	0.958	0.974	0.767	0.819	0.013	0.969	0.959
			MP					*MPNT*		
Bias	0.004	0.103	0.125	0.056	0.042	0.089	0.065	0.682	0.040	0.056
ESE	0.037	0.122	0.161	0.260	0.137	0.046	0.114	0.178	0.320	0.167
SE	0.040	0.123	0.163	0.266	0.143	0.050	0.120	0.176	0.329	0.176
CP	0.958	0.860	0.901	0.947	0.946	0.575	0.938	0.042	0.968	0.969
($n = 800$ & $cen = 30\%$)										
			Proposed					*ICNT*		
Bias	0.010	0.023	0.025	0.005	0.002	0.066	0.171	0.811	0.110	0.114
ESE	0.029	0.093	0.119	0.209	0.110	0.035	0.090	0.133	0.260	0.136
SE	0.028	0.100	0.127	0.205	0.112	0.036	0.099	0.137	0.256	0.139
CP	0.943	0.957	0.956	0.950	0.949	0.555	0.618	0.000	0.936	0.908
			MP					*MPNT*		
Bias	0.005	0.113	0.131	0.051	0.048	0.091	0.052	0.675	0.045	0.049
ESE	0.028	0.080	0.106	0.187	0.097	0.034	0.075	0.120	0.230	0.118
SE	0.028	0.085	0.113	0.186	0.099	0.035	0.082	0.122	0.229	0.122
CP	0.945	0.730	0.795	0.936	0.923	0.249	0.935	0.000	0.952	0.945

Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

Table A3: Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_2(a|x_1)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 15%.

	$\log(\eta)$	$\log(\gamma)$	α	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.792	0.182	0.500	1.609	1.792	0.693	0.500	0.500	1.609	1.792	0.693	0.500	0.500
($n = 400$ *&* $cen = 15\%$)													
				Proposed							*ICNT*		
Bias	0.020	0.003	0.005	0.004	0.021	0.026	0.003	0.002	0.083	0.055	0.154	0.182	0.003
ESE	0.202	0.065	0.252	0.045	0.140	0.216	0.291	0.156	0.045	0.117	0.195	0.271	0.150
SE	0.203	0.067	0.254	0.044	0.135	0.208	0.291	0.151	0.041	0.116	0.192	0.272	0.144
CP	0.923	0.954	0.960	0.931	0.939	0.944	0.954	0.941	0.475	0.943	0.871	0.916	0.942
				MP							*MPNT*		
Bias	0.123	0.101	0.008	0.011	0.134	0.151	0.082	0.046	0.102	0.048	0.024	0.118	0.052
ESE	0.154	0.043	0.236	0.045	0.122	0.200	0.265	0.137	0.043	0.100	0.181	0.245	0.132
SE	0.138	0.059	0.209	0.044	0.116	0.194	0.267	0.134	0.040	0.098	0.177	0.245	0.128
CP	0.740	0.628	0.921	0.937	0.759	0.884	0.932	0.916	0.282	0.905	0.937	0.926	0.908
($n = 800$ *&* $cen = 15\%$)													
				Proposed							*ICNT*		
Bias	0.026	0.006	0.019	0.004	0.033	0.033	0.005	0.003	0.084	0.045	0.148	0.166	0.009
ESE	0.132	0.046	0.195	0.031	0.094	0.144	0.104	0.170	0.031	0.080	0.133	0.186	0.101
SE	0.137	0.046	0.202	0.031	0.093	0.144	0.104	0.174	0.029	0.080	0.134	0.189	0.099
CP	0.921	0.945	0.959	0.941	0.926	0.947	0.951	0.952	0.185	0.923	0.797	0.873	0.948
				MP							*MPNT*		
Bias	0.121	0.101	0.096	0.010	0.143	0.156	0.050	0.008	0.103	0.055	0.020	0.105	0.055
ESE	0.105	0.030	0.177	0.031	0.082	0.135	0.092	0.162	0.030	0.068	0.123	0.168	0.090
SE	0.095	0.041	0.186	0.031	0.081	0.135	0.093	0.145	0.028	0.068	0.124	0.171	0.089
CP	0.696	0.267	0.931	0.927	0.558	0.802	0.898	0.913	0.053	0.873	0.947	0.917	0.890

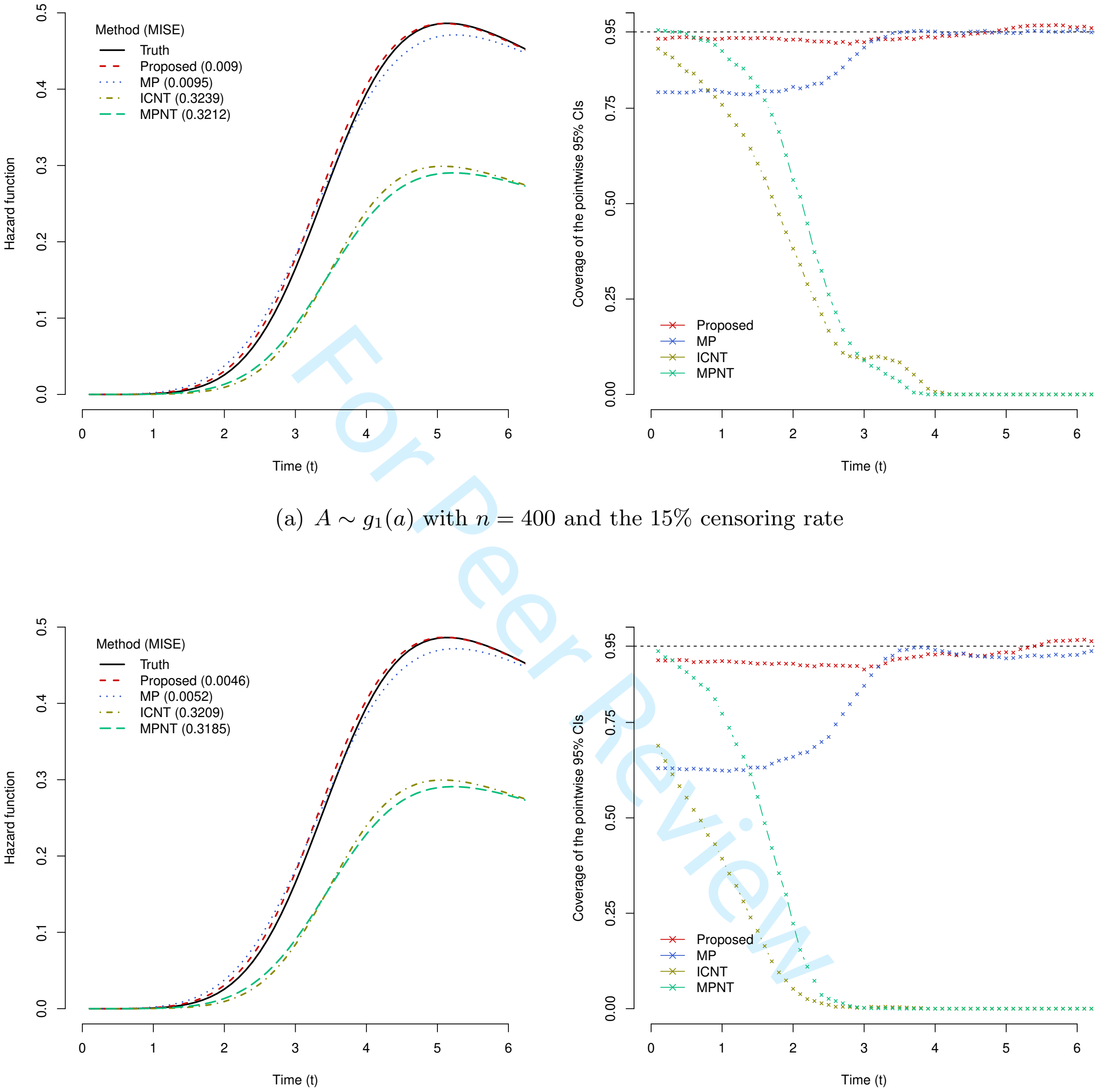
Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

Table A4: Simulation results for estimating $\boldsymbol{\theta}$ with $A_0 \sim g_2(a|x_1)$ from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). The censoring rate is 30%.

	$\log(\eta)$	$\log(\gamma)$	α	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	β_1	β_2
Truth	1.792	0.182	0.500	1.609	1.792	0.693	0.500	0.500	1.609	1.792	0.693	0.500	0.500
($n = 400$ *& cen* $= 30\%$)													
				Proposed							*ICNT*		
Bias	0.012	0.003	0.014	0.004	0.021	0.023	0.002	0.008	0.085	0.067	0.214	0.207	0.012
ESE	0.213	0.069	0.260	0.048	0.149	0.237	0.307	0.163	0.048	0.125	0.216	0.293	0.159
SE	0.213	0.067	0.259	0.047	0.146	0.236	0.308	0.159	0.045	0.125	0.217	0.292	0.154
CP	0.922	0.935	0.955	0.941	0.941	0.960	0.958	0.945	0.523	0.944	0.830	0.913	0.935
				MP							*MPNT*		
Bias	0.120	0.099	0.003	0.010	0.134	0.144	0.081	0.039	0.106	0.040	0.077	0.139	0.039
ESE	0.157	0.045	0.242	0.047	0.130	0.219	0.278	0.142	0.045	0.107	0.201	0.261	0.139
SE	0.143	0.060	0.212	0.048	0.126	0.222	0.282	0.142	0.044	0.106	0.202	0.261	0.136
CP	0.754	0.646	0.915	0.942	0.778	0.915	0.942	0.928	0.332	0.925	0.929	0.934	0.926
($n = 800$ *& cen* $= 30\%$)													
				Proposed							*ICNT*		
Bias	0.020	0.001	0.005	0.008	0.024	0.018	0.018	0.004	0.082	0.063	0.218	0.187	0.008
ESE	0.137	0.045	0.176	0.033	0.103	0.163	0.215	0.113	0.034	0.087	0.151	0.205	0.110
SE	0.142	0.047	0.177	0.033	0.101	0.162	0.214	0.110	0.032	0.087	0.150	0.203	0.107
CP	0.931	0.948	0.955	0.942	0.931	0.948	0.946	0.943	0.291	0.899	0.693	0.871	0.943
				MP							*MPNT*		
Bias	0.120	0.097	0.008	0.007	0.136	0.137	0.098	0.041	0.102	0.043	0.082	0.122	0.041
ESE	0.106	0.029	0.166	0.033	0.090	0.152	0.196	0.100	0.033	0.074	0.141	0.184	0.097
SE	0.098	0.041	0.146	0.033	0.087	0.153	0.197	0.098	0.031	0.074	0.140	0.183	0.095
CP	0.692	0.307	0.921	0.948	0.630	0.874	0.927	0.910	0.098	0.898	0.889	0.911	0.911

Bias, empirical bias; ESE, empirical standard error of the parameter estimator; SE, average of the standard error estimator; CP, coverage of the 95% confidence interval with the Normal approximation.

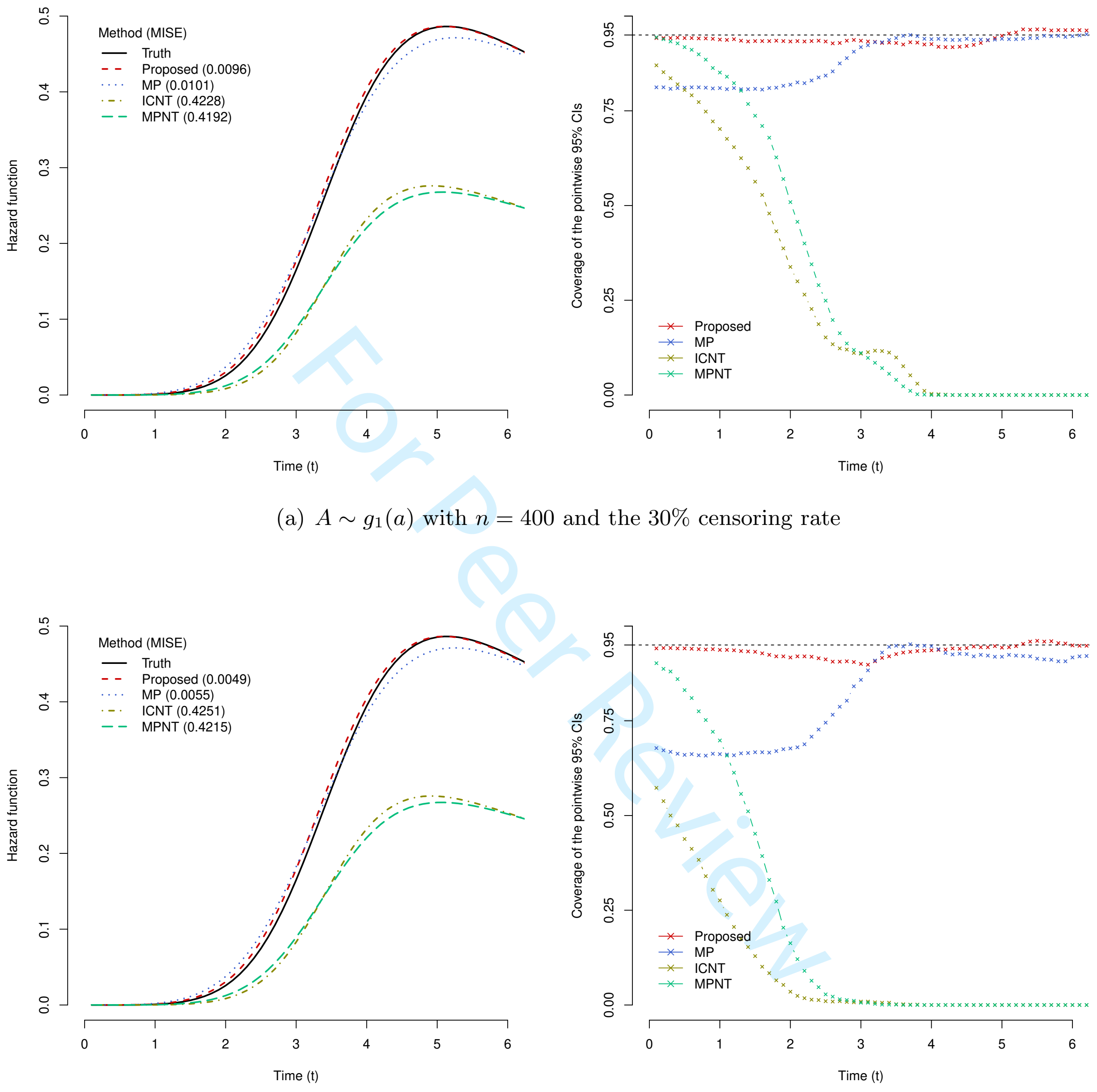
A2 The simulation results of hazard functions



(a) $A \sim g_1(a)$ with $n = 400$ and the 15% censoring rate

(b) $A \sim g_1(a)$ with $n = 800$ and the 15% censoring rate

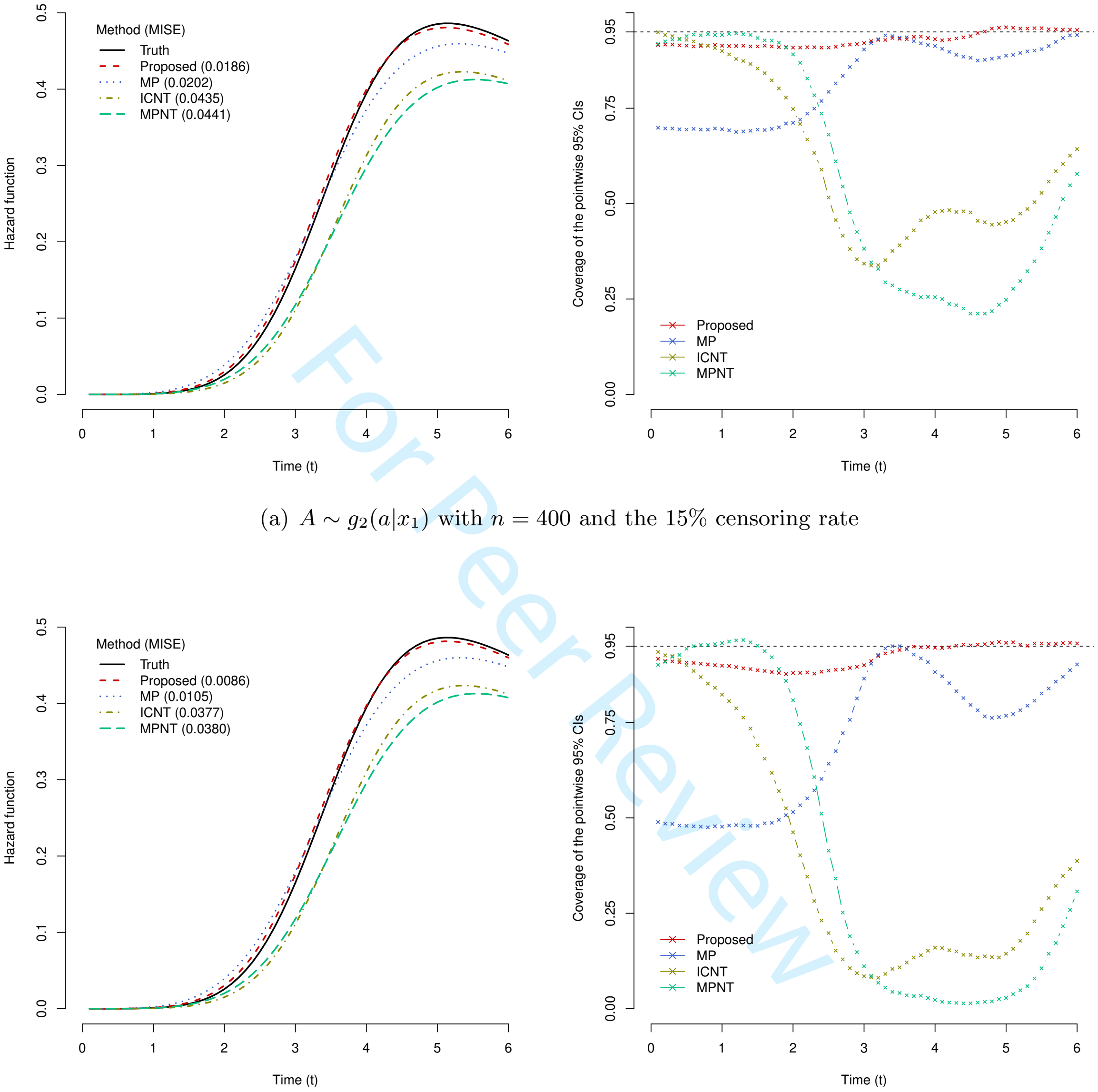
Figure A1: For $A \sim g_1(a)$, the average estimates (the left side) and the empirical coverage probability of the 95% pointwise confidence intervals (the right side) of the hazard functions at $x_1 = 1$ and $x_2 = 0.5$. The Monte Carlo mean of integrated squared error (MISE) is calculated by $\int_0^\infty \{\hat{h}(t|x_1 = 1, x_2 = 0.5) - h(t|x_1 = 1, x_2 = 0.5)\}^2 dt$. The sample sizes are $n = 400$ (a) and $n = 800$ (b) with the 15% censoring rate.



(a) $A \sim g_1(a)$ with $n = 400$ and the 30% censoring rate

(b) $A \sim g_1(a)$ with $n = 800$ and the 30% censoring rate

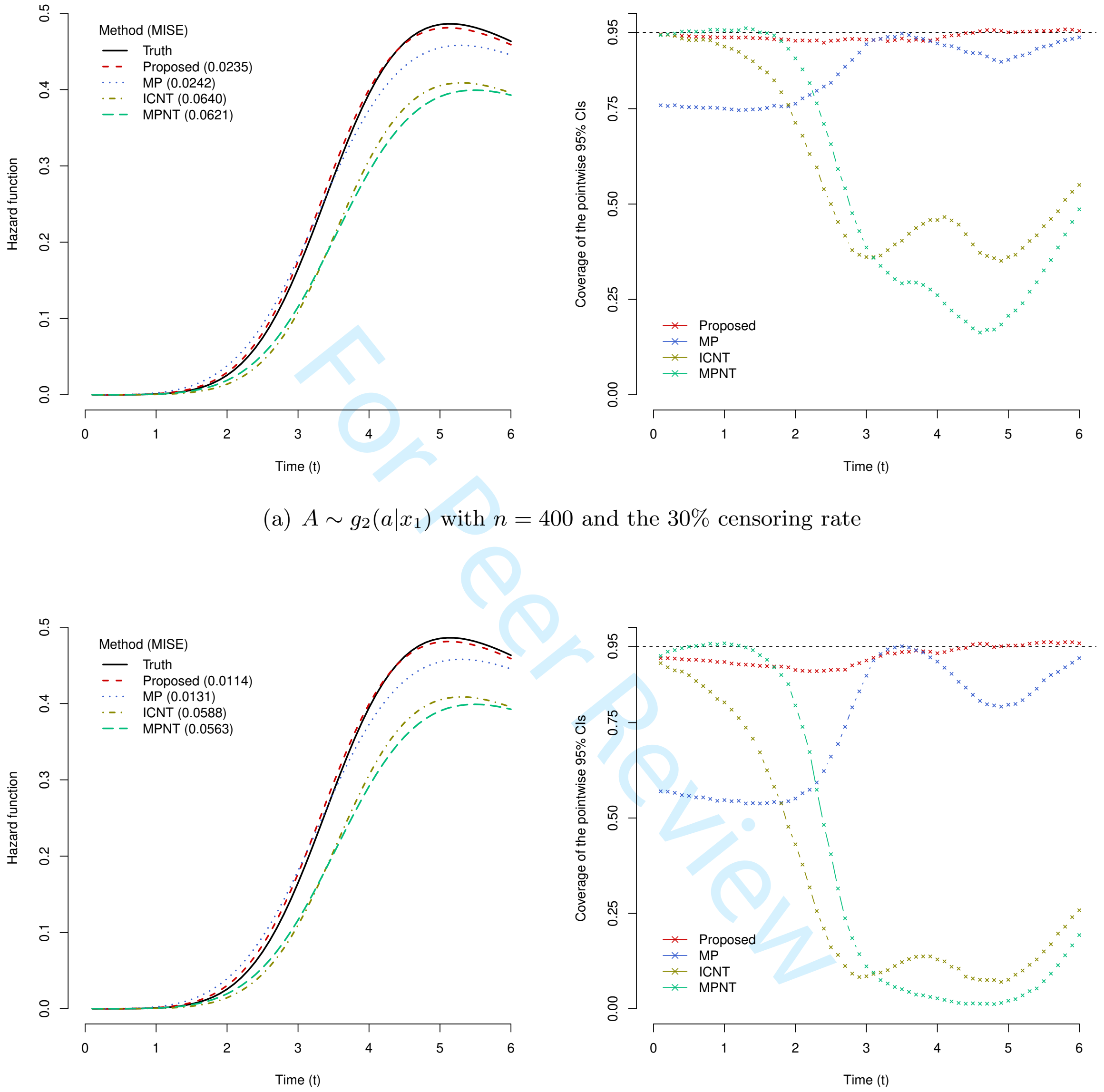
Figure A2: For $A \sim g_1(a)$, the average estimates (the left side) and the empirical coverage probability of the 95% pointwise confidence intervals (the right side) of the hazard functions at $x_1 = 1$ and $x_2 = 0.5$. The Monte Carlo mean of integrated squared error (MISE) is calculated by $\int_0^\infty \{\hat{h}(t|x_1 = 1, x_2 = 0.5) - h(t|x_1 = 1, x_2 = 0.5)\}^2 dt$. The sample sizes are $n = 400$ (a) and $n = 800$ (b) with the 30% censoring rate.



(a) $A \sim g_2(a|x_1)$ with $n = 400$ and the 15% censoring rate

(b) $A \sim g_2(a|x_1)$ with $n = 800$ and the 15% censoring rate

Figure A3: For $A \sim g_2(a|x_1)$, the average estimates (the left side) and the empirical coverage probability of the 95% pointwise confidence intervals (the right side) of the hazard functions at $x_1 = 1$ and $x_2 = 0.5$. The Monte Carlo mean of integrated squared error (MISE) is calculated by $\int_0^\infty \{\hat{h}(t|x_1 = 1, x_2 = 0.5) - h(t|x_1 = 1, x_2 = 0.5)\}^2 dt$. The sample sizes are $n = 400$ (a) and $n = 800$ (b) with the 15% censoring rate.



(a) $A \sim g_2(a|x_1)$ with $n = 400$ and the 30% censoring rate

(b) $A \sim g_2(a|x_1)$ with $n = 800$ and the 30% censoring rate

Figure A4: For $A \sim g_2(a|x_1)$, the average estimates (the left side) and the empirical coverage probability of the 95% pointwise confidence intervals (the right side) of the hazard functions at $x_1 = 1$ and $x_2 = 0.5$. The Monte Carlo mean of integrated squared error (MISE) is calculated by $\int_0^\infty \{\hat{h}(t|x_1 = 1, x_2 = 0.5) - h(t|x_1 = 1, x_2 = 0.5)\}^2 dt$. The sample sizes are $n = 400$ (a) and $n = 800$ (b) with the 30% censoring rate.

A3 Data analysis

Table A5: Data analysis results for the HIV/AIDS cohort data from four approaches: the proposed method (*Proposed*), the middle-point imputation approach adjusting for left-truncation (*MP*), the approach considering interval censoring but ignoring left-truncation (*ICNT*), and the middle-point imputation approach ignoring left-truncation (*MPNT*). For *Proposed* and *MP*, the density for A_0 was a conditional Weibull distribution with the effect α for GENDER.

	$\log(\eta)$	$\log(\gamma)$	α	$\log(\lambda)$	$\log(\phi)$	$\log(\rho)$	GENDER	AGE
Proposed								
Estimate	-5.209	1.795	0.072	2.576	0.176	-1.148	-0.460	0.443
SE	0.008	0.016	0.052	0.319	0.143	2.570	0.406	0.263
pvalue	0.000	0.000	0.163	0.000	0.218	0.655	0.257	0.092
MP								
Estimate	1.272	0.414	0.069	2.547	0.687	-0.114	-0.478	0.706
SE	0.065	0.048	0.243	0.226	0.282	1.136	0.451	0.364
pvalue	0.000	0.000	0.777	0.000	0.015	0.920	0.289	0.053
ICNT								
Estimate	-	-	-	2.453	1.073	0.673	-0.546	0.844
SE	-	-	-	0.193	0.236	0.716	0.525	0.434
pvalue	-	-	-	0.000	0.000	0.347	0.299	0.052
MPNT								
Estimate	-	-	-	2.473	1.052	0.581	-0.530	0.761
SE	-	-	-	0.161	0.190	0.619	0.488	0.363
pvalue	-	-	-	0.000	0.000	0.348	0.278	0.036