Statistica Sinica Preprint No: SS-2016-0231.R2	
Title	Smoothed Rank Regression for the Accelerated Failure Time Competing Risks Model with Missing Cause of Failure
Manuscript ID	SS-2016-0231.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0231
Complete List of Authors	Alan Wan Zhiping Qiu Yong Zhou and Peter Gilbert
Corresponding Author	Alan Wan
E-mail	msawan@cityu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Smoothed Rank Regression for the Accelerated Failure Time Competing Risks Model with Missing Cause of Failure

Zhiping Qiu[1,2], Alan T. K. Wan[3], Yong Zhou[4,5], and Peter B. Gilbert[6]

[1] School of Mathematical Sciences, Huaqiao University, Quanzhou 362021, China

[2] Research Center for Applied Statistics and Big Data, Huaqiao University, Xiamen 361021, China

[3] City University of Hong Kong, Kowloon, Hong Kong

[4] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

[5] Institute of Applied Mathematics, Chinese Academy of Science, Beijing 100190, China

[6] Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

Abstract

This paper examines the accelerated failure time competing risks model with missing cause of failure using the monotone class rank-based estimating equations approach. We handle the non-smoothness of the rank-based estimating equations using a kernel smoothed estimation method, and estimate the unknown selection probability and the conditional expectation by non-parametric techniques. Under this setup, we propose three methods for estimating the unknown regression parameters: inverse probability weighting, estimating equations imputation, and augmented inverse probability weighting. We also obtain the associated asymptotic theories of the proposed estimators and investigate their small sample behaviour in a simulation study. A direct plug-in method is suggested for estimating the asymptotic variances of the proposed estimators. A data application based on a HIV vaccine efficacy trial study is considered.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Keywords: Accelerated failure time model, competing risks, imputation, inverse probability weighting, missing at random, monotone estimating equation, rank-based estimator, U-statistic

1 Introduction

In many areas of research, investigators are interested in studying the effects of different factors on the hazards or failures from a specific cause when failures can result from multiple causes. This leads to the problem of competing risks. This problem arises most frequently in clinical trials where patients may fail from causes other than the disease under investigation. Studies on competing risks that focus on the covariate effects on the cause-specific hazard function for the failure type of interest include Cheng, Fine, and Wei (1998), Shen and Cheng (1999), and Scheike and Zhang (2003). Some authors have also considered direct modeling of the sub-distribution of a competing risk (Fine and Gray (1999); Sun et al., (2006)).

The majority of studies on competing risks to-date assume that the cause of failure is known and observed when the cause of failure may be unknown. For example, in the "Mashi trial" study concerning HIV-related death of infants born to HIV-infected mothers in Botswana considered by Sun, Wang, and Gilbert (2012), the causes of death of live-born infants were known for 50 and missing for 61 of the 111 observations in the study sample. In another study concerning survival times of HIV patients, Bakoyannis, Siannis, and Touloumi (2010) also reported missing causes of death for some sample observations. In our example in Section 6 based on the HVTN 502 'Step' Phase IIb HIV vaccine efficacy trial study, HIV sequences were missing for 23 out of 88 infected participants (Buchbinder

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

et al. (2008)). Methods that account for the missing failure causes to-date assume that the causes of failure are missing at random (MAR), the probability of missingness is only related to the fully observed variables and not to the partially unobserved cause of failure. Studies that focus on the covariate effects on the cause-specific hazard function assuming multiplicative effects when failure causes may be missing include Goetghebeur and Ryan (1995), Lu and Tsiatis (2001), and Gao and Tsiastis (2005). The former two use data imputation methods to compute fitted values for the missing failure causes, while the latter addresses the missing data issue by an augmented inverse probability weighting approach within the framework of a linear transformation model. Lu and Liang (2008) considered an additive hazards model and developed inverse probability weighting and doubly robust methods for estimating the regression coefficients. Other studies on competing risks with missing failure causes that focus on aspects other than the cause-specific hazard function include Bakoyannis, Siannis, and Touloumi (2010), who concentrated on the modelling of the cumulative incidence function, and Sun, Wang, and Gilbert (2012), who considered quantile regression modeling of the survival time.

Recently, Zheng, Lin, and Yu (2016) analysed the competing risks data with missing causes of failure under the accelerated failure time (AFT) model. The AFT model permits a direct measurement of the effects of the covariates on the survival time instead of the hazard function. This facilitates interpretation of results and is considered to be a major advantage of the AFT model over hazards models. One common approach for fitting AFT models is rank-based estimation developed from the weighted log-rank test (Prentice (1978)). This is also the approach taken by Zheng, Lin, and Yu (2016) in their study. When the data are right-censored, the rank-based approach leads to estimators that are consistent and asymptotically normal (Tsiatis (1990); Ying, (1993)). In a recent

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

paper, Lee and Lewbel (2013) provided general identification conditions and developed a sieve maximum likelihood estimation procedure for the AFT model with competing risks data. A shortcoming of the rank-based approach is that rank estimating functions are discontinuous. This feature poses formidable challenges to the computation of the regression coefficient estimates and subsequent inference. Jin et al. (2003) proposed a method that goes some way towards resolving this difficulty. They suggested a monotone approximation to the rank estimating function and a relatively straightforward linear programming-based procedure for estimating the regression coefficients. As the inference procedure of Jin et al.'s (2003) method involves re-sampling, their method can be demanding on computation time, especially with large datasets and for models with many covariates.

Another approach is to apply smoothing methods to the non-smooth estimating functions. The objective of this approach is to construct a smooth surrogate estimating function that is asymptotically equivalent to the original non-smooth function. The continuous differentiability of the surrogate equation ensures that solutions can be obtained by standard numerical algorithms. Brown and Wang (2005) proposed an induced smoothing method whereby the smoothed estimating functions are obtained by taking expectations with respect to an artificial Gaussian continuous noise variable added to the regression coefficients. Heller (2007) employed a direct approximation of the non-smooth function by a local distribution function. As noted by Johnson and Strawderman (2009), when applying Heller's method, if one uses the standard Gaussian cumulative distribution function as the local distribution function, this method will yield the same smoothed estimating functions as Brown and Wang's method that replaces the covariate-dependent bandwidth by a fixed bandwidth. Brown and Wang's (2005) induced smoothing method

has been generalised to estimating functions with general weight (Chiou, Kang, and Yan (2014)), and extended to AFT models with censored data (Brown and Wang (2007); Zhao, Brown, and Wang (2014)), clusterd data (Johnson and Strawderman (2009)), censored and clustered data (Wang and Fu (2011)), and quantile regression (Pang, Lu, and Wang (2012)). To the best of our knowledge, neither Brown and Wang's nor Heller's methods have been applied to AFT models with missing failure causes, and the purpose of this paper is to take steps in this direction.

In this paper, we consider the AFT competing risk model with MAR causes of failure using the monotone rank estimating equations approach. We overcome the difficulty with regard to the discontinuity of the rank estimating equations using a local distribution function smoothing approach in the spirit of Heller (2007). In this setup, we consider three procedures for estimating the unknown regression coefficients. The first is based on a non-parametric inverted probability weighting (IPW) approach, similar to that developed by Qi, Wang, and Prentice (2005) for the proportional hazards model. This approach uses non-parametric smoothers in estimating the selection probabilities, thus overcoming the difficulty with the mis-specification of propensity score frequently encountered with parametric methods. The second method is based on the estimating equation imputation (EEI) approach proposed under a general setup by Zhou, Wan, and Wang (2008). The EEI approach is closely related to the missing information principle; in the context of interest here, studies that apply the missing information principle for the handling of censored data trace back to the work of Buckley and James (1979). The third is an augmented IPW (AIPW) approach in the spirits of Robins, Rotnitzky, and Zhao (1994) who considered a general setup. An important appeal of this approach is that it leads to estimators that are doubly robust. The AIPW approach was considered by Wang and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Chen (2001) for the proportional hazards model.

Although Zheng, Lin, and Yu (2016) also considered the modeling of competing risk data with missing failure causes by a rank-based estimating equations procedure, there are significant differences between their approach and ours. To overcome the difficulty with respect to solving the discontinuous rank estimating equations, Zheng, Lin, and Yu (2016) transformed the problem into an optimisation problem, and the subsequent inference procedure involves re-sampling, which can be demanding on computation time. For the proposed method, as the estimating equations are differentiable with respect to the unknown parameters, estimates of the parameters can be computed by the Newton-Raphson algorithm, and the associated asymptotic variances can be estimated by a plug-in method. We consider the IPW, EEI, and AIPW methods for handling missing data while Zheng, Lin, and Yu (2016) only discussed the IPW and doubly robust methods based on a Martingale with zero mean. In particular, the EEI method we introduce does not require the estimation of the missing probability. We consider the latter a significant advantage. All three missing data handling methods being considered have identical asymptotic properties and comparable finite sample properties.

The remainder of the paper is organised as follows. Section 2 describes the model setup and the smoothed rank estimating equations approach. The three proposed methods for handling missing failure causes and their properties are discussed and examined in Section 3. Section 4 explores the selection of kernel functions and bandwidth parameters, along with a discussion on dimension reduction. Section 5 focuses on the finite sample properties of estimators, while Section 6 considers applications of the proposed methods based on a data set. Some concluding remarks are placed in Section 7. Proofs are contained in the online supplementary file.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

2 Notations, Model Descriptions and A Smoothed Rank Estimating Equations Approach

Let the population contain n independent subjects. For simplicity and without loss of generality, we assume that there are only two mutually exclusive causes of failure, denoted by $J_i = 1, 2$. For the i^{th} $(i = 1, 2, \ldots, n)$ subject, let T_{i1} and T_{i2} be the latent failure times associated with $J_i = 1$ and $J_i = 2$ respectively, $T_i = \min(T_{i1}, T_{i2})$ be the uncensored failure time, C_i the right-censoring time, $\delta_i = I(T_i \leq C_i)$ the censoring indicator such that $\delta_i = 1$ if T_i is observed and $\delta_i = 0$ otherwise, and $\boldsymbol{Z}_i$ be the $p \times 1$ vector of covariates. The observable failure time is thus $\tilde{T}_i = \min(T_i, C_i)$. We assume that C_i, T_{i1} and T_{i2} are mutually independent given $\boldsymbol{Z}_i$.

Suppose that we are only interested in assessing the covariate effect on the failure time of the second type. The AFT model postulates a linear relationship between the natural log of the failure time and the covariates (Kalbfleisch and Prentice (2002)):

$$\log(T_2) = \boldsymbol{Z}^T \boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ is an unknown $p \times 1$ regression coefficient vector, and the error term ϵ has a mean of zero with an unspecified continuous distribution independent of $\boldsymbol{Z}$. When there is no missing cause of failure, the right-censored competing risks data set comprises i.i.d. observations of $(\tilde{T}_i, \delta_i, \delta_i J_i, \boldsymbol{Z}_i), i = 1, \ldots, n$. Let the counting process be $N_i(t) = I\{\log \tilde{T}_i - \boldsymbol{Z}_i^T \boldsymbol{\beta} \leq t, \delta_i J_i = 2\}$, $Y_i(t) = I\{\log \tilde{T}_i - \boldsymbol{Z}_i^T \boldsymbol{\beta} \geq t\}$, and $\lambda(t)$ be the unknown hazard function of ϵ in (1). It can be shown using counting process theory (Fleming and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Harrington (1991)) that

$$M_i(t) = N_i(t) - \int_{-\infty}^{t} Y_i(u)\lambda(u)du, \quad i = 1, 2, \ldots, n,$$

are mean zero martingale processes. By applying arguments as in Tsiatis (1990), we obtain estimating equations for the joint estimation of $\boldsymbol{\beta}$ and $\lambda(t)$:

$$\sum_{i=1}^{n} dM_i(t) = \sum_{i=1}^{n} \big[dN_i(t) - Y_i(t)\lambda(t)dt\big] = 0, \tag{2}$$

$$\sum_{i=1}^{n} \int_{-\infty}^{\tau} \boldsymbol{Z}_i dM_i(t) = \sum_{i=1}^{n} \int_{-\infty}^{\tau} \boldsymbol{Z}_i \big[dN_i(t) - Y_i(t)\lambda(t)dt\big] = 0, \tag{3}$$

where τ is a constant representing the end time of the study. Hence (3) yields

$$\lambda(t)dt = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} Y_i(t)}. \tag{4}$$

Substituting (4) into (3) leads to an estimating equation for $\boldsymbol{\beta}$ in model (1):

$$\sum_{i=1}^{n} \delta_i I(J_i = 2) \left[\boldsymbol{Z}_i - \frac{\sum_{j=1}^{n} \boldsymbol{Z}_j I(\log \tilde{T}_j - \boldsymbol{Z}_j^T\boldsymbol{\beta} \geq \log \tilde{T}_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})}{\sum_{j=1}^{n} I(\log \tilde{T}_j - \boldsymbol{Z}_j^T\boldsymbol{\beta} \geq \log \tilde{T}_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})} \right] = 0. \tag{5}$$

The l.h.s. of (5) is not monotone in $\boldsymbol{\beta}$, and this can produce multiple solutions of $\boldsymbol{\beta}$. To reconcile this difficulty, we consider a monotone rank estimating equation analogous to that proposed by Fygenson and Ritov (1994) for censored data:

$$\tilde{\boldsymbol{U}}_n(\boldsymbol{\beta}) \equiv n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i I(J_i = 2)(\boldsymbol{Z}_i - \boldsymbol{Z}_j) I\big\{ \log(\tilde{T}_j) - \boldsymbol{Z}_j^T\boldsymbol{\beta} \geq \log(\tilde{T}_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta} \big\} = 0. \tag{6}$$

The l.h.s. of (6) is monotone in $\boldsymbol{\beta}$, but it is still discontinuous with respect to $\boldsymbol{\beta}$ due to the presence of an indicator (jump) function in it. A range of well-developed

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

algorithms including the brutal search method, Nelder-Mead method, and linear programming method developed by Jin et al. (2003) can be used for computing $\widehat{\boldsymbol{\beta}}$. However, as the asymptotic covariance matrix of the estimators involves the hazards function of an unspecified error distribution, direct estimation of the covariance matrix requires an estimate of the hazards function. Recognising that this estimate can be highly unstable, Jin et al. (2003) proposed a resampling method to estimate the covariance matrix that eliminates the estimation of the hazards function but the computation efforts involved for the resampling method can be immense, especially with large data-sets. We develop a differentiable estimating equation to approximate (6). Specifically, with $r_i^{\boldsymbol{\beta}} = \log(\tilde{T}_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}, i = 1, 2, \ldots, n$, along the lines of Heller (2007) we consider an approximation to the indicator function $I(r_j^{\boldsymbol{\beta}} \geq r_i^{\boldsymbol{\beta}})$ by a local distribution function $S((r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}})/\sigma_n)$, where $S(u)$ is non-decreasing, $\lim_{u\to\infty} S(u) = 1$, and $\lim_{u\to-\infty} S(u) = 0$, where σ_n is a sequence of strictly positive and decreasing numbers satisfying $\lim_{n\to\infty} \sigma_n = 0$. Clearly, when $r_j^{\boldsymbol{\beta}} > r_i^{\boldsymbol{\beta}}$, $S((r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}})/\sigma_n) \to 1$ as $n \to \infty$, and when $r_j^{\boldsymbol{\beta}} < r_i^{\boldsymbol{\beta}}$, $S((r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}})/\sigma_n) \to 0$ as $n \to \infty$. A smoothed version of (6) is thus given by

$$n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i I(J_i = 2)(\boldsymbol{Z}_i - \boldsymbol{Z}_j) S\Big(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big) = 0. \tag{7}$$

3 Methods for Handling Missing Causes of Failure

When the causes of failure are only partially available, (7) cannot be applied because the J_i are not observed for all i. Let R_i be the complete-case indicator equal to 1 when either $\delta_i = 0$, or $\delta_i = 1$ and J_i is observed, and 0 otherwise. Thus, when the causes of failure

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

are not completely observed, the right-censored competing risks data set comprises i.i.d. observations of $\{(\tilde{T}_i, \delta_i, \boldsymbol{Z}_i, A_i, R_i, R_i\delta_i J_i), i = 1, \ldots, n\}$, where A_i's are some auxiliary covariates that may be useful for predicting the missing failure type.

We assume that the cause of failure is MAR (Rubin (1976)): given $\delta_i = 1$ and $\boldsymbol{W}_i = (\tilde{T}_i, \boldsymbol{Z}_i^T, A_i)^T$, the probability that the failure cause of the i^{th} subject is missing depends only on the observed $\boldsymbol{W}_i$, but not on the unobserved J_i. We assume that the failure cause missing probability is given by

$$r(\boldsymbol{W}_i) = P(R_i = 1|J_i, \delta_i = 1, \boldsymbol{W}_i) = P(R_i = 1|\delta_i = 1, \boldsymbol{W}_i). \tag{8}$$

Although the MAR assumption is more restrictive than nonignorable missingness, MAR is justified in many practical situations, and there is a large collection of literature that uses the MAR assumption as the baseline for analysis. Recent examples include Aerts et al. (2002), Wang and Rao (2002), Chen, Ibrahim, and Shao (2004), Qi, Wang, and Prentice (2005), Lu and Copas (2005), Zhou, Wan and Wang (2008), among others. In the remainder of this section, we develop three methods for dealing with missing data in the context of competing risks data.

3.1 Inverse probability weighting

Write $\boldsymbol{Q}_i = (\boldsymbol{W}_i^T, \delta_i)^T$. From Horvitz and Thompson (1952),

$$M_i^{(1)}(t) \equiv \frac{R_i}{\pi(\boldsymbol{Q}_i)} N_i(t) - \int_{-\infty}^{t} Y_i(u)\lambda(u)du, \ i = 1, 2, \ldots, n,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

are mean zero processes, where $\pi(\boldsymbol{Q}_i) = P(R_i = 1|\delta_i, \boldsymbol{W}_i) = \delta_i r(\boldsymbol{W}_i) + (1 - \delta_i)$. This leads to an inverse probability weighted (IPW) estimating equation for $\boldsymbol{\beta}$:

$$n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{R_i}{\pi(\boldsymbol{Q}_i)} \delta_i I(J_i = 2)(\boldsymbol{Z}_i - \boldsymbol{Z}_j) S\left(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\right) = 0. \tag{9}$$

As $r(\boldsymbol{W}_i)$ is often unknown, we can estimate $r(\boldsymbol{W}_i)$ parametrically as in Gao and Tsiatis (2005), Lu and Liang (2008), and Sun, Wang, and Gilbert (2012), or non-parametrically as in Qi, Wang, and Prentice (2005), Zhou, Wan, and Wang (2008), and Song et al. (2010). Here, we adopt the non-parametric approach that has the advantage over its parametric counterpart of being less prone to biases arising from model mis-specification. We use a kernel method and assume that d is the size of the continuous elements in $\boldsymbol{W}_i$ and $k(u)$ is a rth-order ($r > d$) kernel function with compact support that satisfies $\int k(u)du = 1$, $\int u^m k(u)du = 0$ for $m = 1, 2, \ldots, r-1$, $\int u^r k(u)du \neq 0$, and $\int k^2(u)du < \infty$. As well, for any $\boldsymbol{u} = (u_1, u_2, \ldots, u_d) \in R^d$, let $K_h(\boldsymbol{u}) = \frac{1}{h^d} \prod_{i=1}^{d} k(u_i/h)$, where h is a bandwidth sequence that satisties $nh^{2r} \to 0$ and $nh^{2d} \to \infty$ as $n \to \infty$.

The Nadaraya-Watson estimator (Nadaraya (1964); Watson (1964)) of $r(\boldsymbol{w})$ is then given by

$$\widehat{r}(\boldsymbol{w}) = \widehat{G}_n^{-1}(\boldsymbol{w}) \frac{1}{n} \sum_{i=1}^{n} R_i \delta_i K_h(\boldsymbol{w}_1 - \boldsymbol{W}_{1i}) I(\boldsymbol{W}_{2i} = \boldsymbol{w}_2), \tag{10}$$

where $\widehat{G}_n(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i K_h(\boldsymbol{w}_1 - \boldsymbol{W}_{1i}) I(\boldsymbol{W}_{2i} = \boldsymbol{w}_2)$, $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$, and $\boldsymbol{W}_{1i}$ and $\boldsymbol{W}_{2i}$ are matrices that contain the continuous and discrete elements of $\boldsymbol{W}_i$, respectively. Substituting the estimator $\widehat{r}(\boldsymbol{W}_i)$ into (9) leads to an IPW estimating equation for $\boldsymbol{\beta}$,

$$\boldsymbol{U}_1(\boldsymbol{\beta}) \equiv n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)} \delta_i I(J_i = 2)(\boldsymbol{Z}_i - \boldsymbol{Z}_j) S\left(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\right) = 0, \tag{11}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

where $\widehat{\pi}(\boldsymbol{Q}_i) = \delta_i \widehat{r}(\boldsymbol{W}_i) + (1 - \delta_i)$.

Denote the solution of (11) as $\widehat{\boldsymbol{\beta}}_{IPW}$. The development of an asymptotic theory for $\widehat{\boldsymbol{\beta}}_{IPW}$ requires conditions

(C1) The covariate vector, $\boldsymbol{Z}_1$, is bounded, and there exists a constant M such that, $\| E(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)^T \| < M < \infty$, and the parameter $\boldsymbol{\beta}$ lies in a compact set $\mathcal{B}$.

(C2) The sequence σ_n satisfies the conditions: $n\sigma_n \to \infty$ and $n\sigma_n^4 \to 0$ as $n \to \infty$.

(C3) The local distribution function $S(u)$ is continuous with respect to u, and its first derivative $s(u)$ satisfies the condition $\int u^2 s(u) du < \infty$ and is symmetric about zero.

(C4) The bandwidth h satisfies the conditions: $nh^{2r} \to 0$ and $nh^{2d} \to \infty$ as $n \to \infty$.

(C5) The matrix $\boldsymbol{A} = \nabla \tilde{\boldsymbol{U}}_0(\boldsymbol{\beta}_0)$ exists and is nonsingular, where $\tilde{\boldsymbol{U}}_0(\boldsymbol{\beta}) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \tilde{\boldsymbol{U}}_n(\boldsymbol{\beta})$.

(C6) If $f_{01}(\cdot)$ and $f_{02}(\cdot)$ are the density functions of $\log(T_{11}) - \boldsymbol{Z}_1^T \boldsymbol{\beta}_0$ and $\log(T_{12}) - \boldsymbol{Z}_1^T \boldsymbol{\beta}_0$, respectively, then $f_{01}(\cdot)$, $f'_{01}(\cdot)$, $f_{02}(\cdot)$ and $f'_{02}(\cdot)$ are bounded functions on $\mathcal{R}$ with

$$\int_{-\infty}^{\infty} \left\{ \frac{f'_{01}(t)}{f_{01}(t)} \right\}^2 f_{01}(t) dt < \infty,$$

$$\int_{-\infty}^{\infty} \left\{ \frac{f'_{02}(t)}{f_{02}(t)} \right\}^2 f_{02}(t) dt < \infty.$$

(C7) The distribution of $\log(C_1) - \boldsymbol{Z}_1^T \boldsymbol{\beta}_0$ is absolutely continuous and has a bounded density function $h(\cdot)$ on $\mathcal{R}$.

(C8) The function $g(\boldsymbol{w}) = P(\boldsymbol{W}_1 = \boldsymbol{w}, \delta_1 = 1)$ is bounded away from zero, and has r continuous and bounded partial derivatives with respect to the continuous components of $\boldsymbol{W}_1$, almost surely.

(C9) The conditional probabilities $r(\boldsymbol{w}) = P(R_1 = 1 \big| \delta_1 = 1, \boldsymbol{W}_1 = \boldsymbol{w})$ and $\rho(\boldsymbol{w}) = P(J_1 = 2 \big| \delta_1 = 1, \boldsymbol{W}_1 = \boldsymbol{w})$ are bounded away from zero, and have r continuous and bounded partial derivatives with respect to the continuous components of $\boldsymbol{W}_1$, almost

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

surely.

Let $\boldsymbol{S}_i = (\boldsymbol{Q}_i^T, R_i, J_i)^T, i = 1, 2, \ldots, n$, $\boldsymbol{h}(\boldsymbol{S}_i, \boldsymbol{S}_j) = \delta_i I(J_i = 2)(\boldsymbol{Z}_i - \boldsymbol{Z}_j) I\big\{r_j^{\boldsymbol{\beta}_0} \geq r_i^{\boldsymbol{\beta}_0}\big\}$, and $\boldsymbol{H}(\boldsymbol{S}_i, \boldsymbol{S}_j) = \boldsymbol{h}(\boldsymbol{S}_i, \boldsymbol{S}_j) + \boldsymbol{h}(\boldsymbol{S}_j, \boldsymbol{S}_i)$.

Theorem 1 *If (C1)-(C9) hold, then* $\widehat{\boldsymbol{\beta}}_{IPW} \xrightarrow{p} \boldsymbol{\beta}_0$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{IPW} - \boldsymbol{\beta}_0) \xrightarrow{d} N\big\{0, \boldsymbol{A}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)(\boldsymbol{A}^{-1}(\boldsymbol{\beta}_0))^T\big\},$$

where "$\xrightarrow{p}$" and "$\xrightarrow{d}$" denote convergence in probability and in distribution,

$$\begin{aligned}
\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) &= \boldsymbol{\Gamma}_1(\boldsymbol{\beta}_0) + \boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0), \boldsymbol{\Gamma}_1(\boldsymbol{\beta}_0) = E(\boldsymbol{H}_1(\boldsymbol{S}_1))^{\otimes 2}, \boldsymbol{H}_1(\boldsymbol{S}_1) = E\big(\boldsymbol{H}(\boldsymbol{S}_1, \boldsymbol{S}_2)\big|\boldsymbol{S}_1\big), \\
\boldsymbol{\Gamma}_2(\boldsymbol{\beta}_0) &= E\Big[\big(1 - r(\boldsymbol{W}_i)\big) r^{-1}(\boldsymbol{W}_i)\rho(\boldsymbol{W}_i)\big(1 - \rho(\boldsymbol{W}_i)\big)\delta_i \boldsymbol{\varphi}^{\otimes 2}(\boldsymbol{W}_i)\Big], \quad \textit{and} \\
\boldsymbol{\varphi}(\boldsymbol{w}) &= E\Big\{(\boldsymbol{Z}_1 - \boldsymbol{Z}_2) I\big\{r_2^{\boldsymbol{\beta}_0} \geq r_1^{\boldsymbol{\beta}_0}\big\}\Big|\boldsymbol{W}_1 = \boldsymbol{w}, \delta_1 = 1\Big\}.
\end{aligned}$$

Proof: See the Online Supplementary Material.

Now, for $i, j = 1, 2, \ldots, n$, take

$$e_{ij}^{(1)\boldsymbol{\beta}} = \frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)}\delta_i I(J_i = 2) I\{r_j^{\boldsymbol{\beta}} \geq r_i^{\boldsymbol{\beta}}\}, \quad \text{and} \quad d_{ij}^{(1)\boldsymbol{\beta}} = \Big(\frac{R_i}{\widehat{r}(\boldsymbol{W}_i)} - 1\Big)\delta_i \widehat{\rho}(\boldsymbol{W}_i) I\{r_j^{\boldsymbol{\beta}} \geq r_i^{\boldsymbol{\beta}}\},$$

where $\widehat{\rho}(\boldsymbol{W}_i)$ is defined in (13). Then from the proof of Theorem 1 and the theory of U-statistic (van der Vaart (2000, Ch.12)), we can show that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{IPW}$ can be consistently estimated by $n^{-1}\widehat{\boldsymbol{A}}_{1n}^{-1}(\widehat{\boldsymbol{\beta}}_{IPW})\widehat{\boldsymbol{\Sigma}}_{1n}(\widehat{\boldsymbol{\beta}}_{IPW})(\widehat{\boldsymbol{A}}_{1n}^{-1}(\widehat{\boldsymbol{\beta}}_{IPW}))^T$, where

$$\begin{aligned}
\widehat{\boldsymbol{A}}_{1n}(\boldsymbol{\beta}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)}\sigma_n^{-1}\delta_i I(J_i = 2)\big(\boldsymbol{Z}_i - \boldsymbol{Z}_j\big)^{\otimes 2} s\Big(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big), \text{and} \\
\widehat{\boldsymbol{\Sigma}}_{1n}(\boldsymbol{\beta}) &= \frac{1}{n^3}\sum_{i}\sum_{j}\sum_{k\neq j}(\boldsymbol{Z}_i - \boldsymbol{Z}_j)(\boldsymbol{Z}_i - \boldsymbol{Z}_k)^T(e_{ij}^{(1)\boldsymbol{\beta}} - e_{ji}^{(1)\boldsymbol{\beta}} - d_{ij}^{(1)\boldsymbol{\beta}})(e_{ik}^{(1)\boldsymbol{\beta}} - e_{ki}^{(1)\boldsymbol{\beta}} - d_{ik}^{(1)\boldsymbol{\beta}}).
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

3.2 Estimating equations imputation

We consider the estimating equations imputation (EEI) approach of Zhou, Wan, and Wang (2008). Let $\rho(\boldsymbol{W}_i) = P(J_i = 2|\delta_i = 1, \boldsymbol{W}_i) = P(J_i = 2|R_i = 1, \delta_i = 1, \boldsymbol{W}_i)$. As $E\big[R_iN_i(t) + (1 - R_i)E\big\{N_i(t)\big|\boldsymbol{Q}_i\big\}\big] = E\big[N_i(t)\big]$, it can be shown that

$$\begin{aligned} M_i^{(2)}(t) &\equiv R_iN_i(t) + (1 - R_i)E\big\{N_i(t)|\boldsymbol{Q}_i\big\} - \int_{-\infty}^{t} Y_i(u)\lambda(u)du \\ &= R_iN_i(t) + (1 - R_i)\delta_i\rho(\boldsymbol{W}_i)N_i^*(t) - \int_{-\infty}^{t} Y_i(u)\lambda(u)du, i = 1, 2, \ldots, n, \end{aligned}$$

are mean zero processes, where $N_i^*(t) = I\big\{\log(T_{i2}) - \boldsymbol{Z}_i^T\boldsymbol{\beta} \le t\big\}$. We can then obtain an estimating equation as

$$n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\big[R_iI(J_i = 2) + (1 - R_i)\rho(\boldsymbol{W}_i)\big](\boldsymbol{Z}_i - \boldsymbol{Z}_j)S\Big(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big) = 0. \tag{12}$$

In practice, $\rho(\boldsymbol{W}_i)$ may be unknown. Analogous to the kernel estimator of $r(\boldsymbol{w})$ in Section 3.1, the estimator of $\rho(\boldsymbol{w})$ is

$$\widehat{\rho}(\boldsymbol{w}) = \widehat{M}_n^{-1}(\boldsymbol{W}_i)\frac{1}{n}\sum_{i=1}^{n} I(J_i = 2)R_i\delta_iK_h(\boldsymbol{w}_1 - \boldsymbol{W}_{1i})I(\boldsymbol{W}_{2i} = \boldsymbol{w}_2), \tag{13}$$

where $\widehat{M}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} R_i\delta_iK_h(\boldsymbol{w}_1 - \boldsymbol{W}_{1i})I(\boldsymbol{W}_{2i} = \boldsymbol{w}_2)$. Thus, the EEI estimator $\widehat{\boldsymbol{\beta}}_{EEI}$ is the solution of the estimating equation

$$\boldsymbol{U}_2(\boldsymbol{\beta}) \equiv n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\big[R_iI(J_i = 2) + (1 - R_i)\widehat{\rho}(\boldsymbol{W}_i)\big](\boldsymbol{Z}_i - \boldsymbol{Z}_j)S\Big(\frac{r_j^{\boldsymbol{\beta}} - r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big) = 0. \tag{14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Theorem 2 *If (C1)-(C9) hold, then* $\widehat{\boldsymbol{\beta}}_{EEI} \xrightarrow{p} \boldsymbol{\beta}_0$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{EEI} - \boldsymbol{\beta}_0) \xrightarrow{d} N\big\{0, \boldsymbol{A}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)(\boldsymbol{A}^{-1}(\boldsymbol{\beta}_0))^T\big\},$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ *is defined in Theorem 1.*

Proof: See the Online Supplementary Material.

It is straightforward to show, from the proof of Theorem 2, that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{EEI}$ can be consistently estimated by $n^{-1}\widehat{\boldsymbol{A}}_{2n}^{-1}(\widehat{\boldsymbol{\beta}}_{EEI})\widehat{\boldsymbol{\Sigma}}_{2n}(\widehat{\boldsymbol{\beta}}_{EEI})(\widehat{\boldsymbol{A}}_{2n}^{-1}(\widehat{\boldsymbol{\beta}}_{EEI}))^T$, where

$$\begin{aligned}
\widehat{\boldsymbol{A}}_{2n}(\boldsymbol{\beta}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\big[R_iI(J_i=2)+(1-R_i)\widehat{\rho}(\boldsymbol{W}_i)\big]\sigma_n^{-1}\big(\boldsymbol{Z}_i-\boldsymbol{Z}_j\big)^{\otimes 2}s\Big(\frac{r_j^{\boldsymbol{\beta}}-r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big),\\
\widehat{\boldsymbol{\Sigma}}_{2n}(\boldsymbol{\beta}) &= \frac{1}{n^3}\sum_{i}\sum_{j}\sum_{k\neq j}(\boldsymbol{Z}_i-\boldsymbol{Z}_j)(\boldsymbol{Z}_i-\boldsymbol{Z}_k)^T(e_{ij}^{(2)\boldsymbol{\beta}}-e_{ji}^{(2)\boldsymbol{\beta}}+d_{ij}^{(2)\boldsymbol{\beta}})(e_{ik}^{(2)\boldsymbol{\beta}}-e_{ki}^{(2)\boldsymbol{\beta}}+d_{ik}^{(2)\boldsymbol{\beta}}),\\
e_{ij}^{(2)\boldsymbol{\beta}} &= \delta_i\big[R_iI(J_i=2)+(1-R_i)\widehat{\rho}(\boldsymbol{W}_i)\big]I\{r_j^{\boldsymbol{\beta}}\geq r_i^{\boldsymbol{\beta}}\}, i,j=1,2,\ldots,n, \quad \text{and}\\
d_{ij}^{(2)\boldsymbol{\beta}} &= \big[I(J_i=2)-\rho(\boldsymbol{W}_i)\big]R_i\delta_i\frac{1-\widehat{r}(\boldsymbol{W}_i)}{\widehat{r}(\boldsymbol{W}_i)}I\{r_j^{\boldsymbol{\beta}}\geq r_i^{\boldsymbol{\beta}}\}, i,j=1,2,\ldots,n.
\end{aligned}$$

3.3 Augmented inverse probability weighted estimator

Another common approach for handling data with missing values is the augmented inverse probability weighted (AIPW) method. The AIPW estimator has the double robustness property that the estimator is consistent provided that either $\rho(\boldsymbol{W}_i)$ or $r(\boldsymbol{W}_i)$ is specified correctly (Robins, Rotnitzky, and Zhao (1994); Wang and Chen (2001)).

Using Robins, Rotnitzky, and Zhao (1994), and noting that $E\Big[\frac{R_i}{\pi(\boldsymbol{Q}_i)}N_i(t)+\big(1-\frac{R_i}{\pi(\boldsymbol{Q}_i)}\big)E\big\{N_i(t)|\boldsymbol{Q}_i\big\}\Big]=E\big[N_i(t)\big]$, it follows that

$$M_i^{(3)}(t)\equiv\frac{R_i}{\pi(\boldsymbol{Q}_i)}N_i(t)+\big(1-\frac{R_i}{\pi(\boldsymbol{Q}_i)}\big)E\big\{N_i(t)|\boldsymbol{Q}_i\big\}-\int_{-\infty}^{t}Y_i(u)\lambda(u)du, i=1,2,\ldots,n,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

are mean zero processes. We have the AIPW estimating equations for $\boldsymbol{\beta}$ as:

$$\begin{aligned}\boldsymbol{U}_3(\boldsymbol{\beta}) \;\equiv\; n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\big[\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)}I(J_i=2)\\ +(1-\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)})\widehat{\rho}(\boldsymbol{W}_i)\big](\boldsymbol{Z}_i-\boldsymbol{Z}_j)S\left(\frac{r_j^{\boldsymbol{\beta}}-r_i^{\boldsymbol{\beta}}}{\sigma_n}\right)=0, \qquad (15)\end{aligned}$$

where $\widehat{\pi}(\boldsymbol{Q}_i)=\delta_i\widehat{r}(\boldsymbol{W}_i)+(1-\delta_i)$ and $\widehat{\rho}(\boldsymbol{W}_i)$ are defined in (10) and (13), respectively.

Let the solution of (15) be $\widehat{\boldsymbol{\beta}}_{AIPW}$.

Theorem 3 *If (C1)-(C9) hold, then* $\widehat{\boldsymbol{\beta}}_{AIPW}\xrightarrow{p}\boldsymbol{\beta}_0$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{AIPW}-\boldsymbol{\beta}_0)\xrightarrow{d}N\big\{0,\boldsymbol{A}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)(\boldsymbol{A}^{-1}(\boldsymbol{\beta}_0))^T\big\},$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ *is specified in Theorem 1.*

Proof: See the Online Supplementary Material.

From the proof of Theorem 3, a consistent estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{AIPW}$ is given by $n^{-1}\widehat{\boldsymbol{A}}_{3n}^{-1}(\widehat{\boldsymbol{\beta}}_{AIPW})\widehat{\boldsymbol{\Sigma}}_{3n}(\widehat{\boldsymbol{\beta}}_{AIPW})(\widehat{\boldsymbol{A}}_{3n}^{-1}(\widehat{\boldsymbol{\beta}}_{AIPW}))^T$, where

$$\begin{aligned}\widehat{\boldsymbol{A}}_{3n}(\boldsymbol{\beta}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\big[\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)}I(J_i=2)+(1-\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)})\widehat{\rho}(\boldsymbol{W}_i)\big]\sigma_n^{-1}\big(\boldsymbol{Z}_i-\boldsymbol{Z}_j\big)^{\otimes 2}s\Big(\frac{r_j^{\boldsymbol{\beta}}-r_i^{\boldsymbol{\beta}}}{\sigma_n}\Big),\\
\widehat{\boldsymbol{\Sigma}}_{3n}(\boldsymbol{\beta}) &= \frac{1}{n^3}\sum_{i}\sum_{j}\sum_{k\neq j}(\boldsymbol{Z}_i-\boldsymbol{Z}_j)(\boldsymbol{Z}_i-\boldsymbol{Z}_k)^T(e_{ij}^{(3)\boldsymbol{\beta}}-e_{ji}^{(3)\boldsymbol{\beta}})(e_{ik}^{(3)\boldsymbol{\beta}}-e_{ki}^{(3)\boldsymbol{\beta}}),\quad\text{and}\\
e_{ij}^{(3)\boldsymbol{\beta}} &= \delta_i\big[\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)}I(J_i=2)+(1-\frac{R_i}{\widehat{\pi}(\boldsymbol{Q}_i)})\widehat{\rho}(\boldsymbol{W}_i)\big]I\{r_j^{\boldsymbol{\beta}}\geq r_i^{\boldsymbol{\beta}}\},i,j=1,2,\ldots,n.\end{aligned}$$

Remark 1 The estimators $\widehat{\boldsymbol{\beta}}_{IPW}$, $\widehat{\boldsymbol{\beta}}_{EEI}$, and $\widehat{\boldsymbol{\beta}}_{AIPW}$ are asymptotically equivalent, surprising because we would expect the AIPW method, combining the IPW and EEI approaches, to have improved efficiency.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Remark 2 The implementation of the IPW and AIPW methods requires the estimation of the missing probability $\pi(\boldsymbol{Q}_i)$. Unless one wants to estimate the asymptotic variance directly, the EEI procedure does not involve the estimation of $\pi(\boldsymbol{Q}_i)$. If we resort to a re-sampling method for estimating the asymptotic variance, then the estimation of $\pi(\boldsymbol{Q}_i)$ is not required for the EEI method. Thus, from a computational point of the view, the EEI method has an advantage over the IPW and AIPW methods.

4 Selection of Kernel Functions and Smoothing Parameters, and Dimension Reduction

4.1 Selection of kernel functions and smoothing parameters

In this section, we discuss the selection of the kernel functions $S(\cdot)$ and $k(\cdot)$ and the smoothing parameters σ_n and h. In our numerical studies and the data example, we use the standard Gaussian cumulative distribution function as $S(u)$, the local distribution function. A recent study on the AFT model under length-biased sampling by Qiu, Qin, and Zhou (2016) shows that the finite sample properties of estimators are generally insensitive to the choice of the local distribution function. As for the choice of σ_n, there exist many studies, including Song et al. (2007), Ma and Huang (2007), Lin and Peng (2013), and Qiu, Qin, and Zhou (2016), that show that smoothing approximation techniques similar to ours are applicable under a wide range of choices of σ_n. Here, we use the "rule of thumb" approach along the lines of Heller (2007), to select this smoothing parameter. Specifically, we take $\sigma_n = \widehat{c}n^{-0.26}$, where $\widehat{c} = \Big[\sum_{i=1}^n R_i I\{J_i\delta_i = 2\}(r_i^{\widehat{\boldsymbol{\beta}}_I} - \overline{r}^{\widehat{\boldsymbol{\beta}}_I})^2 \Big/ (\sum_{i=1}^n R_i I\{J_i\delta_i = 2\} - 1)\Big]^{1/2}$, $r_i^{\widehat{\boldsymbol{\beta}}_I} = \log\tilde{T}_i - \boldsymbol{Z}_i^T\widehat{\boldsymbol{\beta}}_I$, $\widehat{\boldsymbol{\beta}}_I$ is the initial

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

estimator obtained by solving the estimating equation

$$n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n} R_i\delta_i I(J_i=2)(\boldsymbol{Z}_i-\boldsymbol{Z}_j)I\{r_j^{\boldsymbol{\beta}}\geq r_i^{\boldsymbol{\beta}}\}=0,$$

and $\overline{r}^{\widehat{\boldsymbol{\beta}}_I}=\sum_{i=1}^{n}R_iI\{J_i\delta_i=2\}r_i^{\widehat{\boldsymbol{\beta}}_I}/\sum_{i=1}^{n}R_iI\{J_i\delta_i=2\}$; the purpose of imposing the power constant -0.26 for n in σ_n is to satisfy condition (C2).

The generalised cross-validation method can be used for choosing the bandwidth h when estimating $r(\boldsymbol{W}_i)$ and $\rho(\boldsymbol{W}_i)$. Here, following Wang and Wang (2001) and Qi, Wang, and Prentice (2005), we set $h=O(n^{-1/q})$ with $q>2d$ and the smallest even integer for r such that $r\geq q-d$. More specifically, when the number of continuous elements in $\boldsymbol{W}_i$ is equal to 1, we use the univariate second-order Epanechnikov kernel $k(u)=\frac{3}{4\sqrt{5}}(1-\frac{1}{5}u^2)I(u^2<5)$ and the bandwidth $h=4\sigma_{\tilde{T}}n^{-1/3}$, where $\sigma_{\tilde{T}}$ is the sample standard deviation of the observed survival times. When $d=2$, we use the fourth-order Epanechnikov kernel $k(u)=\frac{3}{4\sqrt{5}}(\frac{15}{8}-\frac{7}{8}u^2)(1-\frac{1}{5}u^2)I(u^2<5)$ and the bandwidth $(h_1,h_2)^T=(4\sigma_{\tilde{T}}n^{-1/5},4\sigma_Z n^{-1/5})^T$, where σ_Z is the sample standard deviation of Z_i. We use this method to select the kernel function $k(u)$ and the smoothing parameters h in the simulation studies and the data example.

4.2 Dimension reduction

The three proposed methods are based on non-parametric regression. Such methods suffer from the curse of dimensionality, and this limits their usefulness. An alternative is estimate $\pi(\boldsymbol{W}_i)$ and $\rho(\boldsymbol{W}_i)$ parametrically, along the lines of Gao and Tsiatis (2005), Lu and Liang (2008), Sun, Wang, and Gilbert (2012), Zheng, Lin, and Yu (2016), and others. These methods can result in substantially biased estimators when the correctness

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

of the parametric specifications is called into question (Han (2014)).

Dimension reduction is one way to circumvent the problem. The objective is to seek low dimensional variables $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ in the observed data such that $E(R|\boldsymbol{U}_1) = E(R|\delta, \boldsymbol{W}) = E(R|\boldsymbol{Q})$ and $P(J = 2|\boldsymbol{U}_2) = P(J = 2|\delta, \boldsymbol{W}) = P(J = 2|\boldsymbol{Q})$. If we replace $E(R|\boldsymbol{Q})$ and $P(J = 2|\boldsymbol{Q})$ in the estimating equations pertaining to our methods by $E(R|\boldsymbol{U}_1)$ and $P(J = 2|\boldsymbol{U}_2)$, the estimating equations remain unbiased.

Many methods have been developed for the selection of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$. For example, one could assume $E(R|\boldsymbol{Q}) = g_1(\boldsymbol{Q}^T\theta_1)$, and $P(J = 2|\boldsymbol{Q}) = g_2(\boldsymbol{Q}^T\theta_2)$, where $g_1(\cdot)$ and $g_2(\cdot)$ are unknown functions and θ_1 and θ_2 are parameters that can be estimated, for example, by sliced inverse regression (Li (1991)); then $g_1(\cdot)$ and $g_2(\cdot)$ can be estimated by univariate kernel smoothing techniques. Other flexible parametric models, such as the generalised additive and the partially linear models, can be used to model the conditional probabilities $E(R|\boldsymbol{Q})$ and $P(J = 2|\boldsymbol{Q})$. The asymptotic properties of the estimators resulting from these procedures differ from those developed in Section 3. An interesting topic for future research is to develop the asymptotic properties of estimators under this alternative approach.

5 A Simulation Study

In this section, we focus on finite sample properties and identify, in the context of simulations, estimation and inference properties of the methods developed here. We also draw comparisons of our methods with the complete-case analysis that uses only observations that have the failure cause observed.

Experiment 1

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Consider a model containing only one covariate,

$$\log(T_{i2}) = Z_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where T_{i2} is the failure time associated with the cause of interest, Z_i is a covariate that has a Bernoulli(0.5) or $U[0,1]$ distribution, and ϵ_i, the error term, is one of the $N(0, 0.5^2)$, $U[-0.5, 0.5]$, and Generalised Extreme Value $GEV(0, 0.5, 0)$ distributions. All observations of ϵ_i were converted into mean deviation form in our simulations. Given Z_i, we let $\Phi(\log t - \gamma Z_i)$ be the conditional distribution function of the failure time of the other cause T_{i1}, where $\Phi(\cdot)$ is the standard normal cumulative distribution function with γ chosen such that the failure of interest arose approximately 60% of the time. The censoring time C_i was generated from $U[0, c]$, c being a constant that controls the censoring percentage. In all cases, we chose c such that the censoring percentage was about 30%. Depending on the distributional settings of Z_i and ϵ_i, the percentages of failures due to the cause of interest and the other cause varied between 40% and 42%, and between 28% and 30%, respectively. We set $n = 200$ when Z_i was Bernoulli(0.5), and $n = 400$ when Z_i was $U[0,1]$. We considered two missing data scenarios: $r(\boldsymbol{W}_i) = \exp(\tilde{T}_i - Z_i)/\{1+\exp(\tilde{T}_i - Z_i)\}$, and $r(\boldsymbol{W}_i) = 0.5$. In the first case, the missing percentage varied between 70% and 72% depending on the setting of ϵ_i, and under the second, the missing percentage was approximately 50%.

Our simulation results based on 1000 replications are reported in Table 1, where FULL, CC, IPW, EEI, and AIPW refer to results based on the full data set with no missing failure cause, complete-case study, inverse probability weighting, estimating equations imputation, and augmented inverse probability weighting, respectively, and BIAS, SE, SD, and CP denote the empirical bias, the mean of estimated standard error, the empir-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

ical standard deviation and the proximity of empirical coverage probability of confidence interval (C.I.) corresponding to the nominal 95% level.

Our results show that, by and large, the CC method results in the largest bias and smallest C.I. coverage probability. Of the three proposed methods, the IPW method frequently exhibits the largest bias, but it also yields C.I. coverage probability that is as accurate as those produced by the EEI and AIPW approaches. The biases resulting from the EEI and AIPW approaches are usually quite small and the two approaches also achieve very accurate C.I. coverages. Our results do not suggest any clear preference between the EEI and AIPW approaches; generally speaking, there is little to choose between them. In all cases, the SE's and their corresponding SD's are close, indicating that the various non-parametric procedures we use at different stages perform well. As expected, the benchmark estimator based on the full set of data with no missing cause of failure performs best under all performance dimensions being considered. There are no obvious differences between the three types of error distributions, *ceteris paribus*.

Experiment 2

We considered a model with two covariates:

$$\log(T_{i2}) = \beta_{01} Z_{i1} + \beta_{02} Z_{i2} + \epsilon_i,$$

where $Z_{i1} \sim U[0,1]$, Z_{i2} ~Bernoulli(0.5) and ϵ_i was an error term as in Experiment 1. The distribution function of T_{i1}, given Z_{i1} and Z_{i2}, was $\Phi(\log t - \boldsymbol{\gamma}^T \boldsymbol{Z}_i)$, where $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})^T$. The censoring time C_i was generated from $U[0,c]$, c a constant parameter that controls the censoring percentage. As in Experiment 1, we chose $\boldsymbol{\gamma}$ and c such that, on average, 40% of failures were due to the cause of interest, 30% of failures were due to the other cause and the censoring percentage is about 30%. We considered three

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

missing data scenarios: $r(\boldsymbol{W}_i) = \exp(4Z_{i1} + 3Z_{i2} - \tilde{T}_i)/\{1 + \exp(4Z_{i1} + 3Z_{i2} - \tilde{T}_i)\}$; $r(\boldsymbol{W}_i) = 0.5$; $r(\boldsymbol{W}_i) = 1/\{1 + \exp(Z_{i1}^2 - 2Z_{i2})\}$. For the first, the missing probability was approximately 65.9% when $\epsilon_i \sim N(0, 0.25)$ and 64.3% when $\epsilon_i \sim GEV(0, 0.5, 0)$ and $\epsilon_i \sim U[-0.5, 0.5]$. For the other two settings, the missing probabilities were approximately 50% and 59%, respectively. In all cases, we set $n = 400$ and the number of replications to 1000.

The results are presented in Table 2. By and large, the comments made for Experiment 1, where the model contains a single covariate, also apply to the two-covariate case in broad terms. Specifically, the CC method results in estimates with the largest bias in the majority of cases; the IPW, EEI, and AIPW methods generally yield comparable results although the IPW method tends to result in slightly larger estimator bias than the other two methods. Other things being equal, the form of the error distributions does not appear to impact the results significantly.

6 Application to the HVTN 502 Phase IIb 'Step' HIV Vaccine Efficacy Trial

We applied our methods to the HVTN 502 'Step' Phase IIb trial, a randomised, placebo-controlled, preventive vaccine efficacy trial that enrolled HIV-1 uninfected men who had sex with men who were at high risk for acquiring HIV-1 infection to assess whether the incidence of HIV-1 infection differed between the two treatment groups [active vaccination with the Merck adenovirus type 5 (Ad5) vector vaccine (named MRKAd5) vs. placebo](Buchbinder et al. (2008)). The Step trial enrolled 1836 HIV-1 uninfected men, of whom 88 acquired the primary study endpoint of HIV-1 infection (52 in the vaccine

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

group and 36 in the placebo group). The primary analysis assessed the vaccine effect on the time to HIV-1 infection with a Cox model, yielding an estimated hazard ratio (vaccine vs. placebo) of 1.50 (95% C.I.: 0.95–2.41, p-value = 0.06), suggesting that the vaccine elevated the risk of HIV-1 infection.

HIV-1 is extraordinarily genetically diverse, with many genetic types of HIV-1 exposing participants in the Step trial, and a secondary objective of the Step trial was to assess the vaccine effect on the time to HIV-1 infection with specific genetic types of HIV-1. Based on measurement of the HIV-1 sequences from Step participants who had the HIV-1 infection endpoint, there are many ways to define genetic types. Once a definition is specified– such that there are K mutually exclusive and exhaustive genetic types– then the objective at hand is a standard competing risks failure time problem, where T is the time to the first HIV-1 infection and J is the genetic type of the HIV-1 infection, $J \in \{1, \cdots, K\}$. However, HIV-1 sequences were successfully obtained only from 65 of the 88 HIV-1 infected participants, with J missing for 23 participants. A method handling missing failure causes is needed, fitting the purpose for which our methods were designed. In addition to needing a method to handle the missing outcome type J from 23 HIV-1 infected participants, a method was needed to account for the fact that the vaccine effect on the incidence of HIV-1 infection appeared to wane over time (Duerr et al. (2012)). This casts doubt about the suitability of the Cox model and motivates use of the AFT model developed here. Because previous analyses applied a Cox model to address the secondary objective (Rolland et al. (2011)), our methods may yield a better fit to the application.

It is of particular interest to study the vaccine effect on infection with the HIV-1 genetic type defined by high amino acid dissimilarity to a 'hotspot' span of 30 contiguous

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

amino acids in the Gag HIV-1 protein sequence inside the vaccine construct that was targeted by vaccine-induced T cell responses (Hertz et al. (2013)). We took $J = 2$ where there were 2 or more mismatches of the HIV-1 infected participants hotspot sequence with the corresponding hotspot sequence in the vaccine (based on a multiple sequence alignment). Then all HIV-1infections with genetic types with 0 or 1 mismatches have J=1. The distribution of J across the 88 endpoints was 7 for $J = 1$, 32 for $J = 2$, and 14 missing for HIV-1 infected vaccine recipients, 5 for $J = 1$, 21 for $J = 2$, and 9 missing for HIV-1 infected placebo recipients.

We employed the AFT model to evaluate the effect of Treatment (Treatment=1, if the participant was assigned to receive the MRKAd5 vaccine, Treatment=0 placebo), on the failure time T, where T was the number of days from randomisation to diagnosis of HIV-1 infection due to the genotype of interest, $J = 2$. We also included in the model the demographic factors Age (in years at study entry) and WhiteRace (indicator of reporting white race).

Table 3 reports the estimation results. By all methods, Treatment is statistically significant whereas WhiteRace is not. The results for Treatment show that vaccine recipients have a shorter mean time to diagnosis with genotype $J = 2$ HIV-1 infection than placebo recipients, suggesting that vaccination increased susceptibility to acquisition of $J = 2$ HIV-1 genotypes. In addition, the EEI and AIPW methods found that Age was non-significant, but the CC and IPW methods suggest that Age was significant at the 5% level. The EEI and AIPW methods tended to produce estimates of similar magnitudes, and the same was observed for the CC and IPW methods. For a given coefficient, there was no sign difference in the estimates produced by any of the methods.

In conclusion, the analysis suggests that recipients of the MRKAd5 vaccine may have

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

elevated risk of acquiring HIV-1 infection with HIV-1 genetic types that have too many mismatches to the genetic type represented inside the vaccine construct when these mismatches occur in the HIV-1 Gag hotspot location to which the vaccine predominantly directs T cell responses. This highlights the importance of designing new HIV-1 vaccine regimens that direct immune responses to many different genetic types of HIV-1, to maximize overall vaccine efficacy of future HIV-1 vaccines.

7 Concluding Remarks

Competing risks are commonplace in clinical trial studies. We have examined the AFT competing risk model with missing cause of failure using the monotone rank estimating equations approach combined with local distribution function smoothing, and developed three methods for estimating unknown regression coefficients. Our simulation study shows that the three methods work well, and the methods have been applied to a data set on HIV vaccine efficacy. We have also discussed methods of dimension reduction that can be undertaken in conjunction with the methods developed when the number of covariates is large. Our proposed methods can be extended to other semi-parametric models, such as the generalized transformation models and the mean residual lifetime model. These remain for future work.

8 Supplementary Materials

Additional simulation results and proofs of the main theorems are contained in the Online Supplementary Materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Acknowledgements

Qiu's work was supported by the Education and Scientific Research Projects of Young and Middle-aged Teachers in Fujian Province, China (No. JAT160027). Wan's work was supported by a General Research Fund and a Theme-Based Research Scheme from the Hong Kong Research Grants Council (No. 9042086 and No. T32-102/14N). Zhou's work was supported by National Natural Science Foundation of China (NSFC) (No. 71271128), the State Key Program of National Natural Science Foundation of China (No. 71331006), NCMIS, Key Laboratory of RCSDS, CAS and Shanghai First-class Discipline A and IRTSHUFE, PCSIRT (No. IRT13077). Gilbert's work was supported by the National Institute Of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Numbers R37AI054165 and UM1AI068635. We thank the Editor, an associate editor, and two referees for their comments and suggestions on an earlier version of this paper. The usual disclaimer applies.

References

Aerts, M., Claeskens, G., Hens, N. and Molenberghs, G. (2002). Local multiple imputation. *Biometrika* **89**, 375-388.

Bakoyannis, G., Siannis, F. and Touloumi, G. (2010). Modelling competing risks data with missing cause of failure. *Statist. Med.* **29**, 3172-3185.

Brown, B.M. and Wang, Y.G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika* **92**, 149-158.

Brown, B. M. and Wang, Y. G. (2007). Induced smoothing for rank regression with censored survival times. *Statist. Med.* **26**, 828-836.

Buchbinder, S. P., Mehrotra, D. V., Duerr, A., Fitzgerald, D. W., Mogg, R., Li, D., Gilbert, P. B., Lama, J. R., Marmor, M., Del Rio C, McElrath, M. J., Casimiro, D. R., Gottesdiener, K. M., Chodakewitz, J. A., Corey, L. and Robertson, M. N. (2008). Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet* **372(9653)**, 1881-1893.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429-436.

Chen, M. H., Ibrahim, J. G. and Shao, Q. M. (2004). Propriety of the posterior distribution and existence of the MLE for regression models with covariates missing at random. *J. Amer. Statist. Assoc.* **99**, 421-438.

Cheng, S. C., Fine, J. P. and Wei, L. J. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **54**, 219-228.

Chiou, S. H., Kang, S. and Yan, J. (2014). Fast accelerated failure time modeling for case-cohort data. *Statist. Comput.* **24**, 559-568.

Duerr A., Huang Y., Buchbinder S., Coombs, R. W., Sanchez J., del Rio, C., Casapia, M., Santiago, S., Gilbert, P. B., Corey, L. and Robertson, M. N. (2012). Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study). *J. Infect. Dis.* **206**, 258-266.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Amer. Statist. Assoc.* **94**, 496-509.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Fygenson, M. and Ritov, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22**, 732-746.

Gao, G. and Tsiatis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika* **92**, 875-891.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821-834.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *J. Amer. Statist. Assoc.* **109**, 26-41.

Heller, G. (2007). Smoothed rank regression with censored data. *J. Amer. Statist. Assoc.* **102**, 552-559.

Hertz, T., Ahmed, H., Friedrich, D. P., Casimiro, D. R., Self, S. G., Corey, L., McElrath, M. J., Buchbinder, S., Horton, H., Frahm, N., Robertson, M. N., Graham, B. S. and Gilbert, P. B. (2013). HIV-1 Vaccine-induced T-Cell reponse cluster in epitope hotspots that differ from those induced in natural infection with HIV-1. *PLoS Patho.* **9**, e1003404.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Hyun, S., Lee, J. and Sun, Y. (2012). Proportional hazards model for competing risks data with missing cause of failure. *J. Stat. Plan. Infer.* **142**, 1767-1779.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.

Johnson, L. M. and Strawderman, R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika* **96**, 577-590.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., New York: Wiley.

Lee, S. and Lewbel, A. (2013). Nonparametric identification of accelerated failure time competing risks models. *Economet. Theor.* **29**, 905-919.

Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Lu, G., and Copas, J. B. (2005). Missing at random, likelihood ignorability and model completeness. *Ann. Statist.* **32**, 754-765.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57**, 1191-1197.

Lu, W. and Liang, Y. (2008). Analysis of competing risks data with missng cause of failure under additive hazards model. *Statist. Sinica* **18**, 219-234.

Nadaraya, E. A. (1964). On estimating regression. *Theor. Probab. Appl.* **9**, 141-142.

Pang, L., Lu, W. and Wang, H. (2012). Variance estimation in censored quantile regression via induced smoothing. *Comput. Statist. Data An. 56*, 785-796.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-180.

Qi, L., Wang, Y. C. and Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* **100**, 1250-1263.

Qiu, Z., Qin, J. and Zhou, Y. (2016). Composite Estimating Equation Method for the Accelerated Failure Time Model with Length-biased Sampling Data. *Scand. J. Statist.* **43**, 396-415.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Rolland M., Tovanabutra S., deCamp A. C., Frahm N., Gilbert P. B., Sanders-Buell E., Heath L., Magaret C. A., Bose M., Bradfield A., O'Sullivan A., Crossler J., Jones T., Nau M., Wong K., Zhao H., Raugi D. N., Sorensen S., Stoddard J. N., Maust B. S., Deng W., Hural J., Dubey S., Michael N. L., Shiver J., Corey L., Li F., Self S. G., Kim J., Buchbinder S., Casimiro D. R., Robertson M. N., Duerr A., McElrath M. J., McCutchan F. E., Mullins J. I. (2011). Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat. Med.* **17**, 366-371.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Scheike, T. H. and Zhang, M. J. (2003). Extensions and applications of the Cox-Aalen survival model. *Biometrics* **59**, 1036-1045.

Shen, Y. and Cheng, S. C. (1999). Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics* **55**, 1093-1100.

Song, X. Y., Sun, L. Q., Mu, X. Y. and Dinse, G. E. (2010). Additive hazards regression with censoring indicators missing at random. *Can. J. Statist.* **38**, 333-351.

Sun, L. Q., Liu, J., Zhang, M. and Sun, J. (2006). Modeling the subdistribution of a competing risk. *Statist. Sinica* **16**, 1367-1385.

Sun, Y. Q., Wang, H. J. and Gilbert, P. B. (2012). Quantile regression for competing risks data with missing cause of failure. *Statist. Sinica* **22**, 703-728.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354-372.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wang, C. Y. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics* **57**, 414-419.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896-924.

Wang, S. and Wang, C. Y. (2001). A note on kernel-assisted estimators in missing covariate regression. *Statist. Probabil. Lett.* **55**, 439-449.

Wang, Y. G. and Fu, L. (2011). Rank regression for the accelerated failure time model with clustered and censored data. *Comput. Statist. Data An.* **55**, 2334-2343.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā (series A)* **26**, 359-372.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76-99.

Zhao, Y. D., Brown, B. M. and Wang, Y. G. (2014). Smoothed rank-based procedure for censored data. *Electron. J. Statist.* **8**, 2953-2974.

Zheng, M., Lin, R. and Yu, W. (2016). Competing risks data analysis under the accelerated failure time model with missing cause of failure. *Ann. Inst. Statist. Math.* **68**, 855-876.

Zhou, Y., Wan, A. T. K. and Wang, X. (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc.* **103**, 1187-1199.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Table 1: Simulation results of Experiment 1

		$Z_i \sim$ Bernoulli(0.5)				$Z_i \sim U[0,1]$			
		BIAS	SE	SD	CP	BIAS	SE	SD	CP
Scenario 1									
$\epsilon_i \sim$ N(0,0.25)	FULL	-0.008	0.094	0.091	93.5%	0.001	0.112	0.111	95.1%
	CC	0.043	0.110	0.104	92.0%	0.061	0.130	0.129	92.2%
	IPW	-0.000	0.104	0.097	93.7%	0.016	0.125	0.119	93.4%
	EEI	-0.002	0.102	0.097	94.4%	0.017	0.118	0.119	94.9%
	AIPW	-0.006	0.104	0.097	93.9%	0.004	0.125	0.119	94.0%
$\epsilon_i \sim$ U[-0.5,0.5]	FULL	-0.001	0.055	0.055	94.9%	-0.001	0.070	0.067	94.2%
	CC	0.017	0.069	0.065	92.9%	0.020	0.086	0.081	91.9%
	IPW	0.001	0.063	0.062	94.7%	0.003	0.080	0.077	94.1%
	EEI	0.012	0.064	0.062	94.1%	0.013	0.078	0.078	94.2%
	AIPW	0.002	0.065	0.062	93.8%	0.001	0.081	0.076	93.1%
$\epsilon_i \sim$ GEV(0,0.5,0)	FULL	-0.005	0.132	0.129	94.0%	0.000	0.159	0.156	95.1%
	CC	0.097	0.143	0.139	88.1%	0.110	0.176	0.171	88.7%
	IPW	0.012	0.141	0.133	93.0%	0.021	0.174	0.166	93.8%
	EEI	-0.010	0.138	0.134	93.3%	0.004	0.166	0.166	95.6%
	AIPW	-0.009	0.143	0.135	92.2%	-0.004	0.175	0.166	94.0%
Scenario 2									
$\epsilon_i \sim$ N(0,0.25)	FULL	-0.002	0.090	0.091	95.2%	0.005	0.113	0.111	94.4%
	CC	0.069	0.129	0.127	90.6%	0.083	0.154	0.154	91.0%
	IPW	-0.004	0.111	0.106	95.0%	0.002	0.137	0.131	94.2%
	EEI	-0.000	0.102	0.107	95.7%	0.022	0.122	0.131	96.1%
	AIPW	-0.004	0.105	0.106	95.4%	0.002	0.134	0.131	94.6%
$\epsilon_i \sim$ U[-0.5,0.5]	FULL	-0.000	0.057	0.055	94.5%	0.003	0.069	0.067	95.1%
	CC	0.034	0.081	0.076	91.2%	0.042	0.096	0.093	92.3%
	IPW	-0.001	0.072	0.070	93.9%	0.003	0.089	0.089	95.1%
	EEI	0.014	0.073	0.071	93.7%	0.020	0.082	0.088	96.1%
	AIPW	-0.002	0.074	0.070	93.2%	0.002	0.088	0.085	94.8%
$\epsilon_i \sim$ GEV(0,0.5,0)	FULL	-0.001	0.132	0.128	94.7%	-0.004	0.158	0.156	94.8%
	CC	0.118	0.190	0.181	88.7%	0.127	0.225	0.218	91.0%
	IPW	-0.001	0.160	0.146	92.5%	-0.000	0.194	0.181	94.1%
	EEI	-0.013	0.146	0.145	95.0%	0.011	0.168	0.176	95.8%
	AIPW	-0.004	0.152	0.146	94.4%	-0.002	0.182	0.178	94.6%

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0231

Table 2: Simulation results of Experiment 2

		$\beta_{01}=1$				$\beta_{02}=1$			
		BIAS	SE	SD	CP	BIAS	SE	SD	CP
Scenario 1									
$\epsilon_i \sim$ N(0,0.25)	FULL	-0.004	0.116	0.111	94.0%	-0.001	0.066	0.064	94.2%
	CC	0.039	0.141	0.133	92.0%	-0.004	0.080	0.075	93.7%
	IPW	-0.032	0.139	0.129	92.0%	-0.001	0.076	0.073	94.8%
	EEI	-0.008	0.125	0.126	95.7%	-0.004	0.070	0.072	95.7%
	AIPW	-0.009	0.137	0.126	93.2%	-0.005	0.073	0.073	95.1%
$\epsilon_i \sim$ U[-0.5,0.5]	FULL	0.003	0.068	0.068	95.2%	-0.001	0.038	0.039	95.1%
	CC	0.018	0.080	0.078	93.7%	-0.008	0.045	0.044	94.1%
	IPW	-0.024	0.081	0.084	95.0%	-0.010	0.044	0.046	95.4%
	EEI	0.012	0.078	0.085	96.8%	0.005	0.045	0.047	96.6%
	AIPW	-0.006	0.088	0.082	95.3%	-0.011	0.048	0.047	96.0%
$\epsilon_i \sim$ GEV(0,0.5,0)	FULL	0.005	0.161	0.158	94.4%	0.003	0.091	0.090	95.2%
	CC	0.097	0.202	0.202	92.5%	0.002	0.114	0.113	95.0%
	IPW	-0.007	0.191	0.191	95.1%	0.022	0.104	0.106	94.8%
	EEI	-0.002	0.171	0.177	96.0%	-0.001	0.096	0.100	96.1%
	AIPW	0.008	0.189	0.177	94.2%	0.010	0.101	0.102	95.5%
Scenario 2									
$\epsilon_i \sim$ N(0,0.25)	FULL	0.002	0.111	0.112	95.2%	-0.000	0.066	0.065	95.2%
	CC	0.061	0.158	0.155	92.9%	0.072	0.092	0.090	86.2%
	IPW	-0.002	0.141	0.134	94.0%	0.002	0.081	0.076	93.3%
	EEI	-0.003	0.124	0.133	96.9%	0.005	0.075	0.076	95.7%
	AIPW	-0.002	0.138	0.133	94.7%	-0.000	0.080	0.076	94.0%
$\epsilon_i \sim$ U[-0.5,0.5]	FULL	0.002	0.069	0.068	93.3%	-0.002	0.037	0.039	96.4%
	CC	0.041	0.098	0.094	90.8%	0.030	0.054	0.054	90.7%
	IPW	0.003	0.088	0.088	94.1%	-0.004	0.048	0.049	95.6%
	EEI	0.006	0.083	0.089	96.1%	0.011	0.048	0.050	96.1%
	AIPW	0.002	0.094	0.087	93.4%	-0.005	0.051	0.050	94.7%
$\epsilon_i \sim$ GEV(0,0.5,0)	FULL	0.001	0.164	0.158	94.1%	0.006	0.093	0.090	94.1%
	CC	0.104	0.236	0.225	90.8%	0.128	0.137	0.128	83.0%
	IPW	0.002	0.211	0.186	91.8%	0.009	0.118	0.104	91.5%
	EEI	-0.002	0.177	0.183	95.6%	-0.003	0.107	0.103	94.2%
	AIPW	0.005	0.183	0.117	94.0%	0.006	0.113	0.104	92.8%
Scenario 3									
$\epsilon_i \sim$ N(0,0.25)	FULL	0.002	0.109	0.112	95.8%	-0.001	0.065	0.064	95.7%
	CC	0.060	0.139	0.142	93.0%	-0.018	0.085	0.083	94.2%
	IPW	0.035	0.137	0.134	93.7%	-0.007	0.080	0.075	93.4%
	EEI	0.003	0.120	0.132	97.4%	-0.001	0.076	0.075	94.7%
	AIPW	0.003	0.130	0.131	95.1%	-0.004	0.078	0.075	94.2%
$\epsilon_i \sim$ U[-0.5,0.5]	FULL	0.002	0.068	0.068	95.4%	-0.001	0.040	0.039	94.9%
	CC	0.033	0.089	0.087	92.1%	-0.009	0.053	0.051	93.8%
	IPW	0.024	0.090	0.090	94.7%	-0.005	0.049	0.050	96.2%
	EEI	-0.000	0.085	0.090	96.9%	0.021	0.049	0.050	93.1%
	AIPW	-0.003	0.097	0.086	93.4%	-0.002	0.052	0.049	94.7%
$\epsilon_i \sim$ GEV(0,0.5,0)	FULL	0.007	0.159	0.159	94.6%	0.001	0.090	0.090	95.0%
	CC	0.103	0.210	0.200	90.9%	-0.024	0.124	0.117	93.5%
	IPW	0.053	0.206	0.186	91.0%	-0.008	0.113	0.105	93.4%
	EEI	0.012	0.174	0.181	95.6%	-0.022	0.103	0.104	95.0%
	AIPW	0.012	0.196	0.183	93.6%	-0.006	0.107	0.105	94.7%

Table 3: *Estimation of the effects of WhiteRace, Age and Treatment for the HVTN 502 Step HIV vaccine efficacy trial data*

	WhiteRace			Age			Treatment		
Method	EST	$\widehat{SE}$	P-value	EST	$\widehat{SE}$	P-value	EST	$\widehat{SE}$	P-value
CC	-0.460	0.403	0.254	0.040	0.013	0.002	-0.743	0.152	0.000
IPW	-0.501	0.402	0.213	0.039	0.013	0.002	-0.681	0.167	0.000
EEI	-0.321	0.428	0.453	0.022	0.013	0.084	-0.786	0.132	0.000
AIPW	-0.296	0.426	0.488	0.024	0.013	0.064	-0.772	0.132	0.000