Annals of the Institute of Statistical Mathematics
https://doi.org/10.1007/s10463-023-00874-6

Estimation of complier causal treatment effects with informatively interval-censored failure time data

Yuqing Ma[1] · Peijie Wang[1] · Jianguo Sun[2]

Received: 21 November 2022 / Revised: 27 March 2023 / Accepted: 4 April 2023

Abstract
Estimation of compiler causal treatment effects has been discussed by many authors under different situations but only limited literature exists for interval-censored failure time data, which often occur in many areas such as longitudinal or periodical follow-up studies. Particularly it does not seem to exist a method that can deal with informative interval censoring, which can happen naturally and make the analysis much more challenging. Also, it has been shown that when the informative censoring exists, the analysis without taking it into account would yield biased or misleading results. To address this, we propose an estimated sieve maximum likelihood approach with the use of instrumental variables. The asymptotic properties of the resulting estimators of regression parameters are established, and a simulation study is performed and suggests that it works well. Finally, it is applied to a set of real data that motivated this study.

Keywords Causal inference · Informative censoring · Interval-censored data · Proportional hazards model

1 Introduction

In observational studies, people always want to estimate the causal effects of a treatment, but sometimes the existence of unmeasured confounding can affect the time-to-event outcomes and make the process of estimating much complicate. Applying naive analysis methods in such circumstances may result in severe bias (Zeng 2012). To deal with the endogenous selection into treatments, where treatment choice is bound up with unmeasured confounders of potential outcomes and can greatly complicate the estimation of causal treatment effects, people introduce the instrumental

✉ Peijie Wang
wangpeijie@jlu.edu.cn

[1] School of Mathematics, Jilin University, 2699 Qianjin Street, Changchun 130012, China

[2] Department of Statistics, University of Missouri, MO 65211 Columbia, USA

Published online: 15 May 2023

variables (IVs), the variables that are independent of unmeasured confounders but related to the treatment and influencing the outcome only through its effect on the treatment (Angrist et al. 1996; Baiocchi et al. 2014; Li and Lu 2015; Li and Gray 2016).

Many IV methods have been proposed for the situation with completely observed outcomes or right-censored data. For example, Abadie et al. (2002) proposed a general weighting strategy for estimating complier treatment effect under linear or nonlinear treatment response models. Cheng et al. (2009) further used empirical likelihood techniques to construct semiparametric and nonparametric estimators of the complier outcome distribution. Baker (1998) and Nie et al. (2011) studied the estimation of causal hazard difference in compliers with discrete right-censored survival data without considering covariates. Cuzick et al. (2007) estimated the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. Lin et al. (2014) and Yu et al. (2015) considered causal linear transformation models with right censoring and developed pseudo-likelihood and maximum likelihood based estimation, respectively.

Interval censoring has much more complicated structures than right censoring. Unlike right-censored data that include exact event times for some subjects, interval-censored data only include time intervals where the event times belong to but not exact event times (Sun 2006). It is easy to see that interval-censored data include right-censored data as a special case (Kalbfleisch and Prentice 2011). When analyzing the interval-censored data, the occurrence of informative censoring always makes the problem more complicated, which means that the censoring mechanism is related to the failure time of interest (Huang and Wolfe 2002; Wang et al. 2016; Zhang et al. 2007).

One general situation where informative interval censoring can occur is a periodic follow-up study of certain disease where study subjects may not follow the prespecified visit schedules and instead pay clinical visits according to their disease status or how they feel with respect their treatments. Among others, Ma et al. (2015) discussed this for current status data arising from the proportional hazards model, Zhang et al. (2005) gave some procedures for regression analysis of case II informatively interval-censored case and Wang et al. (2016) considered case K informatively interval-censored data. Many authors have pointed out that if it exists, the analysis that ignores it could yield biased or misleading results or conclusions (Sun 1999).

A special study that motivated this investigation is the HVTN 505 Trial. It is well known that HIV-1 infection is deadly as it causes AIDS for which there is no cure and thus it is important and essential to develop a safe and effective vaccine for the prevention of the infection. The HVTN 505 Trial is designed to assess the efficacy of a DNA prime-recombinant adenovirus type 5 boost (DNA/rAd5) vaccine to prevent human immunodeficiency virus type 1 (HIV-1) infection (Du et al. 2021). The original study consists of 2504 subjects who were examined periodically, and some people did not receive examination as schedules, thus yielding only informative interval-censored data on the time to HIV-1 infection.

As a result, we have to consider how to deal with the informative censoring in the causal inference. Li and Peng (2021) estimated complier causal treatment effect with independent interval-censored data, but there does not seem to exist an established

method for dependent censoring. In this situation, the challenges from interval censoring in causal inference become even more prominent, because we need to deal with the informative censoring and unmeasured confounding at the same time.

The remainder of the paper is organized as follows. Section 2 will introduce the notation, the assumed models under the causal framework and informative censoring, and the likelihood function. The proposed sieve maximum likelihood estimation approach as well as the asymptotic properties of the proposed estimators will be established in Sect. 3. In Sect. 4, we present some results obtained from a simulation study conducted to examine the empirical performance of the proposed method. An illustration of HIV/AIDS study is given in Sect. 5, and Sect. 6 contains some discussion and concluding remarks.

2 Notion, models and likelihood function

Consider a failure time study that consists of n independent subjects. For subject i, $i = 1, 2, \ldots, n$, let T_i denote the failure time of interest, D_i, A_i and X_i denote a binary treatment variable, a binary instrument variable and a $p \times 1$ covariate vector, respectively. The assignment of a treatment in a clinical trial is a simple example of an IV, because it is strongly related to the actual treatment received and not contaminated by unmeasured confounding, which fits the definition of IV well. Besides, many other variables can also be viewed as IV as long as they correspond to its definition.

Let $D_i = 1$ if subject i takes the treatment 1 (treatment group) and $D_i = 0$ if subject i takes treatment 0 (control group). In this paper, we suppose that the observed data arise from a two-arm randomized clinical trial. For all possible instrument variable vectors, define the potential treatment indicator $D_i^a = 1$ if $A_i = a$ and subject i takes up the treatment offered in the treatment group, and $D_i^a = 0$ otherwise. Let T_i^d denote the potential event time for subject i with $D_i = d$. Also, let $D^{\tilde{a}}$ denote the potential treatment given $(A_1, \ldots, A_n)^T = \tilde{a} = (a_1, \ldots, a_n)^T$ and $T^{\tilde{d}}$ the potential treatment given $(T_1, \ldots, T_n)^T = \tilde{d} = (d_1, \ldots, d_n)^T$. Note that the notation with tilde represents the corresponding population analogues.

To give a intuitive explanation, here we use non-compliance and treatment assignment to represent the unmeasured confounders and IV, respectively. The whole population can be divided into four latent compliance subgroups denoted by U. Specifically, we let U_i equal 1 if subject i is an always taker (i.e., $D_i^0 = D_i^1 = 1$), 2 if subject i is a complier (i.e., $D_i^1 > D_i^0$), 3 if subject i is a never taker (i.e., $D_i^0 = D_i^1 = 0$) and 4 if subject i is a defier (i.e., $D_i^1 < D_i^0$). Note that U_i cannot be fully determined by the observed data because D_i^0 and D_i^1 cannot be observed at the same time, so we need an additional model of the latent compliance class U_i.

Suppose that only an interval-censored observation is available and given by $(L_i, R_i, \Delta_{1i}, \Delta_{2i}, \Delta_{3i})$ with $L_i \le R_i$, where L_i and R_i denote two examination times, $\Delta_{1i} = I(T_i \le L_i)$, $\Delta_{2i} = I(L_i < T_i \le R_i)$, and $\Delta_{3i} = 1 - \Delta_{1i} - \Delta_{2i}$. In practice such as in medical follow-up studies, there usually exists an administrative censoring time $\tilde{C}_i$ beyond which the observation process is no longer available. We define $\tilde{\Delta}_i$ to be the event indicator of the ith individual who experiences an event during the study period.

Thus, $\Delta_{3i} = 1 - \tilde{\Delta}_i$. Let $\Delta_i = (\Delta_{1i}, \Delta_{2i}, \Delta_{3i})$. By following Li and Peng (2021) and others, throughout the paper, we will adopt the standard IV assumptions.

Assumption 1 (*stable unit treatment value assumption*). For $i = 1, \ldots, n$, we have that $D_i^{\tilde{a}} = D_i^{\tilde{a}'}$ if $a_i = a_i^{'}$, where $\tilde{a} = (a_1, \ldots, a_n)^T \in \{0, 1\}^n$ and $\tilde{a}' = (a_1^{'}, \ldots, a_n^{'})^T \in \{0, 1\}^n$. Also, we have that $T_i^{\tilde{d}} = T_i^{\tilde{d}'}$ if $d_i = d_i^{'}$, where $\tilde{d} = (d_1, \ldots, d_n)^T \in \{0, 1\}^n$ and $\tilde{d}' = (d_1^{'}, \ldots, d_n^{'})^T \in \{0, 1\}^n$.

Assumption 2 (*random sampling*) $\{ (D_i^0, D_i^1, T_i^0, T_i^1, X_i, A_i), i = 1, \ldots, n \}$ are independent and identically distributed.

Assumption 3 (*independence of the instrument*). For all i, we have that

$$(D_i^0, D_i^1, T_i^0, T_i^1, L_i, R_i, \Delta_i) \perp A_i | X_i,$$

where the symbol '$\perp$' denotes statistical independence.

Assumption 4 (*conditional non-null compliance class*). $P(D^1 > D^0 | X) > 0$.

Assumption 5 (*conditional monotonicity*). $P(D^1 \geq D^0 | X) = 1$.

Assumption 6 (*exclusion restriction*) $P(T^{1d} = T^{0d}) = 1$ for $d = 0, 1$, where T^{ad} denote the potential time to event given $A = a$ and $D = d$ with $a = 0$ or 1.

The previous researches always have an additional assumption about independent censoring that $(T^0, T^1) \perp (L, R, \Delta) | X$, but in this paper, we can relaxing this assumption and give a new method which can deal with informatively interval censoring. We will discuss it in the following.

As discussed above, we need to construct a model for unobserved U, because it is easy to see that the likelihood function will involve the compliance type probabilities $p_{ik} = P(U_i = k | X_i = x_i)$, which is depends on the covariate $X_i, i = 1, \ldots, n, k = 1, 2, 3$. For this, we will assume that they follow the multinomial logistic model

$$\log \frac{P(U = k|X)}{P(U = 2|X)} = \theta_k^T \tilde{X}, \ k = 1, 2, 3 \tag{1}$$

for U, where θ_k denotes a vector of parameters, $\tilde{X} = (1, X)^T$, and $\theta_2 = 0_{(p+1)\times 1}$. Define $\theta = (\theta_1^T, \theta_3^T)^T$ and $p_k(\theta) = P(U = k|X)$. Then under Assumption 5, the model above can be equivalently expressed as

$$p_k(\theta) = \frac{\exp(\theta_k^T \tilde{X})}{1 + \exp(\theta_1^T \tilde{X}) + \exp(\theta_3^T \tilde{X})}, \ k = 1, 2, 3.$$

To construct a model that can explain both the effects of the treatment and covariates to the potential time-to-event outcome T^d and the informative censoring at the same time, we assume that within each latent compliance subgroup, there exists a

latent variable c with mean one and given X, d and c, the hazard function of T^d for subject i has the form

$$\begin{aligned}
\lambda_{T^1}(t|D=1,X,c,U=1) &= \lambda_t(t)\exp(\beta_{01}+\gamma_1^T X)c,\\
\lambda_{T^d}(t|D=d,X,c,U=2) &= \lambda_t(t)\exp(\beta_2 d+\gamma_2^T X)c,\\
\lambda_{T^0}(t|D=0,X,c,U=3) &= \lambda_t(t)\exp(\beta_{03}+\gamma_3^T X)c,
\end{aligned} \tag{2}$$

where $\lambda_t(\cdot)$ is an unknown baseline hazard function, and we assume that c is independent of (L, X, U, A, D). In this model, β_2 represents how the conditional hazard function of the potential event time T^1 given covariates X and c differs from that of T^0 in the complier subgroup. The reason why there is no treatment variable D in the first and last subgroup in model (2) is always-takers (i.e., $U=1$) always receive treatment $D=1$ and never-takers (i.e., $U=3$) always receive treatment $D=0$, which means the treatment has no effect on the potential time-to-event outcome T^d.

In practice, the causal quantity

$$\mathrm{CESP}(t|X) = P(T^1 > t|U=2,X,c) - P(T^0 > t|U=2,X,c)$$

is a function of T given X and c. It is a more direct statement about the complier causal treatment effect than β_2, which represents the causal difference in the survival probability of a complier with covariates X between the two potential scenarios of receiving treatment versus not receiving treatment. Note that from model (2), we have that

$$P(T^d \le t|D=d,X,c,U=2) = 1-\exp\{-\Lambda_t(t)\exp(\beta_2 d+\gamma_2^T X)c\},\ d=0,1,$$

where $\Lambda_t(t)=\int_0^t \lambda_t(u)du$. Then, a natural estimator, denoted by $\widehat{\mathrm{CESP}}(t|X)$, can be obtained by plugging in the estimators of $\Lambda_t(\cdot)$, β_2 and γ_2 defined above.

Another causal quantity that may be of interest is

$$\mathrm{CESP}(t) = P(T^1 > t|U=2,c) - P(T^0 > t|U=2,c)\,,$$

the unconditional counterpart of $\mathrm{CESP}(t|X)$, which is a function of T only based on c. To estimate $\mathrm{CESP}(t)$, note that by applying the Bayes theorem and the independence assumption of c_i described earlier, we have that

$$\mathrm{CESP}(t) = \int \mathrm{CESP}(t|X) f_{X|U=2}(x)\mathrm{d}x = \frac{E\{\mathrm{CESP}(t|X)P(U=2|X)\}}{E\{P(U=2|X)\}},$$

where $f_{X|U=k}(x)$ denotes the density function of X given $U=k$, $k=1,2,3$. Consequently, a plug-in estimator for $\mathrm{CESP}(t)$ is given by

$$\widehat{\mathrm{CESP}}(t) = \frac{n^{-1}\sum_{i=1}^n p_{i2}(\hat\theta_n)\widehat{\mathrm{CESP}}(t|X_i)}{\bar p_2(\hat\theta_n)},$$

where $\bar p_2(\hat\theta_n) = n^{-1}\sum_{i=1}^n p_{i2}(\hat\theta_n)$, $p_{i2}(\hat\theta_n) = \{1+\exp(\hat\theta_{1n}^T\tilde X_i)+\exp(\hat\theta_{3n}^T\tilde X_i)\}^{-1}$, and $\hat\theta_n$ denotes the estimator of θ defined above.

Note that under the IV independence, we have $P(T^d > t|X, c, U = 2) = P(T^d > t|A = d, X, c, U = 2) = P(T > t|D = d, X, c, U = 2)$. In addition, from the definitions of compliers and never-takers, $D = 1$ when $U = 1$ and $D = 0$ when $U = 3$, so $T = T^1$ given $U = 1$ and $T = T^0$ given $U = 3$. Let $\nu = (\nu_1^T, \nu_2^T, \nu_3^T)^T$, where $\nu_1 = (\beta_{01}, \gamma_1^T)^T$, $\nu_2 = (\beta_2, \gamma_2^T)^T$, $\nu_3 = (\beta_{03}, \gamma_3^T)^T$. Also, let $Z_1 = Z_3 = (1, X^T)^T$, and $Z_2 = (D, X^T)^T$. As a result, the formula (2) can be integrated into

$$\lambda(t|D, X, c, U = k) = \lambda_t(t) \exp(\nu_k^T Z_k)c, \ k = 1, 2, 3,$$

where $\lambda(\cdot|D, X, c, U = k)$ denotes the hazard function of T given D, X and c in the latent compliance class $U = k$. The above model give a direct link between the causal estimand and the observed event time T.

As discussed above, dependent interval censoring often occurs. To address this problem, define $W_i = R_i - L_i$, and by following Ma et al. (2016) and others, we assume that the dependent censoring can be characterized by the correlation between the T_i's and W_i's. Furthermore, it is supposed that W_i follows the proportional hazards frailty model given by

$$\lambda_i^{(W)}(t|X_i, c_i) = \lambda_w(t) \exp(\beta_w^T X_i)\, c_i\,, \tag{3}$$

where $\lambda_w(t)$ denotes an unknown baseline hazard functions and β_w is a $p \times 1$ vector of unknown regression parameters. In the following, it will be assumed that the latent variables c_i's have a distribution known up to the unknown variance α and given c_i, W_i is independent of T_i. More comments on this will be given below.

Note that c_i is independent of $(L_i, X_i, U_i, A_i, D_i)$ and the joint distribution of (L_i, X_i) does not involve the parameters of interest. Thus with the observation

$$O = \left\{ O_i = (L_i, R_i, \Delta_i, A_i, D_i, X_i); i = 1, \ldots, n \right\}$$

and under the assumptions above, the likelihood contribution for each i and k conditional on (W_i, L_i, c_i) could be written as

$$\begin{aligned} L_{ik|W_i,L_i,c_i} = & [1 - \exp\{-\Lambda_t(L_i) \exp(\nu_k^T z_{ik})c_i\}]^{\Delta_{1i}} \\ & \times [\exp\{-\Lambda_t(L_i) \exp(\nu_k^T z_{ik})c_i\} - \exp\{-\Lambda_t(R_i) \exp(\nu_k^T z_{ik})c_i\}]^{\Delta_{2i}} \\ & \times [\exp\{-\Lambda_t(R_i) \exp(\nu_k^T z_{ik})c_i\}]^{\Delta_{3i}}. \end{aligned}$$

In addition, the likelihood contribution from the observed W_i conditional on (X_i, c_i) is given by

$$L_{W_i|c_i} = \{\lambda_w(t) \exp(\beta_w^T X_i)c_i \exp\{-\Lambda_w(t) \exp(\beta_w^T X_i)c_i\}\}^{\chi_i},$$

where $\Lambda_w(t) = \int_0^t \lambda_w(u)du$. As stated in Ma et al. (2016), due to the existence of $\tilde{C}_i$, it is possible that R_i is not observed but right-censored at $\tilde{C}_i$ and T_i is right-censored. Thus, $\tilde{\Delta}_i = 0$ implies that the ith individual is right-censored. In this case, one may only have one observation time L_i with $T_i > \tilde{C}_i > L_i$ and $W_i > \tilde{C}_i - L_i$. In other words, one may have $W_i = \infty$ and for this, we define $\chi_i = I(W_i < \infty)$.

For the subject i in subgroup $U_i = k$, we have the following likelihood function

$$f_{ik} = \int L_{ik|W_i,L_i,c_i} L_{W_i|c_i} g(c_i;\alpha) dc_i,$$

where $g(c_i;\alpha)$ denotes the density function of the c_i. If g is the gamma distribution, the function f_{ik} has the closed form

$$\begin{aligned} f_{ik} = & \left(\lambda_w \text{ e x p } (\beta_w^T X_i)\right)^{\chi_i} \left[\left(1 + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right)^{-\alpha^{-1}-\chi_i}\right. \\ & \left. - \left(1 + \alpha\Lambda_t\left(L_i\right) \exp(\beta_t^T X_i) + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right)^{-\alpha^{-1}-\chi_i}\right]^{\Delta_{1i}} \\ & \times \left[\left(1 + \alpha\Lambda_t\left(L_i\right) \exp(\beta_t^T X_i) + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right)^{-\alpha^{-1}-\chi_i}\right. \\ & \left. - \left(1 + \alpha\Lambda_t\left(L_i + W_i\right) \exp(\beta_t^T X_i) + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right)^{-\alpha^{-1}-\chi_i}\right]^{\Delta_{2i}} \\ & \times \left[\left(1 + \alpha\Lambda_t\left(L_i + W_i\right) \exp(\beta_t^T X_i) + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right)^{-\alpha^{-1}-\chi_i}\right]^{\Delta_{3i}}. \end{aligned} \tag{4}$$

In particular, if $T \geq L_i + W_i$, then the likelihood contribution is given by

$$f_{ik} = \left[1 + \alpha\Lambda_t\left(L_i + W_i\right) \exp(\beta_t^T X_i) + \left(\alpha \Lambda_w\left(W_i\right) \exp\left(\beta_w^T X_i\right)\right)\chi_i\right]^{-\alpha^{-1}-\chi_i}.$$

Then, it is easy to see that under the assumptions above, the likelihood contribution from subject i is given by $\sum_{k=1}^{3} f_{ik} p_{ik} I(U_i = k)$.

To derive the observed likelihood function, note that the study subjects can be classified into three groups based on D and A under the assumption 5. For the subjects with $D_i = 1$ and $A_i = 1$, their likelihood contributions are proportional to $f_{i1}p_{i1} + f_{i2}p_{i2}$ and one can show that

$$\begin{aligned} & P(\tilde{\Delta}_i = 1, D_i = 1, A_i = 1 | X_i = x_i, c_i) \\ & = P(L_i < T_i \leq R_i | U_i = 1, X_i = x_i, c_i) P(U_i = 1 | X_i = x_i) P(A_i = 1 | X_i = x_i) \\ & \quad + P(L_i < T_i \leq R_i | U_i = 2, D_i = 1, X_i = x_i, c_i) P(U_i = 2 | X_i = x_i) P(A_i = 1 | X_i = x_i). \end{aligned}$$

In fact, it also holds when we only have one observation time (let $R_i = \infty$). Similarly, the likelihood contributions from the subject i with $D_i = 0$ and $A_i = 1$, $D_i = 1$ and $A_i = 0$, or $D_i = A_i = 0$ are proportional to $f_{i3}p_{i3}$, $f_{i1}p_{i1}$, or $f_{i2}p_{i2} + f_{i3}p_{i3}$, respectively. It follows that the log-likelihood contribution from subject i conditional on $(L_i, \Delta_i, \chi_i, \chi_i W_i, X_i, D_i, A_i)$ could be written as

$$\begin{aligned} l_n(\xi) = \sum_{i=1}^{n} \log\{ & f_{i1}p_{i1} I(d_i = 1, a_i = 0) + (f_{i2}p_{i2} + f_{i3}p_{i3}) I(d_i = a_i = 0) \\ & + f_{i3}p_{i3} I(d_i = 0, a_i = 1) + (f_{i1}p_{i1} + f_{i2}p_{i2}) I(d_i = a_i = 1)\}, \end{aligned}$$

where ξ denotes the vector containing all unknown parameters ν, θ, α, Λ_t and Λ_w.

3 Sieve maximum likelihood estimation

Let $\eta = (\nu^T, \theta^T, \alpha)^T$ and $\Psi = \{(\eta, \Lambda_t, \Lambda_w) \in \mathcal{E} \otimes \mathcal{M}_1 \otimes \mathcal{M}_2\}$ denotes the parameter space. Here, $\mathcal{E} = \{\eta \in R^{p_\eta}, \|\eta\| \le B\}$ with p_η being the dimension of η and B a positive constant, and $\mathcal{M}_j$ is the collections of all bounded and continuous non-decreasing, nonnegative functions over the interval $[\sigma_j, \tau_j]$, where $0 \le \sigma_j < \tau_j < \infty$, $j = 1, 2$. To estimate ξ, a natural approach would be to maximize the log-likelihood function $l_n(\xi)$. On the other hand, it is easy to see that this is not an easy task since $l_n(\xi)$ involves both finite-dimensional and infinite-dimensional parameters. To deal with this, by following Zhou et al. (2016) and others, we will employ the sieve approach.

Specifically, define the sieve space $\Psi_n = \{(\eta, \Lambda_{tn}, \Lambda_{wn}) \in \mathcal{E} \otimes \mathcal{M}_{1n} \otimes \mathcal{M}_{2n}\}$, where

$$\mathcal{M}_{1n} = \Big\{ \Lambda_{tn}(t) = \sum_{p=0}^{m} \phi_{1p}^* B_{1p}^*(t, m, \sigma_1, \tau_1) : \sum_{0 \le p \le m} |\phi_{1p}^*| \le M_n, \\ 0 \le \phi_{10}^* \le \phi_{11}^* \le \cdots \le \phi_{1m}^* \Big\},$$

$$\mathcal{M}_{2n} = \Big\{ \Lambda_{wn}(t) = \sum_{p=0}^{m} \phi_{2p}^* B_{2p}^*(t, m, \sigma_2, \tau_2) : \sum_{0 \le p \le m} |\phi_{2p}^*| \le M_n, \\ 0 \le \phi_{20}^* \le \phi_{21}^* \le \cdots \le \phi_{2m}^* \Big\},$$

with $B_{jp}^*(t, m, \sigma_j, \tau_j)$ denoting the Bernstein basis polynomials defined as

$$B_{jp}^*\big(t, m, \sigma_j, \tau_j\big) = \binom{m}{p} \left(\frac{t - \sigma_j}{\tau_j - \sigma_j}\right)^p \left(1 - \frac{t - \sigma_j}{\tau_j - \sigma_j}\right)^{m-p}, \ p = 0, \ldots, m, j = 1, 2,$$

and $m = o(n^q)$ the degree of the polynomials for some $q \in (0, 1)$. Note that $0 \le \phi_{j0}^* \le \phi_{j1}^* \le \cdots \le \phi_{jm}^*$ can be viewed as a partition here, $j = 1, 2$. Also, note that in the above definition, there is a restriction $0 \le \phi_{j0}^* \le \phi_{j1}^* \le \cdots \le \phi_{jm}^*$, which can be relaxed by letting $\phi_{j0}^* = \exp(\phi_{j0})$, $\phi_{jp}^* = \sum_{s=0}^{p} \exp(\phi_{js})$, $1 \le p \le m$.

Let $B_{jp}(t) = \sum_{i=p}^{m} B_{ji}^*(t, m, \sigma_j, \tau_j)$. Then, we have that

$$\Lambda_{tn}(t) = \sum_{p=0}^{m} \sum_{i=0}^{p} \exp(\phi_{1i}) B_{1p}^*(t, m, \sigma_1, \tau_1) = \sum_{p=0}^{m} \exp(\phi_{1p}) B_{1p}(t),$$

$$\Lambda_{wn}(t) = \sum_{p=0}^{m} \sum_{i=0}^{p} \exp(\phi_{2i}) B_{2p}^*(t, m, \sigma_2, \tau_2) = \sum_{p=0}^{m} \exp(\phi_{2p}) B_{2p}(t),$$

and it is natural to define the sieve maximum likelihood estimator $\hat{\xi}_n = (\hat{\eta}_n, \hat{\Lambda}_{tn}, \hat{\Lambda}_{wn})$ to be the value of ξ that maximizes the log-likelihood function $l_n(\xi)$ over Ψ_n.

Computationally, the algorithm for acquiring the sieve maximum likelihood estimator can be carried out as follows:

Step 1 Set initial values of parameters ξ and cumulative baseline hazards Λ_t and Λ_w. For example, we can set all initial values as $\xi^{(0)}$=($\nu^{(0)}$, $\theta^{(0)}$, $\alpha^{(0)}$, $\phi_{jp}^{(0)}$), $j = 1, 2$, $p = 0, 1, \ldots, m$, to 0.1 and initial values for Λ_t and Λ_w as $\Lambda_t^{(0)}$ and $\Lambda_w^{(0)}$ to $0.1t$ and $0.1w$.

Step 2 According to L, R and the initial values for $\phi_{jp}^{(0)}$, $\Lambda_t^{(0)}$ and $\Lambda_w^{(0)}$, get the Bernstein basis polynomials $B_{jp}^{(0)}$, $j = 1, 2$, $p = 0, 1, \ldots, m$.

Step 3 Based on values of $\xi^{(k)}$, use $\phi_{jp}^{(k)}$ and $B_{jp}^{(k)}$ to express $\Lambda_t^{(k)}$ and $\Lambda_w^{(k)}$ in the likelihood function and update the estimation $\xi^{(k+1)}$=($\nu^{(k+1)}$, $\theta^{(k+1)}$, $\alpha^{(k+1)}$, $\phi_{jp}^{(k+1)}$), $j = 1, 2$, $p = 0, 1, \ldots, m$, by maximizing $l(\xi^{(k)})$ over the sieve space.

Step 4 Repeat Step 3 until the minimum absolute difference of the parameters between two consecutive iterations is less than a given small positive value.

To determine $\hat{\xi}_n$, it is apparent that one needs to choose or select the degree of Bernstein polynomials m, which controls the smoothness of the approximation. For this, a natural approach is to consider several different values of m and choose the value of m that gives the minimum of the AIC

$$\mathrm{AIC} = -2\, l_n(\hat{\xi}_n) + 2\,(6p + 2m + 8)\,.$$

Of course, one may employ other similar criteria such as the BIC and the numerical results indicate that they give similar performance. For the determination of the proposed estimator $\hat{\xi}_n$, in the numerical study below, we employ the fminunc function in MATLAB.

Let $\xi_0 = (\eta_0, \Lambda_{t0}, \Lambda_{w0})$ denote the true value of ξ. In this section, we will establish the asymptotic properties of $\hat{\xi}_n$. Let $\|v\|$ denote the Euclidean norm for a vector v and define the supremum norm $\|f\|_\infty = \sup_t |f(t)|$ for a function f. Furthermore, for any $\xi^1 = (\eta^1, \Lambda_t^1, \Lambda_w^1)$ and $\xi^2 = (\eta^2, \Lambda_t^2, \Lambda_w^2)$ in the parameter space Ψ, define the distance

$$d(\xi^1, \xi^2) = \{||\eta^1 - \eta^2||^2 + ||\Lambda_t^1 - \Lambda_t^2||_2^2 + ||\Lambda_w^1 - \Lambda_w^2||_2^2\}^{1/2}.$$

Here, $||\Lambda_t^1 - \Lambda_t^2||_2^2 = \{\int_{\sigma_1}^{\tau_1} [\Lambda_t^1(t) - \Lambda_t^2(t)]^2 dt\}^{1/2}$ and $||\Lambda_w^1 - \Lambda_w^2||_2^2 = \{\int_{\sigma_2}^{\tau_2} [\Lambda_w^1(t) - \Lambda_w^2(t)]^2 dt\}^{1/2}$. The following theorems establish the asymptotic consistency and normality of the proposed estimators.

Theorem 1 *Suppose that the regularity conditions* (*A*.1)–(*A*.5) *given in Appendix hold. Then, we have that* $d(\hat{\xi}_n, \xi_0) \to 0$ *as* $n \to \infty$.

Theorem 2 *Suppose that the regularity conditions* (*A*.1)–(*A*.5) *given in Appendix hold. Then, we have that* $d(\hat{\xi}_n, \xi_0) = O_p(n^{-min\{(1-q)/2, vq/2\}})$, *where* $q \in (0, 1)$ *such that* $m = o(n^q)$ *and* v *is defined in condition* (*A*.3).

Theorem 3 *Suppose that the regularity conditions* (*A*.1)–(*A*.5) *given in Appendix hold. Then, as* $n \to \infty$ *and if* $q > 1/2v$, *we have that*

$$n^{1/2}(\hat{\eta}_n - \eta_0) \to N(0, \Sigma),$$

where Σ is given in Appendix.

The proofs of the results above are sketched in Appendix. To make inference about η_0, it is apparent that one natural way is to develop a consistent estimate for the asymptotic covariance of $n^{1/2}(\hat{\eta}_n - \eta_0)$. For this, we suggest to use the inverse of the Hessian matrix of the log-likelihood function $l_n(\xi)$ by treating it as a function of the finite-dimensional parameters $(\eta, \phi_{1p}, \phi_{2p})$, $p = 0, \ldots, m$. For the estimation of the variances of CESP($t|X$) and CESP(t), we suggest to employ the delta method, which is time saving compared with the bootstrap procedure. The numerical study below indicates that it works well.

4 A simulation study

In this section, we present some results obtained from a simulation study conducted to assess the performance of the estimation procedure proposed in the previous sections. For the generation of the potential failure time T_i^d's, we first generated the A_i's from the Bernoulli(0.5) distribution and the U_i's under model (1), and then, we can get D_i's naturally. T_i^d's were generated under model (2) with $\lambda_t(t) = 0.3$. Note that here the IV and D are set to be the random treatment assignment and the actual treatment received, respectively.

To generate the observation process and thus the censoring intervals, by mimicking real medical follow-up studies, it was assumed that there exists a sequence of pre-specified observation time points $t_1 < \cdots < t_k$, and each subject is supposed to be observed at each time point with probability p_0. Then, if T was smaller than the smallest real observation time, L was set to be the smallest real observation time, and if T was larger than all of the real observation times, L was taken to be the largest real observation time. Otherwise, L was set to be the largest real observation time that is smaller than T. For the informative censoring mechanism, we set $R = L + W$ with W generated from model (3) with $\lambda_w(t) = 6t$. The simulation results given below are based on $p_0 = 0.8$, sample size $n = 500$ or 1000 with 1000 replications and the degree of Bernstein polynomials being $m = n^{1/4}$.

First, we considered the situation with a single covariate X generated from the continuous uniform distribution over (0, 1). Table 1 presents the results on estimation of the complier causal treatment effect β_2 and other parameters in η obtained by the proposed approach with the continuous covariate and $\beta_{01} = \beta_2 = \beta_{03} = -0.3$, $\beta_w = 0.6$, $\gamma_1 = \gamma_2 = \gamma_3 = 0.6$, $\theta_{01} = \theta_1 = \theta_{03} = \theta_3 = -0.3$ and $\alpha = 0.6$. They include the empirical biases (Bias), the sample standard errors (SE), the average of the estimated standard errors (SEE) and the 95% empirical coverage probabilities (CP). One can see from Table 1 that the proposed estimators seem to be unbiased and the Hessian matrix variance estimation procedure appears to work well. Also, the CP results suggest that the normal approximation to the distribution of the estimator seems

Table 1 Simulation results with covariate distributed from U(0,1)

Para	n=500				n=1000			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\beta_{01} = -0.3$	− 0.0062	0.6839	0.6509	0.942	− 0.0044	0.4525	0.4453	0.948
$\beta_2 = -0.3$	− 0.0263	0.3826	0.3597	0.941	− 0.0138	0.2510	0.2386	0.952
$\beta_{03} = -0.3$	0.0143	0.7293	0.6887	0.943	− 0.0033	0.5004	0.4751	0.944
$\beta_w = 0.6$	0.0107	0.2282	0.2405	0.961	0.0067	0.1582	0.1670	0.959
$\gamma_1 = 0.6$	0.0453	0.7050	0.6813	0.942	− 0.0063	0.4757	0.4696	0.950
$\gamma_2 = 0.6$	0.0364	0.6705	0.6348	0.946	0.0105	0.4437	0.4236	0.949
$\gamma_3 = 0.6$	− 0.0142	0.7412	0.6987	0.934	0.0112	0.4762	0.4782	0.946
$\theta_{01} = -0.3$	0.0133	0.3818	0.3721	0.939	0.0121	0.2623	0.2590	0.956
$\theta_1 = -0.3$	− 0.0158	0.6483	0.6363	0.947	− 0.0154	0.4567	0.4435	0.951
$\theta_{03} = -0.3$	0.0203	0.3825	0.3709	0.956	− 0.0006	0.2719	0.2598	0.946
$\theta_3 = -0.3$	− 0.0136	0.6609	0.6352	0.947	0.0001	0.4628	0.4443	0.949
$\alpha = 0.6$	− 0.0059	0.1213	0.1261	0.967	0.0034	0.0899	0.0909	0.954

to be appropriate. In addition, the estimation results became better when the size increased as expected. We also considered other setups including different values for m and obtained similar results. Table 2 gives the results obtained by the proposed approach with a discrete covariate from Bernoulli(0.5) and $\beta_{01} = \beta_2 = \beta_{03} = -0.2$, $\beta_w = 0.5, \gamma_1 = \gamma_2 = \gamma_3 = 0.5, \theta_{01} = \theta_1 = \theta_{03} = \theta_3 = -0.4$ and $\alpha = 0.6$, and they gave similar conclusions as above.

To further assess the performance of the proposed method, Table 3 provides the results on the estimation of $E_c[\text{CESP}(t|X)]$ ($C(t|X)$ for short) and $E_c[\text{CESP}(t)]$ ($C(t)$

Table 2 Simulation results with covariate distributed from Bernoulli(0.5)

Para	n=500				n=1000			
	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\beta_{01} = -0.2$	− 0.0071	0.4963	0.4672	0.965	0.0065	0.3136	0.3218	0.958
$\beta_2 = -0.2$	− 0.0111	0.3301	0.3148	0.947	− 0.0076	0.2253	0.2163	0.948
$\beta_{03} = -0.2$	− 0.0171	0.5359	0.5111	0.940	0.0149	0.3516	0.3529	0.956
$\beta_w = 0.5$	0.0052	0.1284	0.1399	0.963	− 0.0017	0.0940	0.0968	0.952
$\gamma_1 = 0.5$	0.0110	0.4290	0.4060	0.957	0.0106	0.2915	0.2801	0.946
$\gamma_2 = 0.5$	− 0.0171	0.3720	0.3394	0.932	0.0126	0.2461	0.2311	0.938
$\gamma_3 = 0.5$	0.0172	0.4334	0.4069	0.932	0.0125	0.2785	0.2824	0.948
$\theta_{01} = -0.4$	− 0.0182	0.2593	0.2601	0.955	− 0.0142	0.1805	0.1812	0.950
$\theta_1 = -0.4$	0.0365	0.3557	0.3608	0.962	0.0137	0.2534	0.2513	0.958
$\theta_{03} = -0.4$	− 0.0114	0.2530	0.2596	0.969	− 0.0089	0.1733	0.1814	0.958
$\theta_3 = -0.4$	0.0359	0.3546	0.3596	0.954	0.0136	0.2468	0.2516	0.946
$\alpha = 0.6$	0.0087	0.1231	0.1247	0.956	0.0017	0.0824	0.0862	0.952

Table 3 Second-stage estimations

		n=500				n=1000			
Setup 1	True	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$C(t_1\|X)$	0.0570	-0.0029	0.0682	0.0635	0.936	-0.0020	0.0473	0.0446	0.958
$C(t_2\|X)$	0.0852	0.0042	0.0975	0.0906	0.933	0.0015	0.0666	0.0631	0.944
$C(t_3\|X)$	0.0903	0.0062	0.1040	0.0962	0.933	0.0024	0.0699	0.0665	0.950
$C(t_4\|X)$	0.0757	0.0016	0.0969	0.0935	0.953	-0.0009	0.0632	0.0672	0.950
$C(t_1)$	0.0564	-0.0027	0.0687	0.0637	0.934	-0.0019	0.0472	0.0448	0.958
$C(t_2)$	0.0843	0.0037	0.0966	0.0894	0.932	0.0013	0.0622	0.0626	0.942
$C(t_3)$	0.0887	0.0054	0.1021	0.0943	0.931	0.0021	0.0691	0.0656	0.950
$C(t_4)$	0.0756	0.0009	0.0948	0.0915	0.956	-0.0007	0.0624	0.0661	0.948
Setup 2	True	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$C(t_1\|X)$	0.0342	-0.0027	0.0704	0.0663	0.934	-0.0019	0.0407	0.0398	0.942
$C(t_2\|X)$	0.0554	0.0013	0.0901	0.0860	0.936	0.0003	0.0593	0.0573	0.938
$C(t_3\|X)$	0.0600	0.0014	0.0880	0.0845	0.942	0.0011	0.0632	0.0608	0.944
$C(t_4\|X)$	0.0531	-0.0015	0.0772	0.0791	0.974	-0.0010	0.0587	0.0576	0.948
$C(t_1)$	0.0351	-0.0028	0.0618	0.0585	0.939	-0.0026	0.0415	0.0405	0.942
$C(t_2)$	0.0547	0.0014	0.0853	0.0814	0.941	0.0002	0.0585	0.0566	0.943
$C(t_3)$	0.0581	0.0015	0.0886	0.0849	0.942	0.0011	0.0615	0.0593	0.944
$C(t_4)$	0.0511	-0.0010	0.0816	0.0810	0.962	-0.0008	0.0569	0.0559	0.948

for short) at $t = 1, 2, 3, 4$, respectively, which represent CESP($t|X$) and CESP(t) after integrating out c, respectively. In order to check the performance of estimators easier, we also present the true value (True) of E_c[CESP($t|X$)] and E_c[CESP(t)]. The results at the top half of the table were obtained under the setup given in Table 1 with $X = 0.5$, while those at the bottom half were obtained under the setup given in Table 2 with $X = 1$. Note that at a given t, the true value of E_c[CESP(t)] was obtained via the large-sample approximation based on 100, 000 simulated data. The results suggest that as above, the proposed method seems to perform well and show the asymptotic properties of E_c[CESP($t|X$)] and E_c[CESP(t)], although we do not consider the theoretical proof in this article.

Besides, we consider the situation with bivariate covariates. We assume that $X = (X_1, X_2)^T$, where X_1 is generated from continuous uniform distribution over (0, 1) and X_2 is generated from discrete Bernoulli(0.5) distribution. Let $\beta_{01} = \beta_2 = \beta_{03} = -0.2$, $\beta_w = (\beta_{w1}, \beta_{w2})^T = (0.4, 0.4)^T$, $\gamma_1 = (\gamma_{11}, \gamma_{12})^T = (0.4, 0.4)^T$, $\gamma_2 = (\gamma_{21}, \gamma_{22})^T = (0.4, 0.4)^T$, $\gamma_3 = (\gamma_{31}, \gamma_{32})^T = (0.4, 0.4)^T$, $\theta_1 = (\theta_{01}, \theta_{11}, \theta_{12})^T = (-0.3, -0.3, -0.3)^T$, $\theta_3 = (\theta_{03}, \theta_{31}, \theta_{32})^T = (-0.3, -0.3, -0.3)^T$ and $\alpha = 0.6$. We examine the performance of the estimators under $n = 500$ and $n = 1000$, and the results are given in Table 4. We also examine the performance of E_c[CESP($t|X$)] with $X = (0.5, 1)^T$ and E_c[CESP(t)] at $t = 1, 2, 3, 4$, and we present the results in Table 5. One can see from Tables 4 and 5 that the results are similar to those obtained above and again indicate that the proposed method performs well.

Table 4 Simulation results with binary covariates

	n=500				n=1000			
Para	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\beta_{01} = -0.2$	− 0.0144	0.7360	0.6874	0.928	0.0056	0.5179	0.4746	0.933
$\beta_2 = -0.2$	− 0.0121	0.3327	0.3037	0.937	− 0.0003	0.2222	0.2117	0.938
$\beta_{03} = -0.2$	− 0.0133	0.7610	0.7168	0.929	0.0093	0.5087	0.4950	0.951
$\beta_{w1} = 0.4$	− 0.0092	0.2244	0.2397	0.968	0.0057	0.1586	0.1663	0.964
$\beta_{w2} = 0.4$	− 0.0115	0.1329	0.1390	0.953	− 0.0078	0.0933	0.0967	0.952
$\gamma_{11} = 0.4$	0.0196	0.7452	0.6937	0.941	− 0.0094	0.4998	0.4733	0.944
$\gamma_{12} = 0.4$	0.0114	0.4351	0.3979	0.936	− 0.0091	0.2881	0.2725	0.942
$\gamma_{21} = 0.4$	− 0.0284	0.5847	0.5417	0.941	0.0154	0.4002	0.3753	0.943
$\gamma_{22} = 0.4$	0.0115	0.3520	0.3213	0.946	0.0104	0.2368	0.2220	0.948
$\gamma_{31} = 0.4$	0.0178	0.7164	0.6947	0.942	− 0.0111	0.4929	0.4738	0.946
$\gamma_{32} = 0.4$	− 0.0104	0.4578	0.4015	0.927	− 0.0085	0.2858	0.2742	0.940
$\theta_{01} = -0.3$	0.0193	0.4073	0.4120	0.965	0.0088	0.2812	0.2862	0.963
$\theta_{11} = -0.3$	− 0.0267	0.6367	0.6296	0.941	− 0.0029	0.4077	0.4365	0.957
$\theta_{12} = -0.3$	− 0.0219	0.3691	0.3618	0.945	− 0.0113	0.2609	0.2518	0.948
$\theta_{03} = -0.3$	0.0016	0.4207	0.4125	0.956	0.0006	0.2831	0.2861	0.949
$\theta_{31} = -0.3$	− 0.0109	0.6513	0.6281	0.953	0.0021	0.4406	0.4359	0.950
$\theta_{32} = -0.3$	− 0.0094	0.3766	0.3617	0.944	0.0002	0.2642	0.2517	0.946
$\alpha = 0.6$	− 0.0129	0.1181	0.1206	0.966	0.0019	0.0828	0.0870	0.964

Table 5 Second-stage estimations with binary covariates

		n=600				n=1200			
	True	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$C(t_1\|X)$	0.0446	− 0.0024	0.0734	0.0679	0.935	− 0.0017	0.0507	0.0479	0.948
$C(t_2\|X)$	0.0606	0.0018	0.0916	0.0841	0.931	− 0.0010	0.0621	0.0591	0.935
$C(t_3\|X)$	0.0575	0.0016	0.0875	0.0800	0.936	− 0.0012	0.0580	0.0557	0.939
$C(t_4\|X)$	0.0453	− 0.0019	0.0757	0.0746	0.968	− 0.0012	0.0476	0.0547	0.961
$C(t_1)$	0.0397	− 0.0029	0.0670	0.0615	0.935	− 0.0023	0.0456	0.0434	0.939
$C(t_2)$	0.0570	0.0007	0.0882	0.0806	0.931	− 0.0006	0.0597	0.0569	0.934
$C(t_3)$	0.0575	0.0011	0.0881	0.0807	0.933	− 0.0007	0.0582	0.0567	0.936
$C(t_4)$	0.0484	− 0.0017	0.0790	0.0761	0.963	− 0.0013	0.0523	0.0552	0.959

The intention-to-treat (ITT) principle also can used to estimate the treatment effect, but when patient non-compliance occurs, the ITT estimator measures the effect of the treatment assignment a under the hazard function model $\lambda_{T^a} = \lambda(t)\exp(\beta a + \gamma^T X)c$, rather than the true treatment efficacy as a result of ignoring the existence of non-compliance. Simple alternatives to the ITT analysis such as the as-treated analysis (AT), which only considers the final treatment d under the model $\lambda_{T^d} = \lambda(t)\exp(\beta d + \gamma^T X)c$ simply, and the per-protocol

analysis (PP), which only uses the information from those compliant subjects under the model $\lambda_{T^{d=a}} = \lambda(t)\exp(\beta d + \gamma^T X)c$, could also be biased due to potential confounding by measured or unmeasured factors. As discussed above, the use of these native approaches would yield invalid results for the situation considered here. In other words, it may be of practical interest to compare the proposed method to these naive methods. Some results are presented in Table 6 in which we calculated the empirical biases (Bias) and the empirical standard errors (SE) among the compliers given by the proposed method and the three naive methods under the two settings with $n = 500$, respectively. It is apparent that unlike the proposed method, all of the three naive methods gave biased estimates of the treatment effects. In other words, they can fail to measure the true causal treatment effect in the presence of treatment non-compliance.

Note that in the above, we have assumed that the latent variables c_i's follow the gamma distribution with the unknown variance parameter α. To assess the fault tolerance of the proposed estimation procedure with respect to this, we also considered two other settings for the situation where X follows Uniform(0,1) and Bernoulli(0.5), respectively, with $n = 500$. First, we generated the c_i's from the log-normal distribution with mean 1 and variance 0.5 but assumed that they followed the gamma distribution. Let $\beta_{01} = \beta_2 = \beta_{03} = 0.3$, $\beta_w = 0.3$, $\gamma_1 = \gamma_2 = \gamma_3 = 0.3$, $\theta_{01} = \theta_1 = \theta_{03} = \theta_3 = -0.4$ and $\alpha = 0.5$. In the second setting, we considered a more complex situation that generated c_i's from the gamma distribution with mean 1 and variance 0.5 or log-normal distribution with mean 1 and variance 1 randomly under the same probability but assumed that they followed the gamma distribution. Let $\beta_{01} = \beta_2 = \beta_{03} = 0.5$, $\beta_w = 0.5$, $\gamma_1 = \gamma_2 = \gamma_3 = 0.5$, $\theta_{01} = \theta_1 = \theta_{03} = \theta_3 = -0.5$ and $\alpha = 0.5$. The results obtained on the proposed estimators except α are presented in Table 7, including the Bias, SE, SSE and CP. From Table 7, one can see that the proposed approach seems to perform well as before and thus can deal with the misspecified distribution of the independent latent variable.

Table 6 Comparison with other native methods

	Proposed method		ITT		AT		PP	
Para	Bias	SE	Bias	SE	Bias	SE	Bias	SE
$\beta_2 = -0.3$	− 0.0108	0.3658	0.1581	0.1538	0.1732	0.1502	0.1161	0.1786
$\beta_w = 0.5$	− 0.0022	0.1214	− 0.0035	0.1215	− 0.0054	0.1214	− 0.0093	0.1465
$\gamma_2 = 0.5$	0.0116	0.3806	0.0195	0.1622	0.0195	0.1624	0.0257	0.1875
$r = 0.5$	0.0062	0.1078	− 0.0101	0.1092	− 0.0106	0.1053	0.0077	0.1212
$\beta_2 = 0.2$	0.0103	0.4457	− 0.1132	0.1516	− 0.1209	0.1497	− 0.0798	0.1829
$\beta_w = 0.2$	0.0140	0.1232	0.0153	0.1225	0.0150	0.1223	0.0172	0.1499
$\gamma_2 = 0.2$	− 0.0047	0.4713	− 0.0156	0.1483	− 0.0143	0.1486	− 0.0192	0.1810
$r = 0.5$	0.0191	0.0926	0.0253	0.0888	0.0245	0.0888	0.0160	0.1158

Table 7 Simulation results with misspecified frailty distribution

	$X \sim U(0,1)$				$X \sim Bernoulli(0.5)$			
Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\beta_{01} = 0.3$	0.0247	0.5802	0.5431	0.936	0.0138	0.4196	0.3877	0.948
$\beta_2 = 0.3$	0.0163	0.2869	0.2718	0.958	0.0204	0.2756	0.2694	0.944
$\beta_{03} = 0.3$	0.0218	0.5707	0.5460	0.946	0.0094	0.4488	0.4205	0.938
$\beta_w = 0.3$	− 0.0096	0.1908	0.1955	0.964	0.0071	0.1072	0.1127	0.968
$\gamma_1 = 0.3$	0.0074	0.6214	0.5641	0.934	0.0192	0.3644	0.3225	0.940
$\gamma_2 = 0.3$	0.0267	0.5361	0.4839	0.948	− 0.0104	0.3183	0.2846	0.944
$\gamma_3 = 0.3$	− 0.0136	0.6083	0.5648	0.930	0.0289	0.3332	0.3225	0.932
$\theta_{01} = -0.4$	− 0.0398	0.3269	0.3629	0.968	− 0.0029	0.2370	0.2597	0.960
$\theta_1 = -0.4$	− 0.0351	0.5763	0.6230	0.962	0.0228	0.3408	0.3595	0.959
$\theta_{03} = -0.4$	0.0219	0.3570	0.3636	0.956	− 0.0072	0.2455	0.2593	0.956
$\theta_3 = -0.4$	− 0.0378	0.5994	0.6243	0.960	0.0147	0.3472	0.3599	0.958
$\beta_{01} = 0.5$	0.0233	0.5787	0.5439	0.934	0.0351	0.4285	0.3993	0.932
$\beta_2 = 0.5$	0.0352	0.2779	0.2601	0.940	0.0234	0.2873	0.2695	0.933
$\beta_{03} = 0.5$	0.0380	0.5797	0.5458	0.936	0.0399	0.4642	0.4271	0.936
$\beta_w = 0.5$	0.0162	0.2072	0.2209	0.967	0.0072	0.1253	0.1268	0.962
$\gamma_1 = 0.5$	0.0317	0.6567	0.6141	0.937	0.0157	0.3793	0.3548	0.940
$\gamma_2 = 0.5$	0.0148	0.5049	0.4837	0.933	0.0178	0.3142	0.2785	0.937
$\gamma_3 = 0.5$	0.0422	0.6367	0.6020	0.935	0.0382	0.3688	0.3446	0.932
$\theta_{01} = -0.5$	0.0133	0.3676	0.3591	0.944	0.0010	0.2508	0.2580	0.961
$\theta_1 = -0.5$	− 0.0272	0.6172	0.6187	0.953	− 0.0081	0.3591	0.3569	0.958
$\theta_{03} = -0.5$	0.0011	0.3585	0.3571	0.951	0.0071	0.2569	0.2542	0.952
$\theta_3 = -0.5$	− 0.0039	0.6043	0.6148	0.949	− 0.0126	0.3527	0.3549	0.959

5 An application

In this section, we apply the methodology proposed in the previous sections to the HVTN 505 Trial described above. To test the efficacy of the DNA prime-recombinant adenovirus type 5 boost (DNA/rAd5) vaccine, the HVTN 505 Trial randomly assigned 2504 men or transgender women who have sex with men to receive the DNA/rAd5 vaccine (1253 participants) or placebo (1251 participants) (Hammer et al. 2013; Janes et al. 2017; Youyi et al. 2018). Participants were enrolled at 21 sites in the USA and provided written informed consent. For each subject, four demographic covariates were observed and they are age, race, BMI and behavioral risk.

The primary efficacy end points were HIV infections diagnosed after week 28 following enrollment through the 24-month study visit. For the analysis here, by following Du et al. (2021), failure time of interest here is the time to true HIV-1 infection and for which only interval-censored data with informative censoring are available. For the instrumental variable, we considered the percentage of active treatment for each site every year, which was strongly associated with treatment assignment,

and categorized this instrument into a binary variable according to its median. The IV value of 1 represents a high possibility (above median) of giving vaccine treatment and 0 represents a low possibility (below the median). This proportion has also been used as an IV in Huling et al. (2019), O'Malley et al. (2011) and others.

Table 8 shows the estimated complier causal treatment effect as well as other covariates effects given by the proposed method (EST) along with the estimated standard errors (SE) and the p values for testing no treatment effect. In particular, we have that $\hat{\beta}_{2n} = -0.1366$ with the $p = 0.8585$, which showed lack of vaccine efficacy for the compliers, and the conclusion is same as the results given by Hammer et al. (2013) and others. In addition, the results in Table 8 indicate that the behavior risk seems to be correlated with the HIV infection time, while age, race and BMI did not seem to have any effects on the HIV infection. For comparison, we also applied the three naive methods discussed in the simulation study and include the results in Table 8 and they gave the same results, except that all of the three methods gave smaller estimated complier causal treatment effect. Here for the degree of Bernstein polynomials, we tried several values, including $m = 5, 6, 7, 8$. They gave similar conclusions, and the results given below were obtained based on $m = 7$.

To further see the complier causal difference in the survival probability between the vaccine treatment group versus the usual care group, Table 9 presents the estimates of $E_c[\text{CESP}(t|X)]$ and $E_c[\text{CESP}(t)]$ given by the proposed method along with the estimated standard errors and the p values at $t = 2, 4, 6, 8$. The elements in the vector X were set to be the mean value for continuous covariates and 1 for the binary covariates. The results again suggest that the DNA/rAd5 vaccine

Table 8 Results for the analysis of the HVTN 505 Trial

Covariate	Proposed method			ITT		
	EST	SE	p value	EST	SE	p value
Treatment	− 0.1366	0.7658	0.8585	− 0.1308	0.1858	0.4814
Age	− 0.8244	0.7667	0.2823	− 0.0176	0.0129	0.1710
Race	− 1.5436	0.9495	0.1040	− 0.3697	0.3584	0.3023
BMI	− 0.9968	1.1272	0.3765	− 0.4522	0.2793	0.1054
Behavioral risk	3.4679	1.4211	0.0147	1.5700	0.2636	< 0.0001
r	0.0026	0.7025	0.9970	0.0041	0.0292	0.8882
Covariate	AT			PP		
	EST	SE	p value	EST	SE	p value
Treatment	− 0.0738	0.1825	0.6861	− 0.0832	0.2419	0.7309
Age	− 0.1807	0.1308	0.1673	− 0.1942	0.1899	0.3063
Race	− 0.3615	0.3569	0.3112	− 0.5311	0.5692	0.3508
BMI	− 0.4577	0.2794	0.1014	− 0.5202	0.2972	0.0800
Behavioral risk	1.5648	0.2628	< 0.0001	1.8284	0.4048	< 0.0001
r	0.0002	0.0294	0.9934	0.0010	0.0307	0.9740

Table 9 More results on the analysis for the HVTN 505 Trial

t	$\widehat{E_c[\text{CESP}](t)}$			$\widehat{E_c[\text{CESP}](t\|X)}$		
	EST	SE	p value	EST	SE	p value
1	0.0001	0.0006	0.8600	0.0004	0.0025	0.8602
2	0.0003	0.0016	0.8603	0.0011	0.0060	0.8603
3	0.0005	0.0031	0.8602	0.0019	0.0106	0.8601
4	0.0009	0.0053	0.8602	0.0029	0.0167	0.8602

regimen did not reduce the rate of HIV-1 acquisition in the population studied. These again are consistent with the existing conclusions.

6 Discussion and concluding remarks

This paper discussed how to estimate the causal treatment effect for compliers when only interval-censored data with informative censoring are available for the time-to-event outcome of interest. In the proposed method, we introduced the instrumental variables to deal with the problems caused by the unmeasured confounders, and the latent variable approach was employed to describe the relationship between the failure time of interest and the observation process. The Bernstein polynomials were employed, and a sieve maximum likelihood estimation approach was developed. The proposed estimators were shown to be consistent and asymptotically normally distributed, and they work well in practical situations.

Note that although the focus above has been on the proportional hazards model for the description of the complier causal treatment effect, similar methods could be developed for other models such as the additive hazard model and accelerated failure time model. But the computation and theoretical justification may be different. Also instead of the latent variable method used above to deal with informative censoring, one may try other method such as the copula model approach (Sun 2006). For example, one can directly model the joint distribution of the failure time of interest and censoring variables. However, the limitation is that one usually needs to assume that the underlying copula function is known.

Also, note that in the above, we have assumed all-or-nothing compliance, while in longitudinal studies, a subject may not follow the assigned treatment over some periods and thus it is meaningful to generalize the proposed method to partial compliance. Furthermore, we only considered binary instrumental variables and two arms treatment without changing but one may face more complicated situations in practice. For example, such case could occur when treatments are allocated sequentially or depend possibly on previous treatments and covariate history as well as previous outcomes. In addition, it has been assumed that the latent variable is independent of covariates and in reality, it is apparent that this may not be true and requires the development of a new method.

Acknowledgements The authors wish to thank the Associate Editor and two reviewers for their many helpful comments and suggestions that greatly improved the paper. This work was partially supported by the Natural Science Foundation of Jilin Province (Grant No. 20230101002JC) and the National Nature Science Foundation of China (Grant No. 11801212, Grant No. 12071176).

Appendix: Proofs of Theorems 1–3

In this Appendix, we will sketch the proofs for the asymptotic properties of $\hat{\xi}_n$ given in Theorems 1–3. To establish the asymptotic properties of $\hat{\xi}_n$, we need the following regularity conditions, which are commonly used in the studies of interval-censored data and usually satisfied in practice (Huang and Wolfe 2002; Ma et al. 2016; Zhang et al. 2010; Zhou et al. 2016).

(A.1) The true value for η, denoted as η_0, is in the interior of a compact set $\mathcal{B}$ in R^{p_η}, $\|\eta_0\| \le B$ for a constant $B > 0$, and $P(R - L > \varepsilon) = 1$ for some $\varepsilon > 0$.

(A.2) The distribution of the covariate X has a bounded support in R^p and is not concentrated on any proper subspace of R^p.

(A.3) The first derivative of $\Lambda_{t0}(\cdot)$ and $\Lambda_{w0}(\cdot)$, denoted by $\Lambda_{t0}^{(1)}(\cdot)$ and $\Lambda_{w0}^{(1)}(\cdot)$, is Holder continuous with exponent $s \in (0, 1]$. That is, there exists a constant $K > 0$ such that $|\Lambda_{t0}^{(1)}(t_1) - \Lambda_{t0}^{(1)}(t_2)| \le K|t_1 - t_2|^s$ for all $t_1, t_2 \in [\sigma_1, \tau_1]$, where $0 < \sigma_1 < \tau_1 < \infty$, and $\Lambda_{w0}(\cdot)$ has the similar properties. Let $v = 1 + s$.

(A.4) There exists a constant $K > 0$ such that for every ξ in a neighborhood of ξ_0, $P\{l(\xi, O) - l(\xi_0, O)\} \preceq -Kd^2(\xi, \xi_0)$, where O is the observation data and $\preceq$ means 'smaller than, up to a constant.'

(A.5) The matrix $E(l^*(\eta_0, O)^{\otimes 2})$ is finite and positive definite, where $a^{\otimes 2} = aa^T$ for a vector a, and $l^*(\eta, O)$ is the efficient score for η based on the observation O and given in the proof of Theorem 3.

For the proof, we will mainly employ the empirical process theory and some nonparametric techniques. Let $Pf = \int f(y)dP$ denote the expectation of $f(Y)$ under the probability measure P, and $P_n f = n^{-1}\sum_{i=1}^n f(Y_i)$, the expectation of $f(Y)$ under the empirical measure P_n. Define the covering number of the class $\mathcal{L}_n = \{l(\xi, O) : \xi \in \Psi_n\}$, where $l(\xi, O)$ is the log-likelihood function based on a single observation O. Also for any $\epsilon > 0$, define the covering number $N(\epsilon, \mathcal{L}_n, L_1(P_n))$ as the smallest positive integer κ for which there exists $\{\xi^{(1)}, \ldots, \xi^{(\kappa)}\}$ such that

$$\min_{j \in \{1,\ldots,\kappa\}} \frac{1}{n}\sum_{i=1}^n |l(\xi, O_i) - l(\xi^{(j)}, O_i)| < \epsilon$$

for all $\xi \in \Psi_n$, where $\{O_1, \ldots, O_n\}$ represent the observed data and for $j = 1, \ldots, \kappa$, $\xi^{(j)} = (\eta^{(j)}, \Lambda_t^{(j)}, \Lambda_w^{(j)}) \in \Psi_n$. If no such κ exists, define $N(\epsilon, \mathcal{L}_n, L_1(P_n)) = \infty$. Also for the proof, we need first to establish the following two lemmas, whose proofs are similar to those for Lemmas 1 and 2 in Zhou et al. (2016).

Lemma 1 *Assume that the regularity conditions* (A.1)-(A.3) *given above hold. Then, we have that the covering number of the class* $\mathcal{L}_n = \{l(\xi, O) : \xi \in \Psi_n\}$ *satisfies*

$$N\Big(\epsilon, \mathcal{L}_n, L_1(P_n)\Big) \leq KM_n^{2(m+1)}\epsilon^{-(p_\eta+2(m+1))}$$

for a constant K, *where* $m = o(n^q)$ *with* $q \in (0, 1)$ *is the degree of Bernstein polynomials, and* $M_n = O(n^a)$ *with* $a > 0$ *controls the size of the sieve space* Ψ_n.

Lemma 2 *Assume that the regularity conditions* (A.1)–(A.3) *given above hold. Then, we have that*

$$\sup_{\xi\in\Psi_n} |P_n l(\xi, O) - Pl(\xi, O)| \to 0$$

almost surely.

Now, we are ready to prove Theorems 1–3.

Proof of Theorem 1 We first prove the strong consistency of $\hat{\xi}_n$. Let $l(\xi, O)$ denote the log-likelihood function based on a given single observation O and consider the class of functions $\mathcal{L}_n = \{l(\xi, O) : \xi \in \Psi_n\}$. By Lemma 1, the covering number of $\mathcal{L}_n$ satisfies

$$N\Big(\epsilon, \mathcal{L}_n, L_1(P_n)\Big) \leq KM_n^{2(m+1)}\epsilon^{-(p_\eta+2m+2)}.$$

Furthermore, by Lemma 2, we have

$$\sup_{\xi\in\Psi_n} |P_n l(\xi, O) - Pl(\xi, O)| \to 0 \text{ almost surely.} \tag{5}$$

Note that $Pl(\xi, O) = P\{pl(\xi, O)\} = Pl(\xi, O)$ and ξ_0 maximizes $Pl(\xi, O)$. Let $M(\xi, O) = -l(\xi, O)$, and define $K_\epsilon = \{\xi : d(\xi, \xi_0) \geq \epsilon, \xi \in \Psi_n\}$ for $\epsilon > 0$ and

$$\zeta_{1n} = \sup_{\xi\in\Psi_n} |P_n M(\xi, O) - PM(\xi, O)|, \zeta_{2n} = P_n M(\xi_0, O) - PM(\xi_0, O).$$

Then,

$$\begin{aligned} \inf_{K_\epsilon} PM(\xi, O) &= \inf_{K_\epsilon} \Big\{ PM(\xi, O) - P_n M(\xi, O) + P_n M(\xi, O) \Big\} \\ &\leq \zeta_{1n} + \inf_{K_\epsilon} P_n M(\xi, O). \end{aligned} \tag{6}$$

If $\hat{\xi}_n \in K_\epsilon$, then we have

$$\inf_{K_\epsilon} P_n M(\xi, O) = P_n M(\hat{\xi}_n, O) \leq P_n M(\xi_0, O) = \zeta_{2n} + PM(\xi_0, O). \tag{7}$$

Define $\delta_\epsilon = \inf_{K_\epsilon} PM(\xi, O) - PM(\xi_0, O)$. Under Condition (A.4), we have $\delta_\epsilon > 0$. It follows from (6) and (7) that

$$\inf_{K_\epsilon} PM(\xi, O) \le \zeta_{1n} + \zeta_{2n} + PM(\xi_0, O) = \zeta_n + PM(\xi_0, O),$$

with $\zeta_n = \zeta_{1n} + \zeta_{2n}$, and hence, $\zeta_n \ge \delta_\epsilon$. This gives $\{\hat{\xi}_n \in K_\epsilon\} \subseteq \{\zeta_n \ge \delta_\epsilon\}$, and by (5) and the strong law of large numbers, we have both $\zeta_{1n} \to 0$ and $\zeta_{2n} \to 0$ almost surely. Therefore, $\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{\hat{\xi}_n \in K_\epsilon\} \subseteq \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{\zeta_n \ge \delta_\epsilon\}$, which proves that $d(\hat{\xi}_n, \xi_0) \to 0$ almost surely. □

Proof of Theorem 2 We will show the convergence rate of $\hat{\xi}_n$ by using Theorem 3.4.1 of Van and Wellner (1996). First note from Theorem 1.6.2 of Lorentz (1986) that exists Bernstein polynomials Λ_{tn0} and Λ_{wn0} such that $\|\Lambda_{tn0} - \Lambda_{t0}\|_\infty = O(m^{-v/2})$ and $\|\Lambda_{wn0} - \Lambda_{w0}\|_\infty = O(m^{-v/2})$. Then, we have $d(\xi_{n0} - \xi_0) = O(n^{-vq/2})$. For any $s > 0$, define the class of functions $\mathcal{F}_s = \{l(\xi, O) - l(\xi_{n0}, O) : \xi \in \Psi_n, s/2 < d(\xi - \xi_{n0}) \le s\}$. One can easily show that $P\{l(\xi_0, O) - l(\xi_{n0}, O)\} \le Kd^2(\xi_0, \xi_{n0}) \le Kn^{-vq}$. Hence, under Condition (A.4), we have for large n, $P\{l(\xi, O) - l(\xi_{n0}, O)\} = P\{l(\xi, O) - l(\xi_0, O)\}$ $+P\{l(\xi_0, O) - l(\xi_{n0}, O)\} \le -Ks^2 + Kn^{-vq} = -Ks^2$, for any $l(\xi, O) - l(\xi_{n0}, O) \in \mathcal{F}_s$.

Following the calculations in Shen and Wong (1994), we can establish that for $0 < \epsilon < s$, $\log N_{[]}(\epsilon, \mathcal{F}_s, L_2(P)) \le KN \log(s/\epsilon)$ with $N = 2(m+1)$. Moreover, some algebraic manipulations yield that $P\{(l(\xi, O) - l(\xi_{n0}, O)\}^2 \le Ks^2$ for any $l(\xi, O) - l(\xi_{n0}, O) \in \mathcal{F}_s$. It is easy to see that $\mathcal{F}_s$ is uniformly bounded. Therefore, by Lemma 3.4.2 of Van and Wellner (1996), we obtain

$$E_P \left\| n^{1/2}(P_n - P) \right\|_{\mathcal{F}_s} \le KJ_{[]}(s, \mathcal{F}_s, L_2(P)) \left\{ 1 + \frac{J_{[]}(s, \mathcal{F}_s, L_2(P))}{s^2 n^{1/2}} \right\},$$

where $J_{[]}(s, \mathcal{F}_s, L_2(P)) = \int_0^s \{1 + \log N_{[]}(\epsilon, \mathcal{F}_s, L_2(P))\}^{1/2} d\epsilon \le KN^{1/2}s$. This yields $\phi_n(s) = N^{1/2}s + N/n^{1/2}$. It is easy to see that $\phi_n(s)/s$ is decreasing in s, and $v_n^2 \phi_n(1/v_n) = v_n N^{1/2} + v_n^2 N/n^{1/2} \le Kn^{1/2}$, where $v_n = N^{-1/2} n^{1/2} = n^{(1-q)/2}$.

Finally, note that $P_n\{l(\hat{\xi}_n, O) - l(\xi_{n0}, O)\} \ge 0$ and $d(\hat{\xi}_n, \xi_{n0}) \le d(\hat{\xi}_n, \xi_0) + d(\xi_0, \xi_{n0}) \to 0$ in probability. Thus by applying Theorem 3.4.1 of Van and Wellner (1996), we have $n^{(1-q)/2} d(\hat{\xi}_n, \xi_{n0}) = O_p(1)$. This together with $d(\xi_{n0}, \xi_0) = O(n^{-vq/2})$ yields that $d(\hat{\xi}_n, \xi_0) = O_p(n^{-(1-q)/2} + n^{-vq/2})$ and the proof is completed. □

Proof of Theorem 3 Now, we will prove the asymptotic normality of $\hat{\eta}_n$. Let V denote the linear span of $\Psi - \xi_0$ and define the Fisher inner product for $u, \tilde{u} \in V$ as $< u, \tilde{u} >= P\{\dot{l}(\xi_0, O)[u]\dot{l}(\xi_0, O)[\tilde{u}]\}$ and the Fisher norm for $u \in V$ as $||u||^2 =< u, u >$, where

$$\dot{l}(\xi_0, O)[u] = \frac{dl(\xi_0 + su, O)}{ds}\Big|_{s=0}$$

denotes the first-order directional derivative of $l(\xi, O)$ at the direction $u \in V$ (evaluated at ξ_0). Also, let $\bar{V}$ be the closed linear span of V under the Fisher norm. Then, $(\bar{V}, ||\cdot||)$ is a Hilbert space. Furthermore, for a vector of p_η dimension b with $||b|| \leq 1$ and for any $u \in V$, define a smooth function of ξ as $h(\xi) = b^T\eta$ and

$$\dot{h}(\xi_0)[u] = \frac{dh(\xi_0 + su)}{ds}\Big|_{s=0}$$

whenever the right-hand side limit is well defined. Then by the Riesz representation theorem, there exists $u^* \in \bar{V}$ such that $\dot{h}(\xi_0)[u] =< u, u^* >$ for all $u \in \bar{V}$ and $||u^*|| = ||\dot{h}(\xi_0)||$. Also, note that $h(\xi) - h(\xi_0) = \dot{h}(\xi_0)[\xi - \xi_0]$. It thus follows from the Cramér–Wold device that to prove the asymptotic normality for $\hat{\eta}_n$, i.e., $n^{1/2}(\hat{\eta}_n - \eta_0) \to N(0, I^{-1}(\eta_0))$ in distribution, it suffices to show that

$$n^{1/2} < \hat{\xi}_n - \xi_0, u^* > \to_d N(0, b^T I^{-1}(\eta_0) b)$$

since $b^T(\hat{\eta}_n - \eta_0) = h(\hat{\xi}_n) - h(\xi_0) = \dot{h}(\xi_0)[\xi - \xi_0] =< \hat{\xi}_n - \xi_0, u^* >$. In fact, the above holds since one can show that $n^{1/2} < \hat{\xi}_n - \xi_0, u^* > \to_d N(0, ||u^*||^2)$ and $||u^*||^2 = b^T I^{-1}(\eta_0) b$.

We first prove that $n^{1/2} < \hat{\xi}_n - \xi_0, u^* > \to_d N(0, ||u^*||^2)$. Let $\delta_n = n^{-\min\{(1-q)/2, vq/2\}}$ denote the rate of convergence obtained in Theorem 2, and for any $\xi \in \Psi$ such that $d(\xi, \xi_0) \leq \delta_n$, define the first-order directional derivative of $l(\xi, O)$ at the direction $u \in V$ as

$$\dot{l}(\xi, O)[u] = \frac{dl(\xi + su, O)}{ds}\Big|_{s=0}$$

and the second-order directional derivative at the direction $u, \tilde{u} \in V$ as

$$\ddot{l}(\xi, O)[u, \tilde{u}] = \frac{d^2 l(\xi + su + \tilde{s}\tilde{u}, O)}{d\tilde{s} ds}\Big|_{s=0}\Big|_{\tilde{s}=0} = \frac{d\dot{l}(\xi + \tilde{s}\tilde{u}, O)[\tilde{u}]}{d\tilde{s}}\Big|_{\tilde{s}=0}.$$

Note that by Condition (A.3) and Theorem 1.6.2 of Lorentz (1986), there exists $\Pi_n u^* \in \Psi - \xi_0$ such that $||\Pi_n u^* - u^*|| = O(n^{-qv})$. Furthermore, under the assumption $q > 1/2v$, we have $\delta_n ||\Pi_n u^* - u^*|| = o(n^{-1/2})$. Define $v[\xi - \xi_0, O] = l(\xi, O) - l(\xi_0, O) - \dot{l}(\xi_0, O)[\xi - \xi_0]$ and let ε_n be any positive sequence satisfying $\varepsilon_n = o(n^{-1/2})$. Then by the definition of $\hat{\xi}_n$, we have

$$\begin{aligned}
0 \leq & P_n\big[l\big(\hat{\xi}_n, O\big) - l\big(\hat{\xi}_n \pm \varepsilon_n \Pi_n u^*, O\big)\big] \\
= & \pm \varepsilon_n P_n \dot{l}\big(\xi_0, O\big)\big[\Pi_n u^*\big] + \big(P_n - P\big)\big\{v\big[\hat{\xi}_n - \xi_0, O\big] - v\big[\hat{\xi}_n \pm \varepsilon_n \Pi_n u^* - \xi_0, O\big]\big\} \\
& + P\big\{v\big[\hat{\xi}_n - \xi_0, O\big] - v\big[\hat{\xi}_n \pm \varepsilon_n \Pi_n u^* - \xi_0, O\big]\big\} \\
= & \pm \varepsilon_n P_n \dot{l}\big(\xi_0, O\big)\big[u^*\big] \pm \varepsilon_n P_n \dot{l}\big(\xi_0, O\big)\big[\Pi_n u^* - u^*\big] + \big(P_n - P\big)\big\{v\big[\hat{\xi}_n - \xi_0, O\big] \\
& - r\big[\hat{\xi}_n \pm \varepsilon_n \Pi_n u^* - \xi_0, O\big]\big\} + P\big\{v\big[\hat{\xi}_n - \xi_0, O\big] - v\big[\hat{\xi}_n \pm \varepsilon_n \Pi_n u^* - \xi_0, O\big]\big\} \\
= & \pm \varepsilon_n P_n \dot{l}\big(\xi_0, O\big)\big[u^*\big] \pm I_1 + I_2 + I_3.
\end{aligned}$$

We will investigate the asymptotic behavior of I_1, I_2 and I_3. For I_1, it follows from Conditions (A.1)-(A.3), Chebyshev inequality and $||\Pi_n u^* - u^*|| = o(1)$ that $I_1 = \varepsilon_n \times o_p(n^{-1/2})$. For I_2, by the mean value theorem, we obtain that

$$\begin{aligned} I_2 &= (P_n - P)\big\{l(\hat{\xi}_n, O) - l(\hat{\xi}_n \pm \varepsilon_n \Pi_n u^*, O) \pm \Pi\varepsilon_n \dot{l}(\xi_0, O)[\Pi_n u^*]\big\} \\ &= \pm\varepsilon_n (P_n - P)\big\{(\dot{l}(\tilde{\xi}, O) - \dot{l}(\tilde{\xi}_0, O))[\Pi_n u^*]\big\}, \end{aligned}$$

where $\tilde{\xi}$ lies between $\hat{\xi}_n$ and $\hat{\xi}_n \pm \varepsilon_n \Pi_n u^*$. By Theorem 2.8.3 of Van and Wellner (1996), we know that $\{\dot{l}(\xi, O)[\Pi_n u^*] : ||\xi - \xi_0|| \le \delta_n\}$ is Donsker class. Therefore, by Theorem 2.11.23 of Van and Wellner (1996), we have $I_2 = \varepsilon_n \times o_p(n^{-1/2})$. For I_3, note that

$$\begin{aligned} P\big(v[\xi - \xi_0, O]\big) =& P\big\{l(\xi, O) - l\big(\xi_0, O\big) - \dot{l}\big(\xi_0, O\big)\big[\xi - \xi_0\big]\big\} \\ =& 2^{-1} P\big\{\ddot{l}(\tilde{\xi}, O)\big[\xi - \xi_0, \xi - \xi_0\big] - \ddot{l}\big(\xi_0, O\big)\big[\xi - \xi_0, \xi - \xi_0\big]\big\} \\ &+ 2^{-1} P\big\{\ddot{l}\big(\xi_0, O\big)\big[\xi - \xi_0, \xi - \xi_0\big]\big\} \\ =& 2^{-1} P\big\{\ddot{l}\big(\xi_0, O\big)\big[\xi - \xi_0, \xi - \xi_0\big]\big\} + \varepsilon_n \times o_p\big(n^{-1/2}\big), \end{aligned}$$

where $\tilde{\xi}$ lies between ξ_0 and ξ and the last equation follows from Taylor expansion and Conditions (A.1)-(A.3). Therefore,

$$\begin{aligned} I_3 &= -2^{-1}\left\{\left\|\hat{\xi}_n - \xi_0\right\|^2 - \left\|\hat{\xi}_n \pm \varepsilon_n \Pi_n u^* - \xi_0\right\|^2\right\} + \varepsilon_n \times o_p\big(n^{-1/2}\big) \\ &= \pm\varepsilon_n < \hat{\xi}_n - \xi_0, \Pi_n u^* > + 2^{-1}\left\|\varepsilon_n \Pi_n u^*\right\|^2 + \varepsilon_n \times o_p\big(n^{-1/2}\big) \\ &= \pm\varepsilon_n < \hat{\xi}_n - \xi_0, u^* > + 2^{-1}\left\|\varepsilon_n \Pi_n u^*\right\|^2 + \varepsilon_n \times o_p\big(n^{-1/2}\big) \\ &= \pm\varepsilon_n < \hat{\xi}_n - \xi_0, u^* > + \varepsilon_n \times o_p\big(n^{-1/2}\big), \end{aligned}$$

where the last equality holds due to the facts $\delta_n ||\Pi_n u^* - u^*|| = o(n^{-1/2})$, Cauchy–Schwartz inequality and $||\Pi_n u^*||^2 \to ||u^*||^2$. Combining the above facts, together with $P\dot{l}(\xi_0, O)[u^*] = 0$, we can establish that

$$\begin{aligned} 0 &\le P_n\big\{l\big(\hat{\xi}_n, O\big) - l\big(\hat{\xi}_n \pm \varepsilon_n \Pi_n u^*, O\big)\big\} \\ &= \pm\varepsilon_n P_n \dot{l}\big(\xi_0, O\big)\big[u^*\big] \pm \varepsilon_n < \hat{\xi}_n - \xi_0, u^* > + \varepsilon_n \times o_p\big(n^{-1/2}\big) \\ &= \pm\varepsilon_n \big(P_n - P\big)\{\dot{l}\big(\xi_0, O\big)\big[u^*\big]\} \pm \varepsilon_n < \hat{\xi}_n - \xi_0, u^* > + \varepsilon_n \times o_p\big(n^{-1/2}\big). \end{aligned}$$

Therefore, we obtain $\pm n^{1/2}(P_n - P)\{\dot{l}(\xi_0, O)[u^*]\} \pm n^{1/2} < \hat{\xi}_n - \xi_0, u^* > + o_p(1) \ge 0$ and then $n^{1/2} < \hat{\xi}_n - \xi_0, u^* >= n^{1/2}\big(P_n - P\big)\big\{\dot{l}\big(\xi_0, O\big)[u^*]\big\} + o_p(1) \to_d N(0, ||u^*||^2)$ by the central limit theorem and $||u^*||^2 = ||\dot{l}\big(\xi_0, O\big)[u^*]||^2$.

Next we will prove that $||u^*||^2 = b^T I^{-1}(\eta_0) b$. For each component η_j, $j = 1, 2, \ldots, p_\eta$, we denote by $\phi_j^* = (b_{1j}^*, b_{2j}^*)$ the value of $\phi_j = (b_{1j}, b_{2j})$ minimizing

$$E\big\{l_\eta \cdot e_j - l_{b_1}[b_{1j}] - l_{b_2}[b_{2j}]\big\}^2,$$

where l_η is the score function for η, l_{b_1} and l_{b_2} are the score operator for Λ_t and Λ_w, and e_j is a p_η-dimensional vector of zeros except the j-th element equal to 1.

Define the j-th element of $l^*(\eta, O)$ as $l_\eta \cdot e_j - l_{b_1}[b^*_{1j}] - l_{b_2}[b^*_{2j}]$, $j = 1, 2, \ldots p_\eta$, and $I(\eta)$ as $E(\{l^*(\eta, O)\}^{\otimes 2})$. By Condition (A.5), the matrix $I(\eta_0)$ is positive definite. Furthermore, by following similar calculations in Chen et al. (2006), we obtain

$$||u^*||^2 = ||\dot{h}(\xi_0)||^2 = \sup_{u \in \bar{V}: ||u|| > 0} \frac{|\dot{h}(\xi_0)[u]|^2}{||u||^2} = b^T [E(\{l^*(\eta_0, O)\}^{\otimes 2})]^{-1} b = b^T I^{-1}(\eta_0) b.$$

Thus, we have shown that $n^{1/2}(\hat{\eta}_n - \eta_0) \to N(0, I^{-1}(\eta_0))$ in distribution for the estimator $\hat{\eta}_n$. □

References

Abadie, A., Angrist, J., Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica, 70*, 91–117.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–455.

Baiocchi, M., Cheng, J., Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine, 33*, 2297–2340.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association, 93*, 929–934.

Chen, X., Fan, Y., Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association, 101*, 1228–1240.

Cheng, J., Small, D. S., Tan, Z., Ten Have, T. R. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika, 96*, 19–36.

Cuzick, J., Sasieni, P., Myles, J., Tyrer, J. (2007). Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology), 69*, 565–588.

Du, M., Zhou, Q., Zhao, S., Sun, J. (2021). Regression analysis of case-cohort studies in the presence of dependent interval censoring. *Journal of Applied Statistics, 48*, 846–865.

Hammer, S. M., Sobieszczyk, M. E., Janes, H., Karuna, S. T., Mulligan, M. J., Grove, D., Koblin, B. A., Buchbinder, S. P., Keefer, M. C., Tomaras, G. D., et al. (2013). Efficacy trial of a dna/rad5 hiv-1 preventive vaccine. *New England Journal of Medicine, 369*, 2083–2092.

Huang, X., Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics, 58*, 510–520.

Huling, J. D., Yu, M., O'Malley, A. J. (2019). Instrumental variable based estimation under the semiparametric accelerated failure time model. *Biometrics, 75*, 516–27.

Janes, H. E., Cohen, K. W., Frahm, N., De Rosa, S. C., Sanchez, B., Hural, J., Magaret, C. A., Karuna, S., Bentley, C., Gottardo, R., et al. (2017). Higher t-cell responses induced by dna/rad5 hiv-1 preventive vaccine are associated with lower hiv-1 infection risk in an efficacy trial. *The Journal of Infectious Diseases, 215*, 1376–1385.

Kalbfleisch, J. D., Prentice, R. L. (2011). *The statistical analysis of failure time data*. New York: Wiley.

Li, G., Lu, X. (2015). A bayesian approach for instrumental variable analysis with censored time-to-event outcome. *Statistics in Medicine, 34*, 664–684.

Li, S., Gray, R. J. (2016). Estimating treatment effect in a proportional hazards model in randomized clinical trials with all-or-nothing compliance. *Biometrics, 72*, 742–750.

Li, S., Peng, L. (2021). Instrumental variable estimation of complier causal treatment effect with interval-censored data. *Biometrics, 79*, 253–263.

Lin, H., Li, Y., Jiang, L., Li, G. (2014). A semiparametric linear transformation model to estimate causal effects for survival data. *Canadian Journal of Statistics, 42*, 18–35.

Lorentz, G. G. (1986). *Bernstein polynomials* (2nd ed.). New York: Chelsea Publishing Co.

Ma, L., Hu, T., Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika, 102*, 731–738.
Ma, L., Hu, T., Sun, J. (2016). Cox regression analysis of dependent interval-censored failure time data. *Computational Statistics Data Analysis, 103*, 79–90.
Nie, H., Cheng, J., Small, D. S. (2011). Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. *Biometrics, 67*, 1397–1405.
O'Malley, A. J., Cotterill, P., Schermerhorn, M. L., Landon, B. E. (2011). Improving observational study estimates of treatment effects using joint modeling of selection effects and outcomes: The case of aaa repair. *Medical care, 49*, 1126.
Shen, X., Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics, 22*, 580–615.
Sun, J. (1999). A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology), 61*, 243–250.
Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.
Van, D., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.
Wang, P., Zhao, H., Sun, J. (2016). Regression analysis of case k interval-censored failure time data in the presence of informative censoring. *Biometrics, 72*, 1103–1112.
Youyi, F., Shen, X., Ashley, V. C., Aaron, D., Seaton, K. E., Yu, C., Grant, S. P., Guido, F., Decamp, A. C., Bailer, R. T. (2018). Modification of the association between t-cell immune responses and human immunodeficiency virus type 1 infection risk by vaccine-induced antibody responses in the hvtn 505 trial. *Journal of Infectious Diseases, 217*, 1280–1288.
Yu, W., Chen, K., Sobel, M. E., Ying, Z. (2015). Semiparametric transformation models for causal inference in time to event studies with all-or-nothing compliance. *Journal of the Royal Statistical Society Series B, Statistical Methodology, 77*, 397–415.
Zeng, D. (2012). Estimating treatment effects with treatment switching via semicompeting risks models: An application to a colorectal cancer study. *Biometrika, 99*, 167–184.
Zhang, Y., Hua, L., Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics, 37*, 338–354.
Zhang, Z., Sun, J., Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine, 24*, 1399–1407.
Zhang, Z., Sun, L., Sun, J., Finkelstein, D. M. (2007). Regression analysis of failure time data with informative interval censoring. *Statistics in Medicine, 26*, 2533–2546.
Zhou, Q., Hu, T., Sun, J. (2016). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association, 112*, 664–672.