Communications in Mathematics and Statistics
https://doi.org/10.1007/s40304-022-00322-6

Analysis of Informatively Interval-Censored Case–Cohort Studies with the Application to HIV Vaccine Trials

Mingyue Du[1] · Qingning Zhou[2]

Received: 18 November 2021 / Revised: 4 January 2022 / Accepted: 6 August 2022

Abstract

Case–cohort studies are commonly used in various investigations, and many methods have been proposed for their analyses. However, most of the available methods are for right-censored data or assume that the censoring is independent of the underlying failure time of interest. In addition, they usually apply only to a specific model such as the Cox model that may often be restrictive or violated in practice. To relax these assumptions, we discuss regression analysis of interval-censored data, which arise more naturally in case–cohort studies than and include right-censored data as a special case, and propose a two-step inverse probability weighting estimation procedure under a general class of semiparametric transformation models. Among other features, the approach allows for informative censoring. In addition, an EM algorithm is developed for the determination of the proposed estimators and the asymptotic properties of the proposed estimators are established. Simulation results indicate that the approach works well for practical situations and it is applied to a HIV vaccine trial that motivated this investigation.

Keywords EM algorithm · Informative censoring · Inverse probability weighting · Joint modeling · Transformation model

Mathematics Subject Classification 62N02 · 62N01

✉ Mingyue Du
dummoon@163.com

[1] School of Mathematics, Jilin University, Changchun 130012, People's Republic of China

[2] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, USA

Published online: 23 March 2023

1 Introduction

Case–cohort studies are commonly used in various investigations as a means of reducing the cost in large cohort studies, especially when the disease rate is low and covariate measurements may be expensive [13, 24, 25]. In these studies, instead of collecting the covariate information on all study subjects, the covariate information is collected only on the subjects whose failures are observed and on a subsample of the remaining subjects. Among others, one area where the case–cohort design is often used is epidemiological cohort studies in which the outcomes of interest are times to failure events such as AIDS, cancer, heart disease and HIV infection. For such studies, in addition to the incomplete nature on covariate information, another feature is that the observations are usually interval-censored rather than right-censored due to the periodic follow-up nature of the study [27].

An example of case–cohort studies that motivated this study was given by the HVTN 505 Trial 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 [8–10, 12]. For the study, among others, one variable of interest is the time to true HIV-1 infection but for which only interval-censored data are available since the study subjects were only examined periodically. In the observed data, the information on some demographical covariates is available for all subjects but the information on some biomarkers was available only for HIV infection cases and a small number of non-cases. One goal of the study is the determination of the important or relevant prognostic covariates or biomarkers for HIV-1 infection. In addition to interval censoring, as discussed below, the censoring mechanism may be informative or related to the HIV infection time. More details about the study will be given below.

By interval-censored failure time data, we usually mean that instead of being observed exactly, the failure time of interest is only known or observed to belong to an interval [27]. It is apparent that interval-censored data can have different forms and include right-censored data as a special case. Among them, the most general form is case K interval-censored data, meaning that there exists a sequence of observation times for each subject [28]. It is easy to see that most of medical follow-up studies such as clinical trials will naturally yield interval-censored data. For the analysis of failure time data, one important factor that one has to pay attention is informative censoring, meaning that the failure time of interest and the censoring mechanism are correlated [11, 28, 30]. Among others, Huang and Wolfe [11] and Sun [27] discussed the issue and pointed out that in the presence of informative censoring, the analysis that ignores it may result in biased results or misleading conclusions. More discussion on informatively interval-censored data can be found in [27].

Many authors have investigated the analysis of case–cohort studies but most of the existing methods are for right-censored failure time data [2–4, 14–16, 20, 23–25]. Several methods were also developed for the analysis of interval-censored case–cohort data but they apply only to some special types of interval-censored data [9, 18, 35]. Furthermore, most of the available methods assume that the censoring mechanism is non-informative or independent of the failure time of interest. As discussed by many authors and above, the informative censoring is a serious and difficult issue and the use of the methods that do not take it into account can yield biased results or misleading conclusions [11, 21]. To the best of our knowledge, the only existing method that can deal with informatively interval-censored case–cohort data was given by Du et al. [6], who developed a frailty model approach but only considered a special type of interval-censored data. More specifically, they discussed the situation where each subject only has two observations, and also they assumed that the failure time of interest follows the proportional hazards model, a special case of the model (2.1) discussed below.

Another main assumption behind the most of the existing methods for case–cohort studies, especially all for interval-censored data, is that they suppose that the failure time of interest follows the Cox proportional: hazards model. It is well-known that the proportional hazards assumption may be violated or not true in some applications. To address these and provide more flexible approaches, in the following, we will consider regression analysis of case K informatively interval-censored case–cohort data, the most general type of interval-censored data, under a class of semiparametric transformation models. Semiparametric transformation models have been commonly used for regression analysis of various types of complete and incomplete data partly due to their flexibility, and this is especially the case for failure time data [1, 5, 7, 17, 19, 30, 33, 34]. For example, they include the Cox proportional hazards model and the proportional odds model among others as special cases. In particular, unlike most of the existing methods, the proposed approach allows for informative censoring.

To present the proposed estimation approach, in the following, we will first introduce some background along with some notation and assumptions in Sect. 2. In particular, the assumed models will be described and the latent variable approach will be used to characterize the association between the failure time of interest and observation processes or informative censoring mechanism. In Sect. 3, a two-step inverse probability weighting estimation procedure is proposed and a novel EM algorithm is developed for the determination of the proposed estimators by following Wang et al. [28]. Note that the main novelty lies on the use of Poisson variables, which makes the implementation much easier than the standard EM algorithm. The consistency and asymptotic normality of the resulting estimators of regression parameters are established in Sect. 4, and a weighted bootstrap procedure is also provided for variance estimation. Some results obtained from a simulation study are presented in Sect. 5, and they suggest that the method works well in practical situations. In Sect. 6, we apply the proposed methodology to the HIV vaccine study described above and some concluding remarks are provided in Sect. 7.

2 Case *K* Interval-Censored Data and Semiparametric Transformation Models

In this section, we will first introduce case K interval-censored case–cohort data and the type of semiparametric transformation models that will be used throughout the paper. It is followed by a brief discussion about the estimation problem if regular interval-censored data are available. Consider a failure time study that consists of n independent subjects. For subject i, let T_i denote the failure time of interest and suppose that there exists a p-dimensional vector of covariates denoted by z_i that may affect T_i and a sequence of observation time points denoted by $U_{i0} = 0 < U_{i1} <$

$U_{i2} < \cdots < U_{iK_i} < \infty = U_{iK_i+1}$, where K_i is a random integer, $i = 1, \ldots, n$. Define $\tilde{N}_i(t) = \sum_{j=1}^{K_i} I(U_{ij} \le t)$ and $\delta_{ij} = I(U_{ij-1} < T_i \le U_{ij}), i = 1, \ldots, n, j = 1, \ldots, K_i + 1$. Then, $\tilde{N}_i(t)$ is a point process characterizing the observation process on subject i and jumps only at the observation times. In the following, it will be assumed that one cannot observe T_i and instead only observes the δ_{ij}'s. That is, we have case K interval-censored data.

For case–cohort studies, as mentioned above, the information on covariates is available only for the subjects who either have experienced the failure event of interest or are from the sub-cohort that is a random sample of the entire cohort. Define $\xi_i = 1$ if the covariate z_i is available or observed and 0 otherwise, $i = 1, \ldots, n$. For the selection of the sub-cohort, by following Zhou et al. [35] and others, we will consider the independent Bernoulli sampling with the selection probability $q \in (0, 1)$. Then under the assumption, the probability that the covariate z_i is observed is given by

$$Pr(\xi_i = 1) = \pi_q(\delta_i) = \sum_{j=1}^{K_i} \delta_{ij} + \delta_{iK_i+1} q \,,$$

$i = 1, \ldots, n$, and the observed data have the form

$$O^{\xi} = \{\, O_i^{\xi} = (\tau_i, U_{i0}, U_{ij}, \delta_{ij}, \xi_i, \xi_i z_i, j = 1, \ldots, K_i + 1); \quad i = 1, \ldots, n \,\} .$$

In the above, $\delta_i = (\delta_{i1}, \ldots, \delta_{iK_i+1})$ and τ_i denotes a follow-up time for the ith subject that is assumed to be independent of T_i. Note that if all covariates were observed, the full cohort data would be

$$O = \{\, O_i = (\tau_i, U_{i0}, U_{ij}, \delta_{ij}, z_i, j = 1, \ldots, K_i + 1); \quad i = 1, \ldots, n \,\} .$$

For the description of the covariate effect on T_i, suppose that for subject i, there exists a latent variable u_i and given z_i and u_i, the cumulative hazard function for T_i takes the form

$$\Lambda(t|z_i, u_i) = G\left[\Lambda(t)\exp(z_i^T \beta_1 + u_i \beta_2)\right] . \qquad (2.1)$$

Here, $G(\cdot)$ is a prespecified increasing transformation function, $\Lambda(\cdot)$ is an unknown increasing function, and $\beta = (\beta_1^T, \beta_2)^T$ represents unknown regression parameters. As mentioned above, one advantage of the semiparametric transformation models is their flexibility as they include many commonly used models as special cases. For example, the choices of $G(x) = x$ and $G(x) = \log(1 + x)$ yield the proportional hazards and proportional odds models, respectively. In the following, we will also assume that given z_i and u_i, $\tilde{N}_i(t)$ is a nonhomogeneous Poisson process with the intensity function

$$\lambda_{ih}(t|z_i, u_i) = \lambda_h(t)\exp(z_i^T \alpha + u_i) , \qquad (2.2)$$

and T_i and $\tilde{N}_i(t)$ are independent. In the above, $\lambda_h(t)$ denotes a completely unknown continuous baseline intensity function and α a vector of regression parameters as β.

Before presenting the proposed estimation procedure, first note that if the full-cohort data O were available, then the conditional likelihood function of the observed data has the form

$$L(\beta, \Lambda | u_i' s) = \prod_{i=1}^{n} \prod_{j=1}^{K_i+1} \left\{ \exp\left(-G\left[\Lambda(U_{ij-1}) \exp(z_i^T \beta_1 + u_i \beta_2)\right]\right) \right. \\ \left. - \exp\left(-G\left[\Lambda(U_{ij}) \exp(z_i^T \beta_1 + u_i \beta_2)\right]\right) \right\}^{\delta_{ij}}$$

given the $u_i's$ and $U_{ij}'s$. Also note that for each subject, only one δ_{ij} is unity and the others equal zero. It follows that the conditional likelihood function above can be rewritten as

$$L(\beta, \Lambda | u_i' s) = \prod_{i=1}^{n} \left\{ \exp\left(-G\left[\Lambda(L_i) \exp(z_i^T \beta_1 + u_i \beta_2)\right]\right) \right. \\ \left. - \exp\left(-G\left[\Lambda(R_i) \exp(z_i^T \beta_1 + u_i \beta_2)\right]\right) \right\} ,$$

where $(L_i, R_i]$ denotes the smallest interval that brackets T_i. It is apparent that $L_i = 0$ indicates that the ith subject is left-censored, while $R_i = \infty$ means that the subject is right-censored.

In practice, of course, the u_i's are unknown, and to deal with this, Wang et al. [30] proposed a borrow-strength estimation procedure as follows. Let $\Lambda_h(t) = \int_0^t \lambda_h(s) \mathrm{d}s$ and assume that $\Lambda_h(\tau_0) = 1$, where τ_0 denotes the longest follow-up time. Also, let $s_{(l)}$'s denote the ordered and distinct values of the observation times U_{ij}'s, $d_{(l)}$ the number of the observation times equal to $s_{(l)}$, and $R_{(l)}$ the number of the observation times satisfying $U_{ij} \le s_{(l)} \le \tau_i$ among all subjects. Then, Wang et al. [30] suggested to estimate α by using the estimating equation

$$U(\alpha) = \sum_{i=1}^{n} w_i \tilde{z}_i \left(K_i \hat{\Lambda}_h^{-1}(\tau_i) - E(e^{u_i}) \exp(z_i^T \alpha) \right) = 0 ,$$

where $\tilde{z}_i^T = (1, z_i^T)$, the $w_i's$ are some weights and

$$\hat{\Lambda}_h(t) = \prod_{s_{(l)} > t} \left(1 - \frac{d_{(l)}}{R_{(l)}} \right) ,$$

the estimator of Λ_h. Let $\hat{\alpha}$ denote the estimator of α given by the estimating equation above. Then, one can estimate β and Λ by maximizing $L(\beta, \Lambda | u_i' s)$ with replacing u_i by

$$\hat{u}_i = \log \left\{ \frac{K_i}{\hat{\Lambda}_h(\tau_i) \exp(z_i^T \hat{\alpha})} \right\} .$$

In the next section, we will generalize the approach above to case–cohort studies, and for this, there exist several difficulties. One is that although the generalized method may seem to be straightforward, the determination of the proposed estimators is much more complicated than that considered in Wang et al. [30] due to missing covariates and a novel EM algorithm has to be developed for the purpose. Another difficulty is the establishment of the asymptotic properties of the proposed estimators again due to the missing information. More comments on these are given below.

3 Inverse Probability Weighting Estimation

Now we consider inference procedure for case–cohort studies with the focus on estimation of the regression parameter β. For this, let $\hat{\alpha}$ denote the same estimator defined in the previous section but with the weight

$$w_i = \frac{\xi_i}{\pi_q(\delta_i)} = \frac{\xi_i}{\sum_{j=1}^{K_i} \delta_{ij} + \delta_{iK_i+1} q} .$$

For estimation of β as well as Λ, we propose to maximize the inverse probability weighted log-likelihood function

$$l_n^w(\beta, \Lambda | \hat{u}_i' s) = \sum_{i=1}^{n} w_i \log \Big\{ \exp\Big(-G \Big[\Lambda(L_i) \exp(z_i^T \beta_1 + \hat{u}_i \beta_2) \Big] \Big) - \exp\Big(-G \Big[\Lambda(R_i) \exp(z_i^T \beta_1 + \hat{u}_i \beta_2) \Big] \Big) \Big\} . \qquad (3.1)$$

In the remaining of this section, we will discuss the maximization of $l_n^w(\beta, \Lambda | \hat{u}_i' s)$ by developing an EM algorithm. For this, first note that the transformation function G can be derived by the Laplace transformation of the frailty variable with the support $[0, \infty]$ through

$$\exp\{-G(x)\} = \int_0^\infty \exp(-x\xi) \phi(\xi | r) \mathrm{d}\xi ,$$

where $\phi(\xi | r)$ is the density function of the frailty ξ. More specifically, by letting $\phi(\xi | r)$ be the gamma density function with mean 1 and variance r, we can obtain $G(x) = \log(1 + rx)/r$, the logarithmic transformation function family. In consequence, one can convert the class of transformation models into the proportional hazards frailty model and rewrite the inverse probability weighted log-likelihood function above as

$$l_n^w(\beta, \Lambda|\hat{u}_i's) = \sum_{i=1}^n w_i \log\left\{\int_{\xi_i}\Big(\exp\Big[-\Lambda(L_i)\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\xi_i\Big] - \exp\Big[-\Lambda(R_i)\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\xi_i\Big]\Big)\phi(\xi_i|r)\mathrm{d}\xi_i\right\}. \quad (3.2)$$

Let $0 = t_0 < t_1 < \cdots < t_m$ denote the set of the observation time points consisting of 0 and the unique values of $L_0 > 0$ and $R_i < \infty$ $(i = 1, \ldots, n)$. In the following, we will treat Λ as a step function with nonnegative jumps at the $t_k's$ and the jump size λ_k at t_k. Assume that $\lambda_0 = 0$. Then, the log-likelihood $l_n^w(\beta, \Lambda|\hat{u}_i's)$ can be rewritten as

$$\begin{aligned}
&l_n^w(\beta, \Lambda|\hat{u}_i's)\\
&= \sum_{i=1}^n w_i \log\left\{\int_{\xi_i}\exp\left\{-\sum_{t_k\le L_i}\lambda_k\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\xi_i\right\}\right.\\
&\quad\left.\left[1 - \exp\left\{-\sum_{L_i<t_k\le R_i}\lambda_k\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\xi_i\right\}\right]^{I(R_i<\infty)}\phi(\xi_i|r)\mathrm{d}\xi_i\right\}. \quad (3.3)
\end{aligned}$$

By following Wang et al. [28], we introduce the latent variables $\{P_{ik};\, i = 1, \ldots, n, k = 1, \ldots, m\}$ which, conditional on ξ_i, are independent Poisson random variables with the mean $\lambda_k\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\xi_i$. Then, $l_n^w(\beta, \Lambda|\hat{u}_i's)$ can be equivalently expressed as

$$\begin{aligned}
l_n^w(\beta, \Lambda|\hat{u}_i's) &= \sum_{i=1}^n w_i \log\left\{\int_{\xi_i}\left\{\prod_{t_k\le L_i}p(P_{ik} = 0|\xi_i)\right\}\right.\\
&\quad\left.\left[1 - p\left(\sum_{L_i<t_k\le R_i}P_{ik} = 0|\xi_i\right)\right]^{I(R_i<\infty)}\phi(\xi_i|r)\mathrm{d}\xi_i\right\}. \quad (3.4)
\end{aligned}$$

For the EM algorithm to be developed for maximizing $l_n^w(\beta, \Lambda|\hat{u}_i's)$, we will treat the $\xi_i's$ and $P_{ik}'s$ as missing data. If they were known, the pseudo complete data log-likelihood function would be

$$\begin{aligned}
l_n^{w*}(\beta, \Lambda|\hat{u}_i's) &= \sum_{i=1}^n w_i\Bigg(\sum_{k=1}^m\Big[P_{ik}\log\left\{\xi_i\lambda_k\exp(z_i^T\beta_1 + \hat{u}_i\beta_2)\right\}\\
&\quad -\xi_i\lambda_k\exp(z_i^T\beta_1 + \hat{u}_i\beta_2) - \log P_{ik}!\Big] + \log\phi(\xi_i|r)\Bigg). \quad (3.5)
\end{aligned}$$

In the above, we require that $\sum_{t_k \le L_i} P_{ik} = 0$ and $\sum_{L_i < t_k \le R_i} P_{ik} > 0$ if $R_i < \infty$, and $\sum_{t_k \le L_i} P_{ik} = 0$ if $R_i = \infty$. In the M-step, we calculate

$$\lambda_k = \frac{\sum_{i=1}^n w_i \hat{E}(P_{ik})}{\sum_{i=1}^n w_i \hat{E}(\xi_i) \exp(z_i^T \beta_1 + \hat{u}_i \beta_2)}\,, \qquad (k = 1, \ldots, m)\,, \tag{3.6}$$

where $\hat{E}(\cdot)$ denotes the posterior mean given the observed data. After incorporating (3.6) into the conditional expectation of (3.5), we update β by solving the following equation with the use of the one-step Newton–Raphson method

$$\sum_{i=1}^n w_i \left\{ \sum_{k=1}^m \hat{E}(P_{ik}) \left[\hat{x}_i - \frac{\sum_{j=1}^n w_j \hat{E}(\xi_j) \exp(z_j^T \beta_1 + \hat{u}_j \beta_2) \hat{x}_i}{\sum_{j=1}^n w_j \hat{E}(\xi_j) \exp(z_j^T \beta_1 + \hat{u}_j \beta_2)} \right] \right\} = 0\,, \tag{3.7}$$

where $\hat{x}_i = (z_i^T, \hat{u}_i)^T$.

In the E-step, we evaluate the posterior means $\hat{E}(P_{ik})$ and $\hat{E}(\xi_i)$. For this, note that the posterior density function of ξ_i given the observed data is proportional to $\{\exp(-\xi_i S_{i1}) - \exp(-\xi_i S_{i2})\}\phi(\xi_i|r)$, where $S_{i1} = \sum_{t_k \le L_i} \lambda_k \exp(z_i^T \beta_1 + \hat{u}_i \beta_2)$ and $S_{i2} = \sum_{t_k \le R_i} \lambda_k \exp(z_i^T \beta_1 + \hat{u}_i \beta_2)$. Thus, we have that

$$\begin{aligned} \hat{E}(P_{ik}) = {} & I(R_i < \infty)\lambda_k \exp(z_i^T \beta_1 + \hat{u}_i \beta_2) I(L_i < t_k \le R_i) \\ & \times \frac{\int_{\xi_i} \xi_i \{\exp(-\xi_i S_{i1}) - \exp(-\xi_i S_{i2})\}[1 - \exp\{-\xi_i (S_{i2} - S_{i1})\}]^{-1} \phi(\xi_i|r) \mathrm{d}\xi_i}{\exp\{-G(S_{i1})\} - \exp\{-G(S_{i2})\}} \\ & + I(R_i < \infty)\lambda_k \exp(z_i^T \beta_1 + \hat{u}_i \beta_2) E(\xi_i) I(t_k > R_i) \\ & + I(R_i = \infty)\lambda_k \exp(z_i^T \beta_1 + \hat{u}_i \beta_2) E(\xi_i) I(t_k > L_i), \end{aligned} \tag{3.8}$$

which can be calculated by using the Gaussian–Laguerre quadrature. In addition,

$$\begin{aligned} \hat{E}(\xi_i) = {} & I(R_i < \infty) \frac{\exp\{-G(S_{i1})\} G'(S_{i1}) - \exp\{-G(S_{i2})\} G'(S_{i2})}{\exp\{-G(S_{i1})\} - \exp\{-G(S_{i2})\}} \\ & + I(R_i = \infty) G'(S_{i1})\,, \end{aligned}$$

where

$$G'(x) = \frac{\int_{\xi_i} \xi_i \exp(-x\xi_i) \phi(\xi_i|r) \mathrm{d}\xi_i}{\exp\{-G(x)\}} = \frac{(rx + 1)^{-r^{-1}-1}}{\exp\{-G(x)\}}\,.$$

4 Consistency and Asymptotic Normality

In this section, we will establish the asymptotic properties of the estimators proposed in the previous sections. Let $\hat{\beta}$ and $\hat{\Lambda}$ denote the estimators of β and Λ defined in the previous section and $\beta_0, \alpha_0, \Lambda_0$ and Λ_{0h} the true values of β, α, Λ and Λ_h, respectively.

To establish the asymptotic properties of $\hat{\beta}$ and $\hat{\Lambda}$, we need the following regularity conditions.

(C1) β_0 is an interior point of a known compact set $\mathcal{B} \in R^{p+1}$, and $\Lambda_0(\cdot)$ is continuously differentiable with positive derivatives in $[0, \tau_0]$. Moreover, α_0 lies in the interior of a known compact set $\mathcal{A} \in R^p$ and $\Lambda_{0h}(\cdot)$ is continuously differentiable with positive derivatives in $[0, \tau_0]$ satisfying $\Lambda_{0h}(\tau_0) = 1$.

(C2) The covariate vector z and the latent variable u are bounded and satisfy $E[\exp(u)|z] = E[\exp(u)]$. Moreover, if $h(t) + b_1^T z + b_2 u = 0$ for all $t \in [0, \tau_0]$ with probability 1, then $h(t) = 0$ for $t \in [0, \tau_0]$ and $b_1 = b_2 = 0$.

(C3) The follow-up time τ satisfies that $Pr(\tau \in [\zeta_0, \tau_0]) = 1$ for some constant $\zeta_0 \in (0, \tau_0]$ such that $\Lambda_{0h}(\zeta_0) > 0$. In addition, $Pr(\tau = \tau_0) > 0$.

(C4) The transformation function $G(\cdot)$ is twice continuously differentiable on $[0, \infty)$ with positive first derivatives, and it satisfies $G(0) = 0$ and $G(\infty) = \infty$.

(C5) $0 < q \leq \pi_q(\delta_1, \ldots, \delta_{K+1}) \leq 1$.

Now we can establish the asymptotic properties, including the strong consistency of $\hat{\beta}$ and $\hat{\Lambda}$ in Theorem 4.1 and the asymptotic normality of $\hat{\beta}$ in Theorem 4.2.

Theorem 4.1 *Assume that the regularity conditions* (*C*1)–(*C*5) *given above hold. Then as* $n \to \infty$, *we have that* $\|\hat{\beta} - \beta_0\| + \sup_{t \in [0, \tau_0]} |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0$ *almost surely, where* $\| \cdot \|$ *denotes the Euclidean norm.*

Theorem 4.2 *Assume that the regularity conditions* (*C*1)–(*C*5) *given above hold. Then as* $n \to \infty$, *we have that* $n^{1/2}(\hat{\beta} - \beta_0)$ *converges in distribution to a normal random vector with mean zero.*

The proof of the results above is sketched in the Appendix. For inference about β, it is apparent that one needs to estimate the covariance matrix of $\hat{\beta}$, and for this, it can be seen from the proof that it would be difficult to derive a consistent estimator. Corresponding to this, we propose to employ the weighted bootstraps procedure discussed in Ma and Kosorok [22], which is easy to implement and seems to work well in the numerical studies described below. More specifically, let $\{a_1, \ldots, a_n\}$ denote n independent realizations of a bounded positive random variable a satisfying $E(a) = 1$ and $var(a) = \epsilon_0 < \infty$ and define the new weights $w_i^* = a_i\, w_i$, $i = 1, \ldots, n$. Also, let $\hat{\beta}^*$ denote the estimator of β proposed above with replacing the w_i's by the w_i^*'s. Then if we repeat this B times, one can estimate the covariance matrix of $\hat{\beta}$ by the sample covariance matrix of the $\hat{\beta}^*$'s.

5 A Simulation Study

Now we report some results obtained from a simulation study conducted to evaluate the finite sample performance of the two-step inverse probability weighted estimation procedure proposed in the previous sections. In the study, it was assumed that the covariate z followed the Bernoulli distribution with the success probability of 0.5, and to generate the sub-cohort, as mentioned above, we considered the independent Bernoulli sampling with the selection probability being $q = 0.1$. To generate the true failure times, we first generated the latent variables u_i's by assuming that $u_i^* = \exp(u_i)$ follows the gamma distribution with mean 4 and variance 8. Given the $z_i's$ and $u_i's$, the true failure times were then generated from the transformation model (2.1) with $\Lambda(t) = \log(1 + t/65)$ and $G(x) = \log(1 + rx)/r$ for $r = 0$, 0.5 or 1.

For the generation of the observation process or censoring intervals, it was supposed that $\tilde{N}_i(t)$ follows model (2.2) with $\lambda_h = 1/4$ and the $\tau_i's$ follow the uniform distribution over the interval [3, 4], which gives the proportion of the observed failure events or the event rate of $p_e = 0.05$. Then given z_i, u_i and τ_i, K_i, the number of observation times for subject i, followed the Poisson distribution with mean

$$\Lambda_{ih}(\tau_i | z_i, u_i) = \frac{\tau_i \exp(z_i^T \alpha + u_i)}{4},$$

and the observation times $(U_{i1}, \ldots, U_{iK_i})$ were taken to be the order statistics of a random sample of size K_i from the uniform distribution over $(0, \tau_i)$, $i = 1, 2, \ldots, n$. The results given below are based on $n = 2000$ and $B = 50$ with 1000 replications.

Table 1 presents the results on estimation of the parameters α and β given by the proposed estimation procedure with the true value of all parameters being 0.2. It includes the estimated bias given by the average of the estimates minus the true value (Bias), the sample standard error of the estimates (SSE), the average of the estimated standard errors (ESE), and the 95% empirical coverage probability (CP). The results suggest that the proposed estimators seem to be unbiased and the variance estimation also appears to be reasonable. In addition, they indicate that the normal approximation to the distribution of the proposed estimator seems to be appropriate. Note that the bias on estimation of β_2 may seem to be little large. On the other hand, this is expected since β_2 corresponds to the latent variable effect [6, 28], and more importantly, the main interest here is on estimation of β_1, the covariate effects. We also considered some other set-ups, including different values for the true regression parameters, the event rate of p_e and B as well as different functions for $\Lambda(t)$. All results are similar to the above and suggest that the proposed approach is valid and works well.

Note that in the estimation procedure proposed above, it has been assumed that the observation process follows the Poisson process and it is apparent that sometimes this may not be true. To investigate the dependence of the proposed procedure on the assumption, we repeated the study above by generating the observation times from mixed Poisson processes. In particular, the number of observation times K_i for subject i was generated from the mixed Poisson distribution with mean

$$\Lambda_{ih}(\tau_i | z_i, u_i) = \frac{\gamma_i \tau_i \exp(z_i^T \alpha + u_i)}{4},$$

where the $\gamma_i's$ follow the gamma distribution with mean 1 and variance 0.01. The results on estimation of the parameters α and β obtained by the proposed estimation procedure are given in Table 2 with the other set-ups being the same as in Table 1. One can see that they are similar to those presented in Table 1 and again indicate that the proposed estimation procedure works well for the situations considered and seems to be robust to the Poisson assumption.

Table 1 Simulation results on estimation of regression parameters

r	True value	Bias	SSE	ESE	CP
0	$\alpha = 0.2$	– 0.0260	0.1092	0.1045	0.9270
	$\beta_1 = 0.2$	0.0009	0.2505	0.2469	0.9530
	$\beta_2 = 0.2$	– 0.0490	0.1829	0.1753	0.9360
0.5	$\alpha = 0.2$	– 0.0195	0.1095	0.1049	0.9270
	$\beta_1 = 0.2$	– 0.0013	0.2592	0.2517	0.9380
	$\beta_2 = 0.2$	– 0.0552	0.1841	0.1785	0.9400
1	$\alpha = 0.2$	– 0.0238	0.1078	0.1047	0.9300
	$\beta_1 = 0.2$	0.0112	0.2572	0.2582	0.9480
	$\beta_2 = 0.2$	– 0.0640	0.1781	0.1817	0.9440

Table 2 Simulation results with misspecified observation processes

r	True value	Bias	SSE	ESE	CP
0	$\alpha = 0.2$	– 0.0239	0.1083	0.1058	0.9280
	$\beta_1 = 0.2$	0.0065	0.2473	0.2469	0.9480
	$\beta_2 = 0.2$	– 0.0554	0.1720	0.1752	0.9420
0.5	$\alpha = 0.2$	– 0.0278	0.1077	0.1051	0.9360
	$\beta_1 = 0.2$	0.0063	0.2576	0.2528	0.9380
	$\beta_2 = 0.2$	– 0.0521	0.1705	0.1773	0.9450
1	$\alpha = 0.2$	– 0.0216	0.1056	0.1061	0.9520
	$\beta_1 = 0.2$	0.0086	0.2673	0.2590	0.9330
	$\beta_2 = 0.2$	– 0.0625	0.1821	0.1809	0.9380

6 The Analysis of HVTN 505 Vaccine Trial

In this section, we will apply the methodology proposed in the previous sections to the vaccine study discussed above, the HVTN 505 Trial, which is a randomized, multiple-sites clinical trial of men or transgender women who had sex with men for assessing the efficacy of the DNA/rAd5 vaccine for HIV-1 infection [8, 10, 12]. 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. In the original study, the recruited subjects were randomly assigned to receive either the DNA/rAd5 vaccine or placebo. For the analysis below, by following Du et al. [6], we will focus on the HIV-1 infection time, the failure time of interest, based on the data from the 1253 subjects in the vaccine group. As mentioned above, only interval-censored data are available due to the design of the study.

During the study, for each subject, four demographic covariates were observed and they are age, race, BMI and behavioural risk. In addition, to assess their relationship with the HIV-1 infection, a number of T cell response biomarkers and antibody response biomarkers were measured for a cohort of 150 subjects consisting of all HIV infection cases 25 and other 125 randomly selected subjects among the vaccine recipients. In particular, all of the previous analyses have identified the T cell response biomarker Env CD8+ polyfunctionality score and the antibody response biomarker IgG.Cconenv03140CF.avi, which will be referred below to as Env CD8 Score and IgG, respectively, to have some significant effects on the HIV infection time [6, 8, 10, 12]. In the following, we will focus on these two biomarkers along with race and behavioural risk to assess their possible effects on the HIV-1 infection since all of the existing joint analyses suggested that age and BMI did not seem to have any effects.

Table 3 presents the analysis results given by the proposed approach by taking $G(x) = \log(1 + rx)/r$ with $r = 0.4$ as in the simulation study. Here, the r value was chosen based on the grid search as it gave the smallest AIC value. The table includes the estimated effects on both the HIV-1 infection time and the observation process for each covariate, the estimated standard errors (ESE), and the p value for testing the effect being zero. For comparison, the estimation results obtained under $r = 0$, 0.5 and 1 are also included. One can see from the table that the results seem to be consistent with respect to r and suggest that only the biomarker Env CD8 Score seems to have significant effect on the HIV-1 infection. Also, none of the covariates appears to have any significant effect on the observation process, but the results indicate that the HIV-1 infection time and the observation process seem to be significantly related. In contrast, the application of the method proposed in [35] would suggest that both biomarkers had significant effects on the HIV-1 infection, while Du et al. [6] indicated

Table 3 Estimated covariate effects for the HVTN 505 Trial

r	Covariate	$\hat{\beta}$	SSE	p value	$\hat{\alpha}$	ESE	p value
0	Race	– 1.5993	1.0380	0.1234	– 0.7601	1.0285	0.4599
	Behavioural risk	0.7468	1.4024	0.5944	0.0167	0.7681	0.9827
	Env CD8 Score	– 3.2794	1.7288	0.0578	– 1.1596	0.9746	0.2341
	IgG	– 0.2352	0.6460	0.7157	0.0855	0.7861	0.9134
	$\hat{\beta}_2$	1.5997	0.1963	0.0000			
0.4	Race	– 0.7570	0.8300	0.3618	– 0.7600	1.1959	0.5251
	Behavioural risk	0.3872	1.1773	0.7422	0.0167	1.1287	0.9882
	Env CD8 Score	– 3.0353	1.1585	0.0088	– 1.1596	1.0549	0.2716
	IgG	– 0.3057	0.5389	0.5706	0.0855	0.9285	0.9266
	$\hat{\beta}_2$	1.5944	0.1639	0.0000			
0.5	Race	– 0.6786	0.9268	0.4641	– 0.7601	1.4533	0.6010
	Behavioural risk	0.5050	1.4504	0.7277	0.0167	1.1996	0.9889
	Env CD8 Score	– 3.1005	1.0605	0.0035	– 1.1596	1.1122	0.2971
	IgG	– 0.2623	0.5152	0.6106	0.0855	0.8084	0.9158
	$\hat{\beta}_2$	1.5786	0.1617	0.0000			
1	race	– 0.3916	0.6726	0.5605	– 0.7601	1.4472	0.5994
	Behavioural risk	1.0256	1.1950	0.3908	0.0167	1.2430	0.9893
	Env CD8 Score	– 3.4832	1.0428	0.0008	– 1.1596	0.9863	0.2397
	IgG	– 0.1165	0.5591	0.8349	0.0855	0.8786	0.9225
	$\hat{\beta}_2$	1.5244	0.1658	0.0000			

that in addition to the two biomarkers, behavioural risk also had some effects on the HIV-1 infection. As discussed above, the former ignores the informative censoring and the latter only applies to case II data and models the informative censoring through the length of observation times.

7 Concluding Remarks

This paper discussed inference about semiparametric transformation models when one faces interval-censored failure time data arising from case–cohort studies with informative or dependent censoring. For the problem, a two-step inverse probability weighting estimation approach was proposed, and for the implementation of the approach, an EM algorithm based on Poisson variables was developed. The proposed estimators of regression parameters were shown to be consistent and asymptotically normal, and the numerical study was performed and suggests that the proposed approach works well for practical situations. It is worth to pointing out that although only considered the special cases, Zhou et al. [35] and Du et al. [6] gave more and similar simulation results.

The proposed estimation approach can be seen as generalizations of the methods given in Zhou et al. [35] and Wang et al. [30]. The former considered the data arising from the Cox proportional hazards model and assumed that the censoring is independent. Although the method given in the latter allows for informative censoring, it assumed that the complete information on covariates is available and unlike the case discussed above, one advantage of their situation is that the proposed method can be relatively easily implemented. As mentioned above, Du at al. [6] also investigated the same problem as considered here but their method only applies to a much limited situation.

As discussed above, for the situation considered here, one naive approach is to ignore the informative censoring but the resulting analysis could easily lead to biased estimation and even misleading conclusions. Another commonly used naive approach is to simplify the interval-censored data structure to right-censored data structure by employing, for example, an imputation procedure. As discussed by many authors under different contexts, this could result in biased analysis too and also lose some efficiency. Furthermore, it does not seem to exist an established approach for regression analysis of right-censored case–cohort data with informative censoring.

Note that in the preceding sections, the focus has been on case–cohort designs with time-independent covariates. Instead, sometimes one may want to employ generalized case–cohort designs, and for this, a new estimation would be needed. Also, it is apparent that sometimes there may exist time-dependent covariates, and some modifications would be needed too to apply the proposed method to the time-dependent covariate case. In the proposed method, another assumption used is the intensity model (2.2) for the observation process. Instead, one may prefer to employ the proportional mean or rate model, and in this case, the proposed method should still be valid although some modifications may be needed for the theoretical justification.

There exist several directions for future research. One is that in the proposed estimation procedure, we have assumed that the observation process is a Poisson process.

Although the numerical results indicate that the assumption can be relaxed, it would be helpful to provide some theoretical justification. The focus of this paper has been on univariate analysis of case–cohort studies, and it is apparent that sometimes there may exist more than one failure times of interest. Thus, it would be useful to generalize the proposed estimation procedure to bivariate or multivariate failure time situations. A third direction, also a more difficult task, is the development of a model checking procedure. In the preceding sections, by following others [30, 33, 34], we have employed the AIC to choose the best model under a given class of G. Although many authors have pointed out this need, it does not seem to exist an established procedure for the situation discussed here or similar situations.

Acknowledgements The authors wish to thank two reviewers for their many insightful and helpful comments and suggestions that greatly improved the paper. We also want to thank Dr. Peter Gibert for providing the HVTN 505 Vaccine Trial data. The first author's work was partially supported by the National Natural Science Foundation of China grant # 12101522 and the second author's work in part by the U.S. National Science Foundation grant DMS1916170.

8 Appendix: Proofs of Theorems 4.1 and 4.2

In the following, we will sketch the proofs of Theorems 4.1 and 4.2. Let $\mathbb{P}_n$ denote the empirical measure for n independent observations, $\mathbb{P}$ the true probability measure, and $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$ the empirical process. Let $l(\beta, \Lambda|u)$ be the log-likelihood for a single subject based on the complete data O, given by

$$l(\beta, \Lambda|u) = \sum_{k=1}^{K+1} \delta_k \log \Big\{ \exp(-G[\Lambda(U_{k-1}) \exp(\beta_1^T z + \beta_2 u)]) \\ - \exp(-G[\Lambda(U_k) \exp(\beta_1^T z + \beta_2 u)]) \Big\},$$

and let $l^w(\beta, \Lambda|u) = w\, l(\beta, \Lambda|u)$ be the weighted log-likelihood for a single subject based on the observed data O^ξ under the case–cohort design, where the weight w is given by $w = \xi/\pi_q(\delta_1, \ldots, \delta_{K+1})$. Since $E(w|\delta_1, \ldots, \delta_{K+1}) = 1$, we have $\mathbb{P}\{l^w(\beta, \Lambda|u)\} = \mathbb{P}\{l(\beta, \Lambda|u)\}$.

Proof of Theorem 4.1 We first show that $\limsup_n \hat{\Lambda}(\tau_0 - \epsilon) < \infty$ with probability 1 for any $\epsilon > 0$. By the definition of $(\hat{\beta}, \hat{\Lambda})$, we have

$$\mathbb{P}_n l^w(\hat{\beta}, \hat{\Lambda}|\hat{u}) \geq \mathbb{P}_n l^w(\beta_0, \Lambda_0|\hat{u}).$$

From the consistency of $(\hat{\alpha}, \hat{\Lambda}_h)$ established by Wang et al. [29], we can show that

$$\liminf_n \mathbb{P}_n l^w(\hat{\beta}, \hat{\Lambda}|\hat{u}) \geq \liminf_n \mathbb{P}_n l^w(\beta_0, \Lambda_0|\hat{u}) = \mathbb{P}l(\beta_0, \Lambda_0|u) = O(1)$$

with probability 1. Define $u(\alpha, \Lambda_h; \tau, K, z) = \log\{K/[\Lambda_h(\tau)\exp(\alpha^T z)]\}$. Let $\eta > 0$ be such that $\exp\{\beta_1^T z + \beta_2 u(\alpha, \Lambda_h; \tau, K, z)\} \geq \eta$ for $\beta \in \mathcal{B}$, $\alpha \in \mathcal{A}$, $\tau \in [\zeta_0, \tau_0]$,

$1 \le K \le k_0$, and nondecreasing functions Λ_h such that $\Lambda_h(\zeta_0) \ge \Lambda_{0h}(\zeta_0) - c_0 > 0$ and $\Lambda_h(\tau_0) \le 1$, where $k_0 > 1$ and c_0 are positive constants. Then, we have

$$\begin{aligned}
&\liminf_n \mathbb{P}_n l^w(\hat\beta, \hat\Lambda|\hat u) \\
&\quad \le -\limsup_n \mathbb{P}_n \Big\{ w\delta_{K+1} G[\hat\Lambda(U_K)\exp(\hat\beta_1^T z + \hat\beta_2 \hat u)] \Big\} \\
&\quad \le -\limsup_n \mathbb{P}_n \Big\{ w\delta_{K+1} I(1 \le K \le k_0) G[\hat\Lambda(U_K)\eta] \Big\} \\
&\quad \le -\limsup_n \mathbb{P}_n \Big\{ w\delta_{K+1} I(1 \le K \le k_0, U_K \ge \tau_0 - \epsilon) G[\hat\Lambda(\tau_0 - \epsilon)\eta] \Big\}.
\end{aligned}$$

Hence,

$$\limsup_n \mathbb{P}_n \Big\{ w\delta_{K+1} I(1 \le K \le k_0, U_K \ge \tau_0 - \epsilon) G[\hat\Lambda(\tau_0 - \epsilon)\eta] \Big\} = O(1).$$

Note that as $n \to \infty$, $\mathbb{P}_n\{w\delta_{K+1} I(1 \le K \le k_0, U_K \ge \tau_0 - \epsilon)\} \to \mathbb{P}\{w\delta_{K+1} I(1 \le K \le k_0, U_K \ge \tau_0 - \epsilon)\}$, which is positive under Condition (C3). Thus, by Condition (C4), $\limsup_n \hat\Lambda(\tau_0 - \epsilon) < \infty$ with probability 1 for any $\epsilon > 0$. By Helly's selection theorem and arguing as in the proof of Theorem 4.1 of Zeng et al. [32], for any subsequence of $(\hat\beta, \hat\Lambda)$, we can choose a further subsequence such that $\hat\Lambda$ converges weakly to some function Λ^* on $[0, \tau_0]$ almost everywhere and $\hat\beta$ converges to some constant β^*. The remaining is to show $(\beta^*, \Lambda^*) = (\beta_0, \Lambda_0)$.

Define

$$m(\beta, \Lambda|u) = w \,\log \left\{ \frac{p(\beta, \Lambda|u) + p(\beta_0, \Lambda_0|u)}{2} \right\},$$

where $p(\beta, \Lambda|u) = \exp(l(\beta, \Lambda|u))$. Since $\mathbb{P}_n l^w(\hat\beta, \hat\Lambda|\hat u) \ge \mathbb{P}_n l^w(\beta_0, \Lambda_0|\hat u)$, we have

$$\mathbb{P}_n m(\hat\beta, \hat\Lambda|\hat u) \ge \mathbb{P}_n l^w(\beta_0, \Lambda_0|\hat u) = \mathbb{P}_n m(\beta_0, \Lambda_0|\hat u)$$

and thereby

$$\begin{aligned}
&[\mathbb{P}_n m(\hat\beta, \hat\Lambda|\hat u) - \mathbb{P} m(\beta^*, \Lambda^*|u)] + \mathbb{P} m(\beta^*, \Lambda^*|u) \\
&\quad \ge [\mathbb{P}_n m(\beta_0, \Lambda_0|\hat u) - \mathbb{P} m(\beta_0, \Lambda_0|u)] + \mathbb{P} m(\beta_0, \Lambda_0|u).
\end{aligned}$$

Arguing as in Zeng et al. [32], we can show that $\mathcal{M} = \{m(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z)) : \beta \in \mathcal{B}, \alpha \in \mathcal{A}, \Lambda \in \mathcal{L}, \Lambda_h \in \mathcal{L}_h\}$ is a Glivenko–Cantelli class, where $\mathcal{L}$ is the set of nondecreasing functions Λ on $[0, \tau_0]$ satisfying $\Lambda(0) = 0$ and $\mathcal{L}_h$ is the set of nondecreasing functions Λ_h on $[0, \tau_0]$ satisfying $\Lambda_h(0) = 0$, $\Lambda_h(\zeta_0) \ge \Lambda_{0h}(\zeta_0) - c_0 > 0$ for some positive constant c_0 and $\Lambda_h(\tau_0) \le 1$. Furthermore, based on the asymptotic properties of $(\hat\alpha, \hat\Lambda_h)$ established by Wang et al. [29], we can show that $\mathbb{P}_n m(\beta, \Lambda|\hat u)$ converges to $\mathbb{P} m(\beta, \Lambda|u)$ almost surely for any fixed (β, Λ). Therefore, we have $\mathbb{P}m(\beta^*, \Lambda^*|u) \geq \mathbb{P}m(\beta_0, \Lambda_0|u)$ and further

$$\mathbb{P}\log\left\{\frac{p(\beta^*, \Lambda^*|u) + p(\beta_0, \Lambda_0|u)}{2}\right\} \geq \mathbb{P}\log p(\beta_0, \Lambda_0|u).$$

By the properties of the Kullback–Leibler information, $p(\beta^*, \Lambda^*|u) = p(\beta_0, \Lambda_0|u)$ with probability 1. Thus, for any $t \in [0, \tau_0]$, $\log\{\Lambda^*(t)\}+\beta_1^{*T} z+\beta_2^* u = \log\{\Lambda_0(t)\}+\beta_{01}^T z+\beta_{02}u$. Under Condition (C2), we obtain $\beta^* = \beta_0$ and $\Lambda^* = \Lambda_0$. This completes the proof. □

Proof of Theorem 4.2 Let $\beta = (\beta_1^T, \beta_2)^T$ and $x = (z^T, u)^T$. The score function for β based on the log-likelihood $l(\beta, \Lambda|u)$ is

$$\begin{aligned}
&l_\beta(\beta, \Lambda|u)\\
&= \sum_{k=1}^{K+1} \delta_k \left\{\frac{-\exp(-G[\Lambda(U_{k-1})\exp(\beta^T x)])G'[\Lambda(U_{k-1})\exp(\beta^T x)]\Lambda(U_{k-1})}{M(U_{k-1}, U_k; \beta, \Lambda, x)}\right.\\
&\quad \left.+\frac{\exp(-G[\Lambda(U_k)\exp(\beta^T x)])G'[\Lambda(U_k)\exp(\beta^T x)]\Lambda(U_k)}{M(U_{k-1}, U_k; \beta, \Lambda, x)}\right\}\exp(\beta^T x)x,
\end{aligned}$$

where

$$M(u, v; \beta, \Lambda, x) = \exp(-G[\Lambda(u)\exp(\beta^T x)]) - \exp(-G[\Lambda(v)\exp(\beta^T x)]).$$

The score function for β based on the weighted log-likelihood $l^w(\beta, \Lambda|u)$ is given by

$$l_\beta^w(\beta, \Lambda|u) = w\, l_\beta(\beta, \Lambda|u).$$

To obtain the score operator for Λ, we consider a parametric submodel of Λ defined by $d\Lambda_{\epsilon,h} = (1+\epsilon h)d\Lambda$ for $h \in L_2([0, \tau_0])$. The score function along this submodel based on the log-likelihood $l(\beta, \Lambda|u)$ is

$$\begin{aligned}
&l_\Lambda(\beta, \Lambda|u)(h)\\
&= \frac{\partial}{\partial\epsilon} l(\beta, \Lambda_{\epsilon,h}|u)\Big|_{\epsilon=0}\\
&= \sum_{k=1}^{K+1} \delta_k \left\{\frac{-\exp(-G[\Lambda(U_{k-1})\exp(\beta^T x)])G'[\Lambda(U_{k-1})\exp(\beta^T x)]}{M(U_{k-1}, U_k; \beta, \Lambda, x)}\int_0^{U_{k-1}} h(t)\mathrm{d}\Lambda(t)\right.\\
&\quad \left.+\frac{\exp(-G[\Lambda(U_k)\exp(\beta^T x)])G'[\Lambda(U_k)\exp(\beta^T x)]}{M(U_{k-1}, U_k; \beta, \Lambda, x)}\int_0^{U_k} h(t)\mathrm{d}\Lambda(t)\right\}\exp(\beta^T x).
\end{aligned}$$

The score function along this submodel based on the weighted log-likelihood $l^w(\beta, \Lambda|u)$ is

$$l_\Lambda^w(\beta, \Lambda|u)(h) = w\, l_\Lambda(\beta, \Lambda|u)(h).$$

By the definition of $(\hat{\beta}, \hat{\Lambda})$, we have $\mathbb{P}_n\{l_\beta^w(\hat{\beta}, \hat{\Lambda}|\hat{u})\} = 0$ and $\mathbb{P}_n\{l_\Lambda^w(\hat{\beta}, \hat{\Lambda}|\hat{u})(h)\} = 0$. Also, $\mathbb{P}\{l_\beta^w(\beta_0, \Lambda_0|u)\} = \mathbb{P}\{l_\beta(\beta_0, \Lambda_0|u)\} = 0$ and $\mathbb{P}\{l_\Lambda^w(\beta_0, \Lambda_0|u)(h)\} = \mathbb{P}\{l_\Lambda(\beta_0, \Lambda_0|u)(h)\} = 0$. Therefore,

$$\begin{aligned} & n^{1/2}[\mathbb{P}_n\{l_\beta^w(\hat{\beta}, \hat{\Lambda}|\hat{u})\} - \mathbb{P}\{l_\beta(\hat{\beta}, \hat{\Lambda}|u)\}] \\ & \quad = -n^{1/2}[\mathbb{P}\{l_\beta(\hat{\beta}, \hat{\Lambda}|u)\} - \mathbb{P}\{l_\beta(\beta_0, \Lambda_0|u)\}] \end{aligned} \tag{8.1}$$

and

$$\begin{aligned} & n^{1/2}[\mathbb{P}_n\{l_\Lambda^w(\hat{\beta}, \hat{\Lambda}|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\hat{\beta}, \hat{\Lambda}|u)(h)\}] \\ & \quad = -n^{1/2}[\mathbb{P}\{l_\Lambda(\hat{\beta}, \hat{\Lambda}|u)(h)\} - \mathbb{P}\{l_\Lambda(\beta_0, \Lambda_0|u)(h)\}]. \end{aligned} \tag{8.2}$$

We first consider $\mathbb{P}_n\{l_\beta^w(\beta, \Lambda|\hat{u})\} - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\}$ and $\mathbb{P}_n\{l_\Lambda^w(\beta, \Lambda|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}$ for fixed (β, Λ). Define the functions $H(t) = E[\exp(u)I(\tau \geq t)]$, $R(t) = H(t)\Lambda_{0h}(t)$, $Q(t) = \int_0^t H(s)\mathrm{d}\Lambda_{0h}(s)$, and for $i = 1, \ldots, n$,

$$b_i(t) = \sum_{k=1}^{K_i}\left\{\int_t^{\tau_0} \frac{I(U_{ik} \leq s \leq \tau_i)}{R^2(s)}\mathrm{d}Q(s) - \frac{I(t \leq U_{ik} \leq \tau_0)}{R(U_{ik})}\right\}.$$

In addition, for $i = 1, \ldots, n$, define

$$e_i = -\int \frac{\tilde{w}\tilde{z}kb_i(\tau)}{\Lambda_{0h}(\tau)}\mathrm{d}\mathcal{P}(\tilde{w}, \tilde{z}, k, \tau) + \tilde{w}_i\tilde{z}_i\{K_i\Lambda_{0h}^{-1}(\tau_i) - \exp(\gamma^T\tilde{z}_i)\},$$

where $\tilde{z}_i = (1, z_i^T)^T$, $\gamma = (\log\{E[\exp(u)]\}, \alpha^T)^T$, $\tilde{w}_i$ is the weight given in the estimating equations for α, and $\mathcal{P}(\cdot)$ denotes the joint probability measure of $(\tilde{w}, \tilde{z}, K, \tau)$. From Wang et al. [29], we have $\hat{\Lambda}_h(t) - \Lambda_{0h}(t) = n^{-1}\sum_{i=1}^n \Lambda_{0h}(t)b_i(t) + o_p(n^{-1/2})$ for $\inf\{s : \Lambda_{0h}(s) > 0\} < t < \tau_0$ and $\hat{\alpha} - \alpha_0 = n^{-1}\sum_{i=1}^n f_i(\alpha_0) + o_p(n^{-1/2})$, where $f_i(\alpha) = E[-\partial e_1/\partial\gamma]^{-1}e_i$ without the first entry. Define the function $u(\alpha, \Lambda_h; \tau, K, z) = \log\{K/[\Lambda_h(\tau)\exp(\alpha^T z)]\}$. Then $\hat{u} = u(\hat{\alpha}, \hat{\Lambda}_h; \tau, K, z)$. Furthermore, define

$$l_{\beta\alpha}(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z)) = \frac{\partial}{\partial\alpha}l_\beta(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z))$$

and

$$l_{\beta\Lambda_h}(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z)) = \frac{\partial}{\partial s}l_\beta(\beta, \Lambda|u(\alpha, s; \tau, K, z))\Big|_{s=\Lambda_h(\tau)}.$$

Then, we have

$$\begin{aligned}
&\mathbb{P}_n\{l^w_\beta(\beta, \Lambda|\hat{u})\} - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\}\\
&\quad = \mathbb{P}_n\{l^w_\beta(\beta, \Lambda|u(\hat{\alpha}, \hat{\Lambda}_h; \tau, K, z))\} - \mathbb{P}_n\{l^w_\beta(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau, K, z))\}\\
&\qquad + \mathbb{P}_n\{l^w_\beta(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau, K, z))\} - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\}\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\Big\{\mathbb{P}\{l_{\beta\alpha}(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i))f_i(\alpha_0)\}\\
&\qquad + \mathbb{P}\{l_{\beta\Lambda_h}(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i))\Lambda_{0h}(\tau_i)b_i(\tau_i)\}\\
&\qquad + w_i\, l_\beta(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i)) - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\}\Big\} + o_p(n^{-1/2})\\
&\quad = \frac{1}{n}\sum_{i=1}^{n} c_{\beta i}(\beta, \Lambda) + o_p(n^{-1/2}).
\end{aligned} \tag{8.3}$$

The $c_{\beta i}(\beta, \Lambda)$'s are independent random variables because $c_{\beta i}(\beta, \Lambda)$ depends only on the observed data from the ith subject. It follows from the law of large numbers that for fixed (β, Λ), $\mathbb{P}_n\{l^w_\beta(\beta, \Lambda|\hat{u})\} - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\} \to 0$ almost surely as $n \to \infty$. Furthermore, by the central limit theorem, $n^{1/2}[\mathbb{P}_n\{l^w_\beta(\beta, \Lambda|\hat{u})\} - \mathbb{P}\{l_\beta(\beta, \Lambda|u)\}]$ converges in distribution to a zero-mean normal random vector. Similarly, we can derive the asymptotic properties of $\mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}$. In particular, define

$$l_{\Lambda\alpha}(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z))(h) = \frac{\partial}{\partial\alpha} l_\Lambda(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z))(h)$$

and

$$l_{\Lambda\Lambda_h}(\beta, \Lambda|u(\alpha, \Lambda_h; \tau, K, z))(h) = \frac{\partial}{\partial s} l_\Lambda(\beta, \Lambda|u(\alpha, s; \tau, K, z))(h)\Big|_{s=\Lambda_h(\tau)}.$$

Then, we have

$$\begin{aligned}
&\mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}\\
&\quad = \mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|u(\hat{\alpha}, \hat{\Lambda}_h; \tau, K, z))(h)\} - \mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau, K, z))(h)\}\\
&\qquad + \mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau, K, z))(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}\\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{P}\{[l_{\Lambda\alpha}(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i))(h)]f_i(\alpha_0)\}\\
&\qquad + \mathbb{P}\{[l_{\Lambda\Lambda_h}(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i))(h)]\Lambda_{0h}(\tau_i)b_i(\tau_i)\}\\
&\qquad + w_i\, [l_\Lambda(\beta, \Lambda|u(\alpha_0, \Lambda_{0h}; \tau_i, K_i, z_i))(h)] - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}\} + o_p(n^{-1/2})\\
&\quad = \frac{1}{n}\sum_{i=1}^{n} c_{\Lambda i}(\beta, \Lambda)(h) + o_p(n^{-1/2}).
\end{aligned} \tag{8.4}$$

The $c_{\Lambda i}(\beta, \Lambda)(h)$'s are independent random variables because $c_{\Lambda i}(\beta, \Lambda)(h)$ depends only on the observed data from the ith subject. By the law of large numbers, $\mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\} \to 0$ almost surely as $n \to \infty$, for fixed (β, Λ). By the central limit theorem, $n^{1/2}[\mathbb{P}_n\{l^w_\Lambda(\beta, \Lambda|\hat{u})(h)\} - \mathbb{P}\{l_\Lambda(\beta, \Lambda|u)(h)\}]$ converges in distribution to a zero-mean normal random vector.

On the other hand, arguing as in the proof of Theorem 2 of Zeng et al. [33], we can show that

$$\begin{aligned} &-[\mathbb{P}\{l_\beta(\hat{\beta}, \hat{\Lambda}|u)\} - \mathbb{P}\{l_\beta(\beta_0, \Lambda_0|u)\}] + [\mathbb{P}\{l_\Lambda(\hat{\beta}, \hat{\Lambda}|u)(h^*)\} - \mathbb{P}\{l_\Lambda(\beta_0, \Lambda_0|u)(h^*)\}] \\ &\quad = E[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^T](\hat{\beta} - \beta_0) + O_p(\|\hat{\beta} - \beta_0\|^2 + n^{-2/3}), \end{aligned} \tag{8.5}$$

where $l_\beta = l_\beta(\beta_0, \Lambda_0|u)$, $l_\Lambda(h^*) = l_\beta(\beta_0, \Lambda_0|u)(h^*)$, and h^* is the least favourable direction, a $(p+1)$-vector with components in $L_2([0, \tau_0])$, that solves the normal equation $l^*_\Lambda l_\Lambda(h^*) = l^*_\Lambda l_\beta$ with l^*_Λ being the adjoint operator of l_Λ. The existence of h^* can be established as in Zeng et al. [33]. From Eqs. (8.3)–(8.5), the difference between (8.1) and (8.2) yields

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^n\{c_{\beta i}(\hat{\beta}, \hat{\Lambda}) - c_{\Lambda i}(\hat{\beta}, \hat{\Lambda})(h^*)\} + o_p(1) \\ &\quad = n^{1/2}E[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^T](\hat{\beta} - \beta_0) + O_p(n^{1/2}\|\hat{\beta} - \beta_0\|^2 + n^{-1/6}). \end{aligned} \tag{8.6}$$

The left-hand side of (8.6) can be written as $\mathbb{G}_n\{c_\beta(\hat{\beta}, \hat{\Lambda}) - c_\Lambda(\hat{\beta}, \hat{\Lambda})(h^*)\} + o_p(1)$. As argued in Zeng et al. [26], we can show that $h^*(t)$ is continuously differentiable on $[0, \tau_0]$, and further we are able to prove that $c_\beta(\hat{\beta}, \hat{\Lambda}) - c_\Lambda(\hat{\beta}, \hat{\Lambda})(h^*)$ belongs to a Donsker class and converges in the $L_2(\mathbb{P})$-norm to $c_\beta - c_\Lambda(h^*)$, where c_β and $c_\Lambda(h^*)$ are evaluated at (β_0, Λ_0). In addition, it is easy to show via proof by contradiction that the matrix $E[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^T]$ is invertible. Therefore, (8.6) entails $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$ and yields

$$n^{1/2}(\hat{\beta} - \beta_0) = \left(E[\{l_\beta - l_\Lambda(h^*)\}\{l_\beta - l_\Lambda(h^*)\}^T]\right)^{-1} \mathbb{G}_n\{c_\beta - c_\Lambda(h^*)\} + o_p(1).$$

This implies that $n^{1/2}(\hat{\beta} - \beta_0)$ converges to a zero-mean normal random vector. □

References

1. Chen, K., Jin, Z., Ying, Z.: Semiparametric analysis of transformation models with censored data. Biometrika **89**, 659–668 (2002)
2. Chen, K., Lo, S.H.: Case–cohort and case–control analysis with Cox's model. Biometrika **86**, 755–764 (1999)
3. Chen, K., Sun, L., Tong, X.: Analysis of cohort survival data with transformation model. Stat. Sin. **22**, 489–508 (2012)

4. Chen, Y.H., Zucker, D.M.: Case–cohort analysis with semiparametric transformation models. J. Stat. Plan. Inference **139**, 3706–3717 (2009)
5. Cheng, S.C., Wei, L.J., Ying, Z.: Analysis of transformation models with censored data. Biometrika **82**, 835–845 (1995)
6. Du, M., Zhou, Q., Zhao, S., Sun, J.: Regression analysis of case–cohort studies in the presence of dependent interval censoring. J. Appl. Stat. **48**(5), 846–865 (2021)
7. Fine, J.P., Ying, Z., Wei, L.J.: On the linear transformation model for censored data. Biometrika **85**, 980–986 (1998)
8. Fong, Y., Shen, X., Ashley, V.C., et al.: 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. J. Infect. Dis. **217**, 1280–1288 (2018)
9. Gilbert, P.B., Peterson, M.L., Follmann, D., Hudgens, M.G., Francis, D.P., Gurwith, M., Heyward, W.L., Jobes, D.V., Popovic, V., Self, S.G., et al.: Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. J. Infect. Dis. **191**, 666–677 (2005)
10. Hammer, S.M., Sobieszczyk, M.E., Janes, H., et al.: HVTN 505 study team: efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. N. Engl. J. Med. **369**, 2083–2092 (2013)
11. Huang, X., Wolfe, R.: A frailty model for informative censoring. Biometrics **58**, 510–520 (2002)
12. Janes, H.E., Cohen, K.W., Frahm, N., et al.: Higher T-cell responses induced by DNA/rAd5 HIV-1 preventive vaccine are associated with lower HIV-1 infection risk in an efficacy trial. J. Infect. Dis. **215**, 1376–1385 (2017)
13. Jewell, N.P., van der Laan, M.J.: Case–control current status data. Biometrika **91**, 529–541 (2004)
14. Kang, S., Cai, J.: Marginal hazards model for case–cohort studies with multiple disease outcomes. Biometrika **96**, 887–901 (2009)
15. Keogh, R.H., White, I.R.: Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. Stat. Med. **32**, 4021–4043 (2013)
16. Kim, S., Cai, J., Lu, W.: More efficient estimators for case–cohort studies. Biometrika **100**, 695–708 (2013)
17. Li, S., Hu, T., Zhao, S., Sun, J.: Regression analysis of multivariate current status data with semiparametric transformation frailty models. Stat. Sin. **30**, 1117–1134 (2020)
18. Li, Z., Nan, B.: Relative risk regression for current status data in case–cohort studies. Can. J. Stat. **39**, 557–577 (2011)
19. Lu, W.B., Liu, M.: On estimation of linear transformation models with nested case–control sampling. Lifetime Data Anal. **18**, 80–93 (2012)
20. Lu, W.B., Tsiatis, A.A.: Semiparametric transformation models for the case–cohort study. Biometrika **93**, 207–214 (2006)
21. Ma, L., Hu, T., Sun, J.: Cox regression analysis of dependent interval-censored failure time data. Comput. Stat. Data Anal. **103**, 79–90 (2016)
22. Ma, S., Kosorok, M.R.: Robust semiparametric M-estimation and the weighted bootstrap. J. Multivar. Anal. **96**, 190–217 (2005)
23. Marti, H., Chavance, M.: Multiple imputation analysis of case–cohort studies. Stat. Med. **30**, 1595–1607 (2011)
24. Prentice, R.L.: A case–cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika **73**, 1–11 (1986)
25. Self, S.G., Prentice, R.L.: Asymptotic distribution theory and efficiency results for case–cohort studies. Ann. Stat. **16**, 64–81 (1988)
26. Sun, J.: A nonparametric test for current status data with unequal censoring. J. R. Stat. Soc. B **61**, 243–250 (1999)
27. Sun, J.: The Statistical Analysis of Interval-Censored Failure Time Data. Springer, New York (2006)
28. Wang, L.M., McMahan, C.S., Hudgens, M.G., Qureshi, Z.P.: A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. Biometrics **72**, 222–231 (2016)
29. Wang, M.C., Qin, J., Chiang, C.T.: Analyzing recurrent event data with informative censoring. J. Am. Stat. Assoc. **96**, 1057–1065 (2001)
30. Wang, P., Zhao, H., Du, M.Y., Sun, J.: Inference on semiparametric transformation model with general interval-censored failure time data. J. Nonparametr. Stat. **30**, 753–758 (2018)

31. Wang, P., Zhao, H., Sun, J.: Regression analysis of case K interval-censored failure time data in the presence if informative censoring. Biometrics **72**, 1103–1112 (2016)
32. Zeng, D., Gao, F., Lin, D.Y.: Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. Biometrika **104**, 505–525 (2017)
33. Zeng, D., Mao, L., Lin, D.Y.: Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika **103**, 253–271 (2016)
34. Zhao, X., Zhou, J., Sun, L.: Semiparametric transformation models with time-varying coefficients for recurrent and terminal events. Biometrics **67**, 401–414 (2011)
35. Zhou, Q., Zhou, H., Cai, J.: Case–cohort studies with interval-censored failure time data. Biometrika **104**, 17–29 (2017)