Statistical Methods in Medical Research
XX(X):2–18

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

# Improving estimation efficiency of case-cohort studies with interval-censored failure time data

**Qingning Zhou[1] and Kin Yau Wong[2]**

**Abstract**

The case-cohort design is a commonly used cost-effective sampling strategy for large cohort studies, where some covariates are expensive to measure or obtain. In this paper, we consider regression analysis under a case-cohort study with interval-censored failure time data, where the failure time is only known to fall within an interval instead of being exactly observed. A common approach to analyze data from a case-cohort study is the inverse probability weighting approach, where only subjects in the case-cohort sample are used in estimation, and the subjects are weighted based on the probability of inclusion into the case-cohort sample. This approach, though consistent, is generally inefficient as it does not incorporate information outside the case-cohort sample. To improve efficiency, we first develop a sieve maximum weighted likelihood estimator under the Cox model based on the case-cohort sample, and then propose a procedure to update this estimator by using information in the full cohort. We show that the update estimator is consistent, asymptotically normal, and more efficient than the original estimator. The proposed method can flexibly incorporate auxiliary variables to improve estimation efficiency. A weighted bootstrap procedure is employed for variance estimation. Simulation results indicate that the proposed method works well in practical situations. An application to a Phase 3 HIV vaccine efficacy trial is provided for illustration.

**Keywords**

Cox model, Sieve estimation, Two-phase sampling, Update estimator, Weighted bootstrap.

## 1 Introduction

Two-phase sampling is a commonly used sampling technique that aims at cost reduction and improvement of estimation efficiency.[1–3] Typically, at Phase I, a large random sample is drawn from a target population and information on variables that are cheap or easy to measure is collected. These variables could be outcomes, cheap covariates, or auxiliary variables that are correlated with expensive exposures not available at Phase I. These variables can be used to formulate strata within the Phase I sample. Then at Phase II, a subsample is drawn from each stratum to obtain the variables that are expensive or difficult to measure. The formulation of strata seeks either to oversample subjects with important Phase I variables, or to effectively sample subjects with targeted Phase II variables, or both. In consequence, two-phase sampling achieves efficient access to important variables with less cost.

The case-cohort design is a widely used two-phase sampling design in epidemiological and biomedical studies where the outcomes of interest are times to some rare events, such as HIV infection and the onset of coronary heart disease, and some covariates are expensive to collect or measure. For example, when covariate measurements involve expensive bioassay, genetic measurements, or labor-intensive chart review, it may be economically infeasible to conduct the measurements for all subjects in the study cohort. The case-cohort design, in which only the cases and a subcohort of subjects are selected for the expensive covariate measurements, aims to yield more efficient inference under a certain budget constraint.

The case-cohort design has been adopted for some major biomedical studies, such as the Atherosclerosis Risk in Communities (ARIC) study conducted in four field centers in the United States.[4] The ARIC study is an ongoing longitudinal epidemiological study, where over 15,000 subjects were recruited in 1987 and were periodically examined thereafter. The subjects were followed for disease outcomes of interest, such as coronary heart disease, stroke, diabetes, hypertension, and death. Some covariates, such as high-sensitivity C-reactive protein, DNA alterations, DNA methylation, and metabolites, were only obtained for case-cohort samples.[5] The case-cohort design has also been adopted in major HIV studies. For example, the two Antibody Mediated Prevention trials conducted by HIV Vaccine Trials Network (HVTN), HVTN 703 in sub-Saharan Africa and HVTN 704 in Americas and Europe, were designed to investigate whether a broadly neutralizing antibody (bnAb), VRC01, can be used to prevent HIV-1 acquisition.[6;7] In these trials, the VRC01 measurements were obtained only for a case-cohort sample. Another example is

[1]Department of Mathematics and Statistics, University of North Carolina at Charlotte, USA. [2]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong.

**Corresponding author:**
Qingning Zhou, Department of Mathematics and Statistics, University of North Carolina at Charlotte, USA.
Email: qzhou8@charlotte.edu

the two Phase 3 HIV vaccine efficacy trials, VAX003 in Thailand and VAX004 in North America and the Netherlands.[8] It was of interest to investigate the association between antibody responses to a vaccine containing rgp120 antigens and the incidence of HIV infection. Since the antibody levels were measured by complex assays that are expensive to perform, the case-cohort design was employed to reduce study cost.

Since the initial proposal by Prentice,[9] there has been extensive research in the statistical analysis of case-cohort design and its variations. Under the Cox model, Prentice[9] and Self and Prentice[10] proposed a pseudo-likelihood method for estimation and inference; Chen and Luo[11] developed an estimating equation approach; Marti and Chavance[12] proposed a multiple imputation method; Scheike and Martinussen[13] and Zeng and Lin[3] considered maximum likelihood estimation. Moreover, Kang and Cai,[14] Kim et al.[15] and Kim et al.[5] developed estimating equation approaches for case-cohort design with multiple outcomes. However, most of the existing methods for case-cohort design were developed for right-censored survival data.

In this paper, we consider the case-cohort design with interval-censored failure time data. Interval censoring occurs when the failure time of interest is only known to fall within an interval rather than being exactly observed. Interval-censored data commonly arise in epidemiological and biomedical studies that involve asymptomatic events, such as HIV infection, the onset of AIDS, and the onset of diabetes. For example, in the ARIC study, an outcome of interest is time to the onset of diabetes.[4] The study subjects were examined every three years, and they may miss some examinations or may not come at the scheduled times. If a participant was observed to have diabetes, it was only known that the onset of diabetes occurred between two consecutive visits. Also, in the HIV vaccine trials mentioned above,[6–8] the subjects were tested for HIV infection only at discrete clinic visits and thus only an interval given by the last HIV negative test date and the first HIV positive test date was observed for the HIV infection time. Interval censoring complicates the likelihood and poses challenges for estimation and inference.

Research on the case-cohort design with interval-censored data is limited. Li et al.[16] and Li and Nan[17] studied case-cohort design with grouped survival data and current status data, respectively, which are special cases of interval-censored data. Zhou et al.[18] considered case-cohort design with general interval-censored data and proposed a sieve maximum weighted likelihood method based on inverse probability weighting (IPW). Zhou et al.[19] studied case-cohort design with multiple interval-censored outcomes and developed an IPW method with weights that incorporate information from multiple events. Du et al.[20] considered case-cohort design with informatively interval-censored data, where the monitoring times depend on the failure time, and also employed an IPW method to handle the sampling bias induced by the case-cohort design. It is well known that the IPW method only uses data in the case-cohort sample and is inefficient. We propose an update estimation procedure that improves the efficiency of the IPW estimator in Zhou et al.[18] by using information from the full cohort.

The main idea of the update estimation approach is to find an (asymptotically) mean-zero statistic that is potentially correlated with the original consistent estimator (say, the IPW estimator) and then construct an update estimator as the optimal linear combination of the original estimator with the mean-zero statistic. It can be shown that

the update estimator is still unbiased and at least as efficient as the original estimator. The update estimation approach, in its current general form for regression models, was first formulated by Chen and Chen,[21] and it has thereafter been employed under different settings involving incomplete or imprecise data.[22–27] In this paper, we propose an update estimation procedure to improve the efficiency of the IPW estimator given by Zhou et al.[18] for regression analysis of case-cohort study with interval-censored failure time data. Specifically, we assume a working regression model of the failure time given the cheap covariates and auxiliary variables, if available. We then fit the working model to the case-cohort sample and the full cohort data, respectively, to obtain two estimators of the same limit. We thus can find a mean-zero statistic by taking difference of the two estimators and construct an update estimator as the optimal linear combination of the IPW estimator with the mean-zero statistic.

The remainder of this paper is organized as follows. In Section 2, we describe the design, data structure and model assumptions. In Section 3, we review the IPW estimator given by Zhou et al.[18] In Section 4, we propose an update estimation procedure to improve the IPW estimator using information from the full cohort, and also establish the asymptotic properties of the proposed update estimator. Simulation studies and a real data application are given in Sections 5 and 6, respectively. We conclude in Section 7 with some discussions on extensions or directions for future research.

## 2 Design, Data and Model

Let $T$ be the failure time, $X$ be a vector of expensive covariates, and $Z$ be a vector of cheap covariates. For example, $X$ could consist of biomarkers ascertained by bioassay or genetic analysis, and $Z$ could consist of age and gender. Assume that $T$ follows the Cox model with the conditional cumulative hazard function given by

$$\Lambda(t \mid X, Z) = \Lambda(t) \exp(\beta^{\mathrm{T}} X + \gamma^{\mathrm{T}} Z), \tag{1}$$

where $\vartheta \equiv (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$ is a $d$-vector of regression parameters, and $\Lambda$ is the unspecified cumulative baseline hazard function. Also, let $X^*$ denote a vector of cheap auxiliary variables that could be available in practice and informative to the expensive covariate $X$. Assume that $T$ and $X^*$ are independent given $X$ and $Z$.

Suppose that we observe interval-censored failure time data. Let $U_1, \ldots, U_K$ denote the random examination times such that $0 = U_0 < U_1 < \cdots < U_K < U_{K+1} = \infty$, where $K$ is a random positive integer. Also, define $\Delta_k = I(U_{k-1} < T \leq U_k)$ for $k = 1, \ldots, K+1$, where $I(\cdot)$ is the indicator function. Then the interval-censored failure time data consists of $\{K, U_1, \ldots, U_K, \Delta_1, \ldots, \Delta_K\}$. We assume that the examination times are conditionally independent of the failure time given the covariates.

We consider a two-phase (generalized) case-cohort design based on a full cohort of size $n$. At Phase I, we observe the interval-censored failure time, the cheap covariates, and the auxiliary variables for all cohort members, denoted by $\{K_i, U_{i1}, \ldots, U_{iK_i}, \Delta_{i1}, \ldots, \Delta_{iK_i}, Z_i, X_i^*\}$ for $i = 1, \ldots, n$. At Phase II, we first select a subcohort via independent Bernoulli sampling with a known success probability $q_s \in (0, 1]$, and then select a subset of cases (i.e., subjects with $\sum_{k=1}^{K_i} \Delta_{ik} = 1$) outside

the subcohort also by Bernoulli sampling with a known success probability $q_c \in (0, 1]$. Note that if $q_c = 1$, then all cases are selected, and this is called a case-cohort design; if $q_c \in (0, 1)$, then this is usually referred to as a generalized case-cohort design. Let $\eta_i$ indicate whether the $i$th subject is selected into the subcohort and $\zeta_i$ indicate whether the $i$th subject is a selected case. Note that if $\zeta_i = 1$, then $\eta_i = 0$ and $\sum_{k=1}^{K_i} \Delta_{ik} = 1$. The expensive covariates are observed for subjects in the subcohort and for the selected cases outside the subcohort, that is, for the $i$th subject, $X_i$ is observed if $\eta_i = 1$ or $\zeta_i = 1$. Let $\xi_i$ indicate whether $X_i$ is observed. Then the observed data can be represented by

$$O_i^{\xi} = \{K_i,\, U_{i1}, \ldots, U_{iK_i},\, \Delta_{i1}, \ldots, \Delta_{iK_i},\, Z_i,\, X_i^*,\, \xi_i X_i,\, \xi_i\}, \quad i = 1, \ldots, n,$$

where the notation $\xi_i X_i$ means that if the $i$th subject is selected at Phase II, then $\xi_i = 1$ and $X_i$ is observed; otherwise, $\xi_i = 0$ and $X_i$ is not observed.

## 3 Sieve Maximum Weighted Likelihood Estimator

Let $\theta = (\vartheta, \Lambda)$. For estimation, we first consider the weighted log-likelihood function

$$l_n^w(\theta) = \sum_{i=1}^{n} w_i \Bigg( \sum_{k=1}^{K_i+1} \Delta_{ik} \log \Big[ \exp\left\{-\Lambda(U_{i,k-1}) \exp(\beta^{\mathrm{T}} X_i + \gamma^{\mathrm{T}} Z_i)\right\} - \exp\left\{-\Lambda(U_{ik}) \exp(\beta^{\mathrm{T}} X_i + \gamma^{\mathrm{T}} Z_i)\right\} \Big] \Bigg),$$

where the weight $w_i$ is defined as

$$w_i = \frac{\xi_i}{\pi_q(\Delta_i)} = \frac{\xi_i}{\left(1 - \sum_{k=1}^{K_i} \Delta_{ik}\right) q_s + \left(\sum_{k=1}^{K_i} \Delta_{ik}\right) \{q_s + (1 - q_s) q_c\}}.$$

The weighted likelihood approach is commonly used to handle sampling bias in two-phase studies.[1;2] In this case, the weight $w_i$ is set to be the inverse probability of the $i$th subject being selected into the (generalized) case-cohort sample. Specifically, if the $i$th subject is a case, then the probability of being selected is 1 or $q_s + (1 - q_s)q_c$, under the case-cohort design or the generalized case-cohort design, respectively. If the $i$th subject is not a case, then the probability of being selected is $q_s$ under both designs. In the above, we assume that $q_s$ and $q_c$ are known for simplicity; our approach still works if they are unknown and replaced by consistent estimators.

To estimate the unknown function $\Lambda$, Zhou et al.[18] proposed a sieve method based on Bernstein polynomials. Specifically, the sieve space is defined as

$$\Theta_n = \mathcal{B} \otimes \mathcal{M}_n\,, \tag{2}$$

where $\mathcal{B}$ is a compact set in $R^d$ and

$$\mathcal{M}_n = \left\{ \Lambda_n(t) = \sum_{k=0}^{m} \phi_j B_j(t, m, \sigma, \tau) : \;\; \phi_m \geq \ldots \geq \phi_1 \geq \phi_0 \geq 0,\; \sum_{k=0}^{m} |\phi_j| \leq M_n \right\}$$

with $B_j(t, m, \sigma, \tau)$ being Bernstein basis polynomials of degree $m$, i.e.,

$$B_j(t, m, \sigma, \tau) = \binom{m}{j} \left(\frac{t-\sigma}{\tau-\sigma}\right)^j \left(1 - \frac{t-\sigma}{\tau-\sigma}\right)^{m-j}, \quad j = 0, \ldots, m.$$

Here $\sigma$ and $\tau$ are the lower and upper bounds of the examination times, respectively, with $0 < \sigma < \tau < \infty$. The sieve maximum weighted likelihood estimator $\hat{\theta}_n = (\hat{\vartheta}_n, \hat{\Lambda}_n)$ is defined as the value of $\theta$ that maximizes the weighted log-likelihood function $l_n^w$ over $\Theta_n$. We suggest to select the degree of Bernstein polynomials $m$ based on the AIC. In particular, we consider several candidate values of $m$ and select the one that minimizes

$$\text{AIC} = -2l_n^w(\hat{\theta}_n) + 2(p + m + 1).$$

At the selected $m$, we perform maximization of the weighted log-likelihood function, with respect to the regression parameters and Bernstein coefficients, using the BFGS quasi-Newton algorithm.

It is well known that the IPW estimator is inefficient. To improve estimation efficiency, we consider an update approach that utilizes the available information in the full cohort by fitting a working model relating the cheap covariates or auxiliary variables to the failure time. It can be shown that the update estimator is guaranteed to be asymptotically at least as efficient as the original IPW estimator.

## 4 Proposed Update Estimator

Consider a working Cox model for $T$ given $X^*$ and $Z$ with the conditional cumulative hazard function

$$\Lambda^*(t \mid X^*, Z) = \Lambda^*(t) \exp(\beta^{*\mathrm{T}} X^* + \gamma^{*\mathrm{T}} Z), \tag{3}$$

where $\vartheta^* \equiv (\beta^{*\mathrm{T}}, \gamma^{*\mathrm{T}})^{\mathrm{T}}$ is a $d^*$-vector of regression parameters, and $\Lambda^*$ is the unspecified cumulative baseline hazard function. Note that we can consider a working Cox model given $Z$ only, if $X^*$ is not available.

We first estimate the working Cox model (3) using the Bernstein-polynomial-based sieve method similarly as the above but with covariates $X^*$ and $Z$ instead. Let $\hat{\theta}_n^* = (\hat{\vartheta}_n^*, \hat{\Lambda}_n^*)$ denote the sieve maximum weighted likelihood estimator of $\theta^* = (\vartheta^*, \Lambda^*)$ based on the case-cohort sample. Since $X^*$ and $Z$ are available for all subjects in the cohort, we can also obtain the sieve maximum likelihood estimator of $\theta^* = (\vartheta^*, \Lambda^*)$, denoted by $\bar{\theta}_n^* = (\bar{\vartheta}_n^*, \bar{\Lambda}_n^*)$, based on the full cohort. Let $\theta_0 = (\vartheta_0, \Lambda_0)$ denote the true value of $\theta = (\vartheta, \Lambda)$ in model (1). Let $\Sigma = [\Sigma_{11}, \Sigma_{12}; \Sigma_{21}, \Sigma_{22}]$ be the covariance matrix of the limiting distribution of $\left(\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)^{\mathrm{T}}, \sqrt{n}(\hat{\vartheta}_n^* - \bar{\vartheta}_n^*)^{\mathrm{T}}\right)^{\mathrm{T}}$, and let $\hat{\Sigma} = [\hat{\Sigma}_{11}, \hat{\Sigma}_{12}; \hat{\Sigma}_{21}, \hat{\Sigma}_{22}]$ denote a consistent estimator of $\Sigma$. We define the update estimator of $\vartheta$ as

$$\bar{\vartheta}_n = \hat{\vartheta}_n - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}(\hat{\vartheta}_n^* - \bar{\vartheta}_n^*).$$

The asymptotic distribution of the proposed update estimator $\bar{\vartheta}_n$ is given in the following theorem.

**Theorem 1.** *Under Conditions (C1)–(C6) given in the Appendix, we have*

$$\sqrt{n}(\bar{\vartheta}_n - \vartheta_0) = \sqrt{n}(\hat{\vartheta}_n - \vartheta_0) - \Sigma_{12}\Sigma_{22}^{-1}\sqrt{n}(\hat{\vartheta}_n^* - \bar{\vartheta}_n^*) + o_p(1) \to N(0, \Psi)$$

*in distribution with* $\Psi = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, *where*

$$\Sigma_{11} = I(\vartheta_0)^{-1} E\left\{ \frac{1}{\pi_q(\Delta)} \left[ l(\vartheta_0, \Lambda_0; O) \right]^{\otimes 2} \right\} I(\vartheta_0)^{-1},$$

$$\Sigma_{22} = I^*(\vartheta_0^*)^{-1} E\left\{ \frac{1-\pi_q(\Delta)}{\pi_q(\Delta)} \left[ l^*(\vartheta_0^*, \Lambda_0^*; O^*) \right]^{\otimes 2} \right\} I^*(\vartheta_0^*)^{-1},$$

$$\Sigma_{12} = \Sigma_{21}^T = I(\vartheta_0)^{-1} E\left\{ \frac{1-\pi_q(\Delta)}{\pi_q(\Delta)} \, l(\vartheta_0, \Lambda_0; O) \, l^*(\vartheta_0^*, \Lambda_0^*; O^*)^T \right\} I^*(\vartheta_0^*)^{-1},$$

*and* $l(\vartheta_0, \Lambda_0; O)$, $I(\vartheta_0)$, $l^*(\vartheta_0^*, \Lambda_0^*; O^*)$ *and* $I^*(\vartheta_0^*)$ *are defined in the Appendix.*

**Remark 1.** The asymptotic covariance matrix of $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ is $\Sigma_{11}$, whereas the asymptotic covariance matrix of $\sqrt{n}(\bar{\vartheta}_n - \vartheta_0)$ is $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Thus, $\bar{\vartheta}_n$ is guaranteed to be asymptotically at least as efficient as $\hat{\vartheta}_n$.

**Remark 2.** The proof of Theorem 1 is sketched in the Appendix. The main technical challenge is to establish the existence of the efficient scores $l$ and $l^*$ under the Cox model (1) and the working model (3), respectively. It can be shown that the efficient scores solve some integral equations but do not have explicit forms. In fact, generally for semiparametric regression analysis of interval-censored data, the efficient score and thus the asymptotic covariance matrix do not have explicit forms.[28–30]

Since there is no closed-form expression for the covariance matrix $\Sigma$, we propose to estimate $\Sigma$ using a weighted bootstrap method.[31] Specifically, let $\{u_1, \ldots, u_n\}$ denote $n$ independent realizations of a bounded positive random variable $u$ satisfying $E(u) = \text{var}(u) = 1$. We use the exponential distribution with mean one in the simulation studies and real data analysis. Define the new weights $w_i^b = u_i w_i$ for $i = 1, \ldots, n$. Let $\hat{\theta}_n^b = (\hat{\vartheta}_n^b, \hat{\Lambda}_n^b)$ be the sieve maximum weighted likelihood estimator that maximizes the new weighted log-likelihood function $l_n^{w^b}$ over $\Theta_n$, where $l_n^{w^b}$ is obtained by replacing $w_i$ with $w_i^b$ in $l_n^w$. We generate $B$ samples of $\{u_1, \ldots, u_n\}$ and obtain the corresponding $\hat{\vartheta}_n^b$ as well as $\hat{\vartheta}_n^{*b}$ and $\bar{\vartheta}_n^{*b}$ similarly for $b = 1, \ldots, B$. Then we take $\hat{\Sigma}$ as the sample covariance matrix of $\left(\sqrt{n}(\hat{\vartheta}_n^b - \vartheta_0)^{\mathrm{T}}, \sqrt{n}(\hat{\vartheta}_n^{*b} - \bar{\vartheta}_n^{*b})^{\mathrm{T}}\right)^{\mathrm{T}}$. The asymptotic covariance matrix of $\sqrt{n}(\bar{\vartheta}_n - \vartheta_0)$ can be estimated by $\hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$, which is consistent according to Ma and Kosorok.[31]

## 5 Simulation Studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed method. Assume that the failure time $T$ follows the Cox model

$$\Lambda(t \mid X, Z) = \Lambda(t) \exp(\beta X + \gamma Z),$$

where $\Lambda(t) = 0.2t^2$, $\beta = 0$ or $0.3$, and $\gamma = 0.5$. We consider three setups for the covariates: (i) $X \sim N(0,1)$ and $Z$ is empty; (ii) $X \sim Ber(0.5)$ and $Z$ is empty; and (iii) $(X, Z)$ follow a bivariate normal distribution with zero mean and covariance matrix $[1, 0.2; 0.2, 1]$. For continuous $X$, we generate the auxiliary variable $X^* = X + e$, where $e \sim N(0, \sigma^2)$ and $\sigma = 0.30$, $0.86$, or $1.70$, such that the correlation between $X$ and $X^*$ is $\rho = 0.95$, $0.75$, or $0.50$, respectively. For binary $X$, we generate $X^* = MX + (1 - M)(1 - X)$, where $M \sim Ber(1 - p_m)$ and is independent of $X$, and we consider a misclassification rate of $p_m = 0.05$, $0.10$, or $0.20$. To generate the examination times, we first define a set of scheduled examination times, $s_j = ju/(n_t + 1)$ for $j = 1, \ldots, n_t$, where $n_t$ is the total number of scheduled examinations, and $u$ is the end-of-study time. For the $i$th subject, the actual set of examination times are $\{s_j + \epsilon_{ij} : R_{ij} = 1, j = 1, \ldots, n_t\}$, where $\epsilon_{ij}$'s are i.i.d. $\mathrm{Unif}(-t_d/3, t_d/3)$ variables, $R_{ij}$'s are i.i.d. $\mathrm{Bernoulli}(0.8)$ variables, and $t_d = u/(n_t + 1)$. This is to mimic an actual follow-up study, where subjects may miss a scheduled visit and may also visit at a time different from the scheduled time. The number of scheduled examinations $n_t$ is taken as 12. The value of $u$ is chosen such that the case rate is $p_c = 0.1$, $0.2$, or $0.3$. We consider a case-cohort study for $p_c = 0.1$ or $0.2$ by taking all cases, and also consider a generalized case-cohort study for $p_c = 0.2$ or $0.3$ by taking a subsample of cases outside the subcohort with the sampling probability equal to $q_c = 0.5$. The subcohort is selected via independent Bernoulli sampling with success probability $q_s = 0.2$. The degree of Bernstein polynomials $m$ is selected from $1, 2, \ldots, 5$ based on the AIC. The weighted bootstrap procedure for variance estimation is based on $500$ samples. The sample size is $n = 1000$. The results are based on $1000$ replicates.

Tables 1–3 present the estimation results of the Euclidean parameters, including "Bias": the average point estimate minus the true parameter value, "SSD": the sample standard deviation of point estimates, "ESE": the estimated standard error based on weighted bootstrap, and "CP": the coverage proportion of the $95\%$ confidence interval based on normal approximation. The original IPW estimator of Zhou et al.,[18] denoted by ZZC, is included for comparison with the proposed update estimator. The relative efficiency of the proposed estimator with respect to ZZC, denoted by "RE", is given in Tables 1–3. For all setups considered, the proposed estimator is virtually unbiased, the estimated variance based on weighted bootstrap reflects the true variability, and the coverage proportion is close to the nominal level. In addition, the proposed estimator is more efficient than the ZZC estimator, and its efficiency gain for the estimation of $\beta$ increases with the case rate $p_c$ and with the association between $X$ and $X^*$. Also, when $p_c = 0.2$, relative efficiencies (RE) of the proposed method against ZZC under the generalized case-cohort study ($q_c = 0.5$) are higher than those under the case-cohort study ($q_c = 1$). This could be due to the presence of more out-of-sample information that we can borrow via the update approach under the generalized case-cohort study.

## 6 Application to the VAX004 Trial

We provide an illustrative example using the world's first Phase 3 HIV-1 vaccine efficacy trial, VAX004.[32] This trial was conducted in North America and the Netherlands in $5403$

**Table 1.** Simulation results with $X \sim N(0,1)$

| | | | | $\beta = 0$ | | | | | $\beta = 0.3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_c$ | $q_c$ | Method | $\rho$ | Bias | SSD | ESE | CP | RE | Bias | SSD | ESE | CP | RE |
| 0.10 | 1 | ZZC | | −0.003 | 0.121 | 0.118 | 0.95 | 1.00 | 0.004 | 0.127 | 0.121 | 0.95 | 1.00 |
| | | Proposed | 0.95 | −0.004 | 0.100 | 0.098 | 0.94 | 1.46 | −0.004 | 0.103 | 0.099 | 0.95 | 1.52 |
| | | | 0.75 | −0.003 | 0.108 | 0.106 | 0.95 | 1.26 | 0.000 | 0.113 | 0.110 | 0.95 | 1.26 |
| | | | 0.50 | −0.002 | 0.115 | 0.113 | 0.95 | 1.11 | 0.003 | 0.121 | 0.117 | 0.95 | 1.09 |
| 0.20 | 1 | ZZC | | 0.004 | 0.098 | 0.096 | 0.95 | 1.00 | 0.006 | 0.103 | 0.098 | 0.94 | 1.00 |
| | | Proposed | 0.95 | 0.002 | 0.074 | 0.072 | 0.94 | 1.72 | 0.001 | 0.078 | 0.073 | 0.93 | 1.78 |
| | | | 0.75 | 0.004 | 0.084 | 0.082 | 0.95 | 1.35 | 0.005 | 0.089 | 0.085 | 0.94 | 1.35 |
| | | | 0.50 | 0.004 | 0.092 | 0.090 | 0.95 | 1.13 | 0.006 | 0.098 | 0.092 | 0.94 | 1.13 |
| 0.20 | 0.5 | ZZC | | 0.009 | 0.109 | 0.109 | 0.96 | 1.00 | 0.009 | 0.113 | 0.111 | 0.94 | 1.00 |
| | | Proposed | 0.95 | 0.002 | 0.076 | 0.074 | 0.94 | 2.07 | 0.002 | 0.079 | 0.075 | 0.94 | 2.08 |
| | | | 0.75 | 0.004 | 0.090 | 0.089 | 0.95 | 1.48 | 0.006 | 0.094 | 0.091 | 0.95 | 1.45 |
| | | | 0.50 | 0.007 | 0.101 | 0.101 | 0.95 | 1.17 | 0.008 | 0.106 | 0.103 | 0.95 | 1.15 |
| 0.30 | 0.5 | ZZC | | 0.005 | 0.096 | 0.095 | 0.95 | 1.00 | 0.011 | 0.100 | 0.097 | 0.94 | 1.00 |
| | | Proposed | 0.95 | 0.001 | 0.063 | 0.061 | 0.94 | 2.27 | 0.003 | 0.067 | 0.063 | 0.94 | 2.26 |
| | | | 0.75 | 0.003 | 0.077 | 0.076 | 0.95 | 1.54 | 0.007 | 0.083 | 0.079 | 0.94 | 1.45 |
| | | | 0.50 | 0.004 | 0.088 | 0.087 | 0.95 | 1.19 | 0.009 | 0.094 | 0.089 | 0.94 | 1.14 |

Bias, average estimate minus true value; SSD, sample standard deviation; ESE, estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency of the proposed estimator with respect to the ZZC estimator.

**Table 2.** Simulation results with $X \sim Ber(0.5)$

| | | | | $\beta = 0$ | | | | | $\beta = 0.3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_c$ | $q_c$ | Method | $p_m$ | Bias | SSD | ESE | CP | RE | Bias | SSD | ESE | CP | RE |
| 0.10 | 1 | ZZC | | −0.004 | 0.221 | 0.233 | 0.97 | 1.00 | −0.001 | 0.228 | 0.236 | 0.96 | 1.00 |
| | | Proposed | 0.05 | −0.004 | 0.196 | 0.203 | 0.96 | 1.28 | −0.001 | 0.201 | 0.206 | 0.96 | 1.28 |
| | | | 0.10 | −0.004 | 0.202 | 0.209 | 0.96 | 1.20 | −0.001 | 0.207 | 0.213 | 0.96 | 1.21 |
| | | | 0.20 | −0.003 | 0.210 | 0.220 | 0.97 | 1.11 | −0.000 | 0.215 | 0.223 | 0.96 | 1.13 |
| 0.20 | 1 | ZZC | | −0.002 | 0.180 | 0.189 | 0.96 | 1.00 | −0.003 | 0.180 | 0.189 | 0.97 | 1.00 |
| | | Proposed | 0.05 | −0.002 | 0.145 | 0.149 | 0.96 | 1.55 | −0.003 | 0.145 | 0.151 | 0.96 | 1.54 |
| | | | 0.10 | −0.002 | 0.153 | 0.158 | 0.96 | 1.38 | −0.003 | 0.154 | 0.159 | 0.96 | 1.37 |
| | | | 0.20 | −0.001 | 0.166 | 0.172 | 0.95 | 1.18 | −0.002 | 0.166 | 0.173 | 0.96 | 1.18 |
| 0.20 | 0.5 | ZZC | | −0.004 | 0.204 | 0.215 | 0.96 | 1.00 | −0.004 | 0.206 | 0.216 | 0.96 | 1.00 |
| | | Proposed | 0.05 | −0.004 | 0.152 | 0.156 | 0.96 | 1.79 | −0.006 | 0.154 | 0.157 | 0.95 | 1.78 |
| | | | 0.10 | −0.005 | 0.164 | 0.170 | 0.96 | 1.55 | −0.006 | 0.166 | 0.171 | 0.96 | 1.54 |
| | | | 0.20 | −0.003 | 0.183 | 0.191 | 0.96 | 1.23 | −0.003 | 0.185 | 0.192 | 0.96 | 1.23 |
| 0.30 | 0.5 | ZZC | | −0.002 | 0.187 | 0.188 | 0.95 | 1.00 | −0.002 | 0.187 | 0.189 | 0.96 | 1.00 |
| | | Proposed | 0.05 | −0.003 | 0.131 | 0.131 | 0.95 | 2.04 | −0.003 | 0.134 | 0.132 | 0.94 | 1.95 |
| | | | 0.10 | −0.005 | 0.144 | 0.145 | 0.94 | 1.68 | −0.004 | 0.147 | 0.146 | 0.95 | 1.61 |
| | | | 0.20 | −0.003 | 0.163 | 0.165 | 0.95 | 1.31 | −0.002 | 0.166 | 0.166 | 0.95 | 1.26 |

Bias, average estimate minus true value; SSD, sample standard deviation; ESE, estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency of the proposed estimator with respect to the ZZC estimator.

HIV-1-uninfected volunteers, including 5095 non-injection drug-using men who have sex with men and 308 women at high risk for heterosexual acquisition of HIV-1. This is a randomized, double-blinded, placebo-controlled trial completed in 2003. In this trial, 3598 subjects received the recombinant glycoprotein 120 (rgp120) vaccine and 1805 had placebo. Immunizations were administered by intramuscular injection at months 0, 1, 6, 12, 18, 24 and 30. At each of these visits and at month 36, subjects were tested for HIV-1 infection by standard HIV-1 ELISA and confirmatory immunoblot. For each infected

**Table 3.** Simulation results with $X$ and $Z$ from a bivariate normal distribution

| | | | | $\beta = 0$ | | | | | $\gamma = 0.5$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_c$ | $q_c$ | Method | $\rho$ | Bias | SSD | ESE | CP | RE | Bias | SSD | ESE | CP | RE |
| 0.10 | 1 | ZZC | | −0.007 | 0.133 | 0.127 | 0.94 | 1.00 | 0.021 | 0.135 | 0.130 | 0.94 | 1.00 |
| | | Proposed | 0.95 | −0.006 | 0.106 | 0.102 | 0.94 | 1.56 | 0.003 | 0.110 | 0.100 | 0.93 | 1.53 |
| | | | 0.75 | −0.007 | 0.118 | 0.113 | 0.95 | 1.26 | 0.005 | 0.110 | 0.101 | 0.93 | 1.50 |
| | | | 0.50 | −0.007 | 0.127 | 0.121 | 0.95 | 1.09 | 0.006 | 0.111 | 0.103 | 0.93 | 1.48 |
| 0.20 | 1 | ZZC | | −0.003 | 0.105 | 0.102 | 0.95 | 1.00 | 0.012 | 0.101 | 0.104 | 0.95 | 1.00 |
| | | Proposed | 0.95 | −0.004 | 0.077 | 0.075 | 0.94 | 1.87 | 0.004 | 0.074 | 0.073 | 0.95 | 1.88 |
| | | | 0.75 | −0.004 | 0.089 | 0.087 | 0.95 | 1.37 | 0.005 | 0.074 | 0.074 | 0.95 | 1.86 |
| | | | 0.50 | −0.004 | 0.099 | 0.096 | 0.95 | 1.13 | 0.006 | 0.074 | 0.075 | 0.95 | 1.84 |
| 0.20 | 0.5 | ZZC | | −0.004 | 0.117 | 0.115 | 0.95 | 1.00 | 0.011 | 0.112 | 0.117 | 0.96 | 1.00 |
| | | Proposed | 0.95 | −0.004 | 0.078 | 0.076 | 0.94 | 2.24 | 0.005 | 0.074 | 0.073 | 0.95 | 2.30 |
| | | | 0.75 | −0.005 | 0.095 | 0.094 | 0.95 | 1.51 | 0.007 | 0.074 | 0.074 | 0.95 | 2.27 |
| | | | 0.50 | −0.005 | 0.107 | 0.106 | 0.95 | 1.18 | 0.008 | 0.075 | 0.076 | 0.94 | 2.22 |
| 0.30 | 0.5 | ZZC | | −0.002 | 0.104 | 0.099 | 0.94 | 1.00 | 0.009 | 0.098 | 0.100 | 0.94 | 1.00 |
| | | Proposed | 0.95 | −0.002 | 0.064 | 0.063 | 0.94 | 2.66 | 0.005 | 0.061 | 0.060 | 0.94 | 2.56 |
| | | | 0.75 | −0.003 | 0.081 | 0.079 | 0.94 | 1.65 | 0.006 | 0.062 | 0.062 | 0.94 | 2.52 |
| | | | 0.50 | −0.003 | 0.094 | 0.091 | 0.95 | 1.22 | 0.006 | 0.062 | 0.063 | 0.95 | 2.45 |
| | | | | $\beta = 0.3$ | | | | | $\gamma = 0.5$ | | | | |
| $p_c$ | $q_c$ | Method | $\rho$ | Bias | SSD | ESE | CP | RE | Bias | SSD | ESE | CP | RE |
| 0.10 | 1 | ZZC | | 0.001 | 0.133 | 0.130 | 0.95 | 1.00 | 0.021 | 0.136 | 0.131 | 0.94 | 1.00 |
| | | Proposed | 0.95 | −0.006 | 0.106 | 0.103 | 0.94 | 1.57 | 0.005 | 0.106 | 0.100 | 0.94 | 1.64 |
| | | | 0.75 | −0.002 | 0.119 | 0.115 | 0.94 | 1.24 | 0.005 | 0.108 | 0.102 | 0.94 | 1.58 |
| | | | 0.50 | 0.000 | 0.128 | 0.124 | 0.95 | 1.08 | 0.005 | 0.110 | 0.103 | 0.94 | 1.53 |
| 0.20 | 1 | ZZC | | 0.002 | 0.110 | 0.104 | 0.94 | 1.00 | 0.012 | 0.105 | 0.105 | 0.94 | 1.00 |
| | | Proposed | 0.95 | −0.000 | 0.078 | 0.076 | 0.94 | 1.97 | 0.004 | 0.076 | 0.073 | 0.94 | 1.90 |
| | | | 0.75 | 0.002 | 0.095 | 0.089 | 0.92 | 1.34 | 0.004 | 0.078 | 0.075 | 0.94 | 1.82 |
| | | | 0.50 | 0.003 | 0.105 | 0.097 | 0.93 | 1.10 | 0.004 | 0.079 | 0.076 | 0.94 | 1.76 |
| 0.20 | 0.5 | ZZC | | 0.001 | 0.122 | 0.116 | 0.93 | 1.00 | 0.010 | 0.116 | 0.117 | 0.94 | 1.00 |
| | | Proposed | 0.95 | −0.001 | 0.080 | 0.077 | 0.95 | 2.30 | 0.004 | 0.077 | 0.073 | 0.94 | 2.28 |
| | | | 0.75 | 0.001 | 0.101 | 0.096 | 0.94 | 1.46 | 0.005 | 0.079 | 0.075 | 0.94 | 2.17 |
| | | | 0.50 | 0.001 | 0.114 | 0.108 | 0.94 | 1.15 | 0.006 | 0.080 | 0.077 | 0.94 | 2.09 |
| 0.30 | 0.5 | ZZC | | 0.001 | 0.106 | 0.100 | 0.94 | 1.00 | 0.009 | 0.098 | 0.100 | 0.94 | 1.00 |
| | | Proposed | 0.95 | 0.001 | 0.068 | 0.064 | 0.94 | 2.46 | 0.005 | 0.062 | 0.060 | 0.94 | 2.52 |
| | | | 0.75 | 0.002 | 0.087 | 0.081 | 0.94 | 1.49 | 0.005 | 0.064 | 0.062 | 0.94 | 2.38 |
| | | | 0.50 | 0.001 | 0.099 | 0.092 | 0.94 | 1.15 | 0.006 | 0.065 | 0.064 | 0.94 | 2.28 |

Bias, average estimate minus true value; SSD, sample standard deviation; ESE, estimated standard error; CP, coverage proportion with 95% nominal level; RE, relative efficiency of the proposed estimator with respect to the ZZC estimator.

subject, the time to HIV-1 infection was only known to fall between the last negative and first positive test dates, yielding interval-censored failure time data.

An objective of the VAX004 trial is to evaluate the association between antibody responses to rgp120 and the incidence of HIV-1 infection. Gilbert et al.[8] designed a case-cohort study for this purpose, since the infection rate was low and the antibody responses were expensive to measure. The case-cohort sample consisted of a subcohort of 178 vaccine recipients randomly selected at enrollment, including 11 infected and 167 uninfected subjects, as well as 230 infected vaccine recipients outside the subcohort. For subjects in the case-cohort sample, up to eight antibody variables were potentially measured from the available pre-infection blood samples collected at the month 0, 0.5, 1, 1.5, 6, 6.5, 12, 12.5, 18, 18.5, 24, 24.5, 30, and 30.5 study visits. Lacking of appropriate

**Table 4.** Analysis results for the VAX004 trial

| | ZZC method | | | Proposed method | | |
|---|---|---|---|---|---|---|
| Variable | $\hat{\beta}$ | SE | P-value | $\hat{\beta}$ | SE | P-value |
| MN Neutralization Titer | −0.1455 | 0.0247 | 0.0000 | −0.1431 | 0.0246 | 0.0000 |
| Behavorial Risk Score | 0.3119 | 0.0764 | 0.0000 | 0.3564 | 0.0479 | 0.0000 |
| Age | −0.0101 | 0.0156 | 0.5188 | −0.0105 | 0.0154 | 0.4979 |
| Sex | −0.9351 | 1.3994 | 0.5040 | −1.2079 | 1.3546 | 0.3725 |

SE, estimated standard error; P-value, p-value for testing $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$.

methods to deal with case-cohort interval-censored data, Gilbert et al.[8] analyzed the data with the HIV-1 infection date set as the midpoint of the last negative and first positive test dates.

Here we illustrate the proposed method by analyzing the subset of vaccine recipients who were at risk for HIV-1 infection at the month 6 study visit. We focus on the antibody variable, functional assay MN neutralization titer, and consider its measurements up to the month 6 study visit as the main covariate of interest. Specifically, we evaluate the area under the curve of MN neutralization titer values over time from month 0 to month 6 and fit the Cox model for time to HIV-1 infection against this covariate, along with three baseline covariates: age, sex and behavioral risk score. The cohort in our analysis consists of 3381 vaccine recipients at risk for HIV-1 infection at the month 6 study visit, including 208 infected and 3173 uninfected subjects. The MN neutralization titer values are available in 279 subjects, including 67 infected and 212 uninfected subjects. We apply the proposed update estimation method and also consider the original IPW estimation from Zhou et al.,[18] denoted by ZZC, for comparison. For the proposed method, the working model is taken as the Cox model with the three baseline covariates. The degree of Bernstein polynomials is chosen to be $m = 3$ based on the AIC criterion. The analysis results are presented in Table 4. Both methods suggest that higher MN neutralization titer value and lower behavorial risk score are significantly associated with lower risk of HIV-1 infection. This is consistent with the findings of Gilbert et al.[8] Also, the proposed estimator yields smaller standard errors than the ZZC estimator.

## 7 Discussion

We conclude with some extensions or directions for future research. First, the proposed method can be applied to case-cohort designs where the subcohort is selected by sampling without replacement or where the sampling probability depends on covariates. Also, if the cumulative baseline hazard function $\Lambda$ is of interest, we could update $\hat{\Lambda}_n$ similarly as updating $\hat{\vartheta}_n$ to obtain a more efficient estimator, though it would entail a different and more challenging theoretical treatment as $\hat{\Lambda}_n$ has a rate of convergence slower than $\sqrt{n}$ and is not regular. Moreover, although we focused on the Cox model, the proposed update procedure can easily be adapted to other regression models, such as the proportional odds model and semiparametric transformation models. It can also be extended to update the IPW estimators in Zhou et al.[19] that concerns multiple interval-censored failure times and Du et al.[20] that addresses informative interval censoring. In

addition, the proposed update procedure can be extended to the case of time-dependent covariates. However, in this case, the likelihood function involves integration over the covariate process and thus the estimation procedure would be more computationally intensive. In fact, as long as the original estimation procedure can handle the case of time-dependent covariates, the proposed update approach can be applied to improve estimation efficiency. Lastly, an alternative approach to improve upon the IPW method is to employ the augmented inverse probability weighting (AIPW) method given by Robins et al.[33] Although the AIPW method has been applied in various settings, it is difficult to implement in our case, as there is not a simple estimating function that can readily be used for this approach. Another alternative method is to employ maximum likelihood estimation. However, for this method, we need to model the conditional distribution of $X$ given $X^*$ and $Z$, for which imposing a parametric assumption could be too restrictive and using nonparametric methods such as kernel or sieve estimation usually suffer from the curse of dimensionality. Both directions warrant further investigation.

## Acknowledgements

The authors thank the Editors and reviewers for their helpful comments that improve the paper. The authors also thank the Global Solutions for Infectious Diseases (GSID) and Dr. Peter Gilbert for providing data from the Phase 3 HIV vaccine trial VAX004.

## Appendix — Technical Proofs

In this Appendix, we sketch the proof of Theorem 1. We first describe the regularity conditions needed for the proof. Let $\theta_0 = (\vartheta_0, \Lambda_0)$ denote the true value of $\theta = (\vartheta, \Lambda)$ in model (1). Also, let $\theta_0^* = (\vartheta_0^*, \Lambda_0^*)$ be the value of $\theta^* = (\vartheta^*, \Lambda^*)$ that minimizes the Kullback–Leibler divergence given by

$$KL(\theta^*) = E\left\{\log\left(\frac{L(O^*)}{L(\theta^* \mid O^*)}\right)\right\},$$

where $L(\theta^* \mid O^*)$ denotes the likelihood function at $\theta^*$ under the working model (3) based on the data $O^* = \{K, U_1, \ldots, U_K, \Delta_1, \ldots, \Delta_K, Z, X^*\}$, and $L(O^*)$ denotes the true likelihood of $O^*$. The regularity conditions needed for the proof of Theorem 1 are given below:

(C1) The true value $\vartheta_0$ is an interior point of a compact set $\mathcal{B}$ in $R^d$. Also, $\Lambda_0(\cdot)$ is continuously differentiable up to order $r \geq 1$ with strictly positive derivative $\lambda_0(\cdot)$ on $[\sigma, \tau]$ and $0 < \Lambda_0(\sigma) < \Lambda_0(\tau) < \infty$, where $[\sigma, \tau]$ is the union of the supports of $\{U_k : k = 1, \ldots, K\}$ with $0 < \sigma < \tau < \infty$.

(C2) Let $W = (X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$. The distribution of $W$ has a bounded support in $R^d$. If $a^{\mathrm{T}}W + b = 0$, then $a = 0$ and $b = 0$. For some $\kappa > 0$, $a^{\mathrm{T}} \operatorname{var}(W \mid K, \mathcal{U})\, a \geq \kappa\, a^{\mathrm{T}} E(WW^{\mathrm{T}} \mid K, \mathcal{U})\, a$ for all $a \in R^d$, where $\mathcal{U} = (U_1, \ldots, U_K)$.

(C3) The parameter value $\theta_0^*$ is the unique value of $\theta^*$ that minimizes $KL(\theta^*)$, and $\vartheta_0^*$ is an interior point of a compact set $\mathcal{B}^*$ in $R^{d^*}$. Also, $\Lambda_0^*(\cdot)$ is continuously differentiable up to order $r \geq 1$ with strictly positive derivative $\lambda_0^*(\cdot)$ and $0 < \Lambda_0^*(\sigma) < \Lambda_0^*(\tau) < \infty$.

(C4) Let $W^* = (X^{*\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$. The distribution of $W^*$ has a bounded support in $R^{d^*}$. If $a^{\mathrm{T}}W^* + b = 0$, then $a = 0$ and $b = 0$. For some $\kappa^* > 0$, $a^{\mathrm{T}}\,\mathrm{var}(W^* \mid K, \mathcal{U})\,a \geq \kappa^*\, a^{\mathrm{T}}\, E(W^*W^{*\mathrm{T}} \mid K, \mathcal{U})\,a$ for all $a \in R^{d^*}$, where $\mathcal{U} = (U_1, \ldots, U_K)$.

(C5) The number of examination times $K$ is positive with $E(K) < \infty$. There exists some constant $c > 0$ such that $Pr(\min_{0 \leq k \leq K}(U_{k+1} - U_k) \geq c \mid K, Z) = 1$ with probability one. For $k = 0, \ldots, K$, the conditional densities of $(U_k, U_{k+1})$ given $K$ and $(X, X^*, Z)$, denoted by $g_k(u, v)$, have continuous second-order partial derivatives with respect to $u$ and $v$ on $[\sigma, \tau]$ when $v - u \geq c$, and are continuously differentiable with respect to $(X, X^*, Z)$.

(C6) The degree of Bernstein polynomials satisfies $m = o(n^\nu)$ with $1/(2r) < \nu < 1/2$, and $M_n = O(n^a)$ with $a > 0$ controlling the size of the sieve space $\Theta_n$.

**Remark 3.** Condition (C1) is typical for semiparametric regression models for survival data. Condition (C2) is for the identifiability of model (1). Conditions (C3) and (C4) are similar to the previous two conditions but are for the working model (3). Condition (C5) pertains to the joint distribution of the examination times. It requires that the examinations occur over $[\sigma, \tau]$, and that two adjacent examination times are separated by at least a positive constant $c$, so that the likelihood is bounded away from zero. Condition (C6) controls the rates at which the degree of Bernstein polynomials and the size of the sieve space diverge to infinity.

Next, we present four lemmas needed for the proof. For $\theta_1 = (\vartheta_1, \Lambda_1)$ and $\theta_2 = (\vartheta_2, \Lambda_2)$, define the distance

$$d(\theta_1, \theta_2) = \left(\|\vartheta_1 - \vartheta_2\|^2 + \|\Lambda_1 - \Lambda_2\|^2_{L_2[\sigma,\tau]}\right)^{1/2},$$

where $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_{L_2[\sigma,\tau]}$ denotes the $L_2$-norm on $[\sigma, \tau]$.

**Lemma 1.** *Under Conditions (C1), (C2), (C5) and (C6), we have $d(\hat{\theta}_n, \theta_0) \to 0$ almost surely,*

$$d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{(1-\nu)/2,\, \nu r/2\}}),$$

*and*

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = I(\vartheta_0)^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i l(\vartheta_0, \Lambda_0; O_i)\right\} + o_p(1) \to N(0, \Sigma_{11})$$

*in distribution, with*

$$\Sigma_{11} = I(\vartheta_0)^{-1} E\left\{\frac{1}{\pi_q(\Delta)}\left[l(\vartheta_0, \Lambda_0; O)\right]^{\otimes 2}\right\} I(\vartheta_0)^{-1},$$

*where $l(\vartheta_0, \Lambda_0; O)$ and $I(\vartheta_0)$ are the efficient score and information for $\vartheta$, respectively, based on the complete data $O = \{K, U_1, \ldots, U_K, \Delta_1, \ldots, \Delta_K, Z, X\}$, $v^{\otimes 2} = vv^{\mathrm{T}}$ for a vector $v$, and $\pi_q(\Delta) = \left(1 - \sum_{k=1}^{K_i} \Delta_{ik}\right) q_s + \left(\sum_{k=1}^{K_i} \Delta_{ik}\right) \{q_s + (1 - q_s)q_c\}$.*

Lemma 1 concerns the consistency, rate of convergence, and asymptotic normality of the original estimator under the case-cohort sample. The lemma establishes that the efficient score for the original model, $l$, exists. In fact, $l$ is the solution to some integral equation given in Huang and Wellner,[34] but it does not have an explicit form. Specifically, let

$$\ell(\vartheta, \Lambda) = \sum_{k=1}^{K+1} \Delta_k \log \Big[ \exp\big\{ -\Lambda(U_{k-1}) \exp(\beta^{\mathrm{T}} X + \gamma^{\mathrm{T}} Z) \big\} - \exp\big\{ -\Lambda(U_k) \exp(\beta^{\mathrm{T}} X + \gamma^{\mathrm{T}} Z) \big\} \Big]$$

be the log-likelihood function based on a complete observation, where $\vartheta = (\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$. Let $l_\vartheta$ denote the score function for $\vartheta$, i.e., the partial derivatives of $\ell$ with respect to $\vartheta$. Let $l_\Lambda[h]$ denote the score operator for $\Lambda$ along the direction $h$, i.e., $l_\Lambda[h] = \frac{\partial}{\partial s} \ell(\vartheta, \Lambda_s)|_{s=0}$ with $\Lambda_s(\cdot) = (1 + sh(\cdot))\Lambda(\cdot)$. Then the efficient score is defined as $l = l_\vartheta - l_\Lambda[h_0]$, where $h_0$ is the least favorable direction that minimizes $E\|l_\vartheta - l_\Lambda[h]\|^2$. It has been shown in Huang and Wellner[34] that $h_0$ is the solution to a Fredholm integral equation of the second kind:

$$h_0(t) - \int K(t, x) h_0(x) dx = d(t),$$

where $K(t, x)$ and $d(t)$ are two complicated functions formulated in the proof of Theorem 4.1 in Huang and Wellner.[34]

**Lemma 2.** *Under Conditions (C3)–(C6), we have $d(\bar{\theta}_n^*, \theta_0^*) \to 0$ almost surely,*

$$d(\bar{\theta}_n^*, \theta_0^*) = O_p(n^{-\min\{(1-\nu)/2,\, \nu r/2\}}),$$

*and*

$$\sqrt{n}(\bar{\vartheta}_n^* - \vartheta_0^*) = I^*(\vartheta_0^*)^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l^*(\vartheta_0^*, \Lambda_0^*; O_i^*) \right\} + o_p(1) \to N(0, \bar{\Sigma}^*)$$

*in distribution, with*

$$\bar{\Sigma}^* = I^*(\vartheta_0^*)^{-1} E\left\{ l^*(\vartheta_0^*, \Lambda_0^*; O^*)^{\otimes 2} \right\} I^*(\vartheta_0^*)^{-1},$$

*where $l^*(\vartheta_0^*, \Lambda_0^*; O^*)$ is the efficient score and $I^*(\vartheta_0^*)$ is the expectation of the negative derivative of $l^*(\vartheta^*, \Lambda^*; O^*)$ with respect to $\vartheta^*$ at $(\vartheta_0^*, \Lambda_0^*)$.*

Lemma 2 concerns the asymptotic properties of the estimator based on the (generally misspecified) working model (3) under the full data.

**Lemma 3.** *Under Conditions (C3)–(C6), we have* $d(\hat{\theta}_n^*, \theta_0^*) \to 0$ *almost surely,*

$$d(\hat{\theta}_n^*, \theta_0^*) = O_p(n^{-\min\{(1-\nu)/2,\, \nu r/2\}}),$$

*and*

$$\sqrt{n}(\hat{\vartheta}_n^* - \vartheta_0^*) = I^*(\vartheta_0^*)^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i l^*(\vartheta_0^*, \Lambda_0^*; O_i^*) \right\} + o_p(1) \to N(0, \Sigma^*)$$

*in distribution, with*

$$\Sigma^* = I^*(\vartheta_0^*)^{-1} E \left\{ \frac{1}{\pi_q(\Delta)} \left[ l^*(\vartheta_0^*, \Lambda_0^*; O^*) \right]^{\otimes 2} \right\} I^*(\vartheta_0^*)^{-1},$$

*where* $l^*(\vartheta_0^*, \Lambda_0^*; O^*)$ *and* $I^*(\vartheta_0^*)$ *are defined in Lemma 2, and* $\pi_q(\Delta)$ *is defined as in Lemma 1.*

Similar to Lemma 2, Lemma 3 pertains to the working model, but it concerns the asymptotic properties of the estimator based on the case-cohort sample.

**Lemma 4.** *Under Conditions (C3)–(C6), we have*

$$\sqrt{n}(\hat{\vartheta}_n^* - \bar{\vartheta}_n^*) = I^*(\vartheta_0^*)^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (w_i - 1) l^*(\vartheta_0^*, \Lambda_0^*; O_i^*) \right\} + o_p(1) \to N(0, \Sigma_{22})$$

*in distribution, with*

$$\Sigma_{22} = I^*(\vartheta_0^*)^{-1} E \left\{ \frac{1 - \pi_q(\Delta)}{\pi_q(\Delta)} \left[ l^*(\vartheta_0^*, \Lambda_0^*; O^*) \right]^{\otimes 2} \right\} I^*(\vartheta_0^*)^{-1},$$

*where* $l^*(\vartheta_0^*, \Lambda_0^*; O^*)$ *and* $I^*(\vartheta_0^*)$ *are defined in Lemma 2, and* $\pi_q(\Delta)$ *is defined as in Lemma 1.*

Lemma 4 concerns the difference of the two estimators under the working model.

Lemma 1 follows from Theorems 1 and 2 of Zhou et al.;[18] we can also prove Lemma 3 using arguments similar to the proofs of these two theorems. Lemma 4 is a direct consequence of Lemmas 2 and 3. Theorem 1 follows from Lemmas 1 and 4 and Slutsky's theorem. In the following, we sketch the proof of Lemma 2.

Let $\mathbb{P}$ and $\mathbb{P}_n$ denote the true and empirical measures, respectively. Let $l(\theta^* \mid O^*)$ be the log-likelihood function under the working model (3) based on a single observation $O^* = \{K, U_1, \ldots, U_K, \Delta_1, \ldots, \Delta_K, Z, X^*\}$,

$$\begin{aligned} l(\theta^* \mid O^*) = \sum_{k=1}^{K+1} \Delta_k \Big\{ \log \big[ \exp\left\{ -\Lambda^*(U_{k-1}) \exp(\beta^{*\mathrm{T}} X^* + \gamma^{*\mathrm{T}} Z) \right\} \\ - \exp\left\{ -\Lambda^*(U_k) \exp(\beta^{*\mathrm{T}} X^* + \gamma^{*\mathrm{T}} Z) \right\} \big] \Big\}. \end{aligned} \tag{4}$$

Let $\Theta_n^* = \mathcal{B}^* \otimes \mathcal{M}_n$, where $\mathcal{M}_n$ is defined in (2). Then $\bar{\theta}_n^* = \arg\max_{\theta^* \in \Theta_n^*} \mathbb{P}_n l(\theta^* \mid O^*)$. Recall that $\theta_0^* = (\vartheta_0^*, \Lambda_0^*)$ is the value of $\theta^* = (\vartheta^*, \Lambda^*)$ that minimizes the Kullback-Leibler divergence given by

$$KL(\theta^*) = E\left\{\log\left(\frac{L(O^*)}{L(\theta^* \mid O^*)}\right)\right\},$$

where $L(\theta^* \mid O^*)$ denotes the likelihood of $O^*$ under the working model (3) and $L(O^*)$ is the true likelihood of $O^*$. Then $\theta_0^* = \arg\max_{\theta^*} \mathbb{P}l(\theta^* \mid O^*)$.

To prove the consistency result in Lemma 2, we can first show that $\sup_{\theta^* \in \Theta_n^*} \|\mathbb{P}_n l(\theta^* \mid O^*) - \mathbb{P}l(\theta^* \mid O^*)\| \to 0$ almost surely, following Zhou et al.[30] By the definitions of $\bar{\theta}_n^*$ and $\theta_0^*$ as well as the uniqueness of $\theta_0^*$ under (C3), we can show that $d(\bar{\theta}_n^*, \theta_0^*) \to 0$ almost surely. Also, the rate of convergence $d(\bar{\theta}_n^*, \theta_0^*) = O_p(n^{-\min\{(1-\nu)/2,\, \nu r/2\}})$ can be shown by using Theorem 3.4.1 of van der Vaart and Wellner[35] similarly as in Zhou et al.[30] Furthermore, noting that $\theta_0^* = \arg\max_{\theta^*} \mathbb{P}l(\theta^* \mid O^*)$, we can establish the asymptotic normality in Lemma 2 under Conditions (C3)–(C6) similarly as in Zhou et al.[30] and Wong et al.[36] Similarly as discussed in Remark 2, the efficient score $l^*$ exists and solves some integral equation, but does not have an explicit form.[30;36]

## Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

Qingning Zhou's work was partially supported by the National Science Foundation grant DMS-1916170. Kin Yau Wong's work was partially supported by the Hong Kong Research Grants Council grant PolyU153034/22P.

## References

1. Breslow NE and Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* 2007; 34(1): 86–102.
2. Saegusa T and Wellner JA. Weighted likelihood estimation under two-phase sampling. *Ann Statist* 2013; 41(1): 269–295.
3. Zeng D and Lin DY. Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J Am Statist Assoc* 2014; 109(505): 371–383.
4. The ARIC Investigators. The Atherosclerosis Risk in Community (ARIC) study: design and objectives. *American Journal of Epidemiology* 1989; 129: 687–702.
5. Kim S, Zeng D and Cai J. Analysis of multiple survival events in generalized case-cohort designs. *Biometrics* 2018; 74(4): 1250–1260.
6. Seaton KE, Huang Y, Karuna S et al. Pharmacokinetic serum concentrations of vrc01 correlate with prevention of HIV-1 acquisition. *EBioMedicine* 2023; 93.

7. Corey L, Gilbert PB, Juraska M et al. Two randomized trials of neutralizing antibodies to prevent HIV-1 acquisition. *New England Journal of Medicine* 2021; 384: 1003–1014.
8. Gilbert PB, Peterson ML, Follmann D 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* 2005; 191(5): 666–677.
9. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; 73(1): 1–11.
10. Self SG and Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann Statist* 1988; 16(1): 64–81.
11. Chen K and Lo SH. Case-cohort and case-control analysis with Cox's model. *Biometrika* 1999; 86(4): 755–764.
12. Marti H and Chavance M. Multiple imputation analysis of case–cohort studies. *Statist Med* 2011; 30(13): 1595–1607.
13. Scheike TH and Martinussen T. Maximum likelihood estimation for Cox's regression model under case–cohort sampling. *Scand J Statist* 2004; 31(2): 283–293.
14. Kang S and Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika* 2009; 96(4): 887–901.
15. Kim S, Cai J and Lu W. More efficient estimators for case-cohort studies. *Biometrika* 2013; 100(3): 695–708.
16. Li Z, Gilbert P and Nan B. Weighted likelihood method for grouped survival data in case–cohort studies with application to HIV vaccine trials. *Biometrics* 2008; 64(4): 1247–1255.
17. Li Z and Nan B. Relative risk regression for current status data in case-cohort studies. *Canad J Statist* 2011; 39(4): 557–577.
18. Zhou Q, Zhou H and Cai J. Case-cohort studies with interval-censored failure time data. *Biometrika* 2017; 104(1): 17–29.
19. Zhou Q, Cai J and Zhou H. Semiparametric regression analysis of case-cohort studies with multiple interval-censored disease outcomes. *Statistics in medicine* 2021; 40(13): 3106–3123.
20. Du M, Zhou Q, Zhao S et al. Regression analysis of case-cohort studies in the presence of dependent interval censoring. *Journal of applied statistics* 2021; 48(5): 846–865.
21. Chen YH and Chen H. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2000; 62(3): 449–460.
22. Chen YH. Cox regression in cohort studies with validation sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; 64(1): 51–62.
23. Wang X and Wang Q. Semiparametric linear transformation model with differential measurement error and validation sampling. *Journal of Multivariate Analysis* 2015; 141: 67–80.
24. Yan Y, Zhou H and Cai J. Improving efficiency of parameter estimation in case-cohort studies with multivariate failure time data. *Biometrics* 2017; 73(3): 1042–1052.
25. Yang S and Ding P. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association* 2020; 115(531): 1540–1554.
26. Tong J, Huang J, Chubak J et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association* 2020; 27(2): 244–253.

27. Yin Z, Tong J, Chen Y et al. A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. *Journal of the American Medical Informatics Association* 2022; 29(1): 52–61.
28. Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Ann Statist* 1996; 24(2): 540–568.
29. Zeng D, Mao L and Lin D. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* 2016; 103(2): 253–271.
30. Zhou Q, Hu T and Sun J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association* 2017; 112(518): 664–672.
31. Ma S and Kosorok MR. Robust semiparametric M-estimation and the weighted bootstrap. *J Multivar Anal* 2005; 96(1): 190–217.
32. Harro CD, Judson FN, Gorse GJ et al. Recruitment and baseline epidemiologic profile of participants in the first phase 3 HIV vaccine efficacy trial. *J of Acquir Immune Defic Syndr* 2004; 37(3): 1385–1392.
33. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 1994; 89(427): 846–866.
34. Huang J and Wellner JA. Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*. Springer, pp. 123–169.
35. van der Vaart AW and Wellner JA. *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer, 1996.
36. Wong KY, Zhou Q and Hu T. Semiparametric regression analysis of doubly-censored data with applications to incubation period estimation. *Lifetime Data Analysis* 2023; 29(1): 87–114.