Published in final edited form as:

Lifetime Data Anal. 2020 January ; 26(1): 85–108. doi:10.1007/s10985-019-09461-5.

Semiparametric Inference for a Two-stage Outcome-dependent Sampling Design with Interval-censored Failure Time Data

Qingning Zhou,

Department of Mathematics and Statistics, University of North Carolina at Charlotte

Jianwen Cai,

Department of Biostatistics, University of North Carolina at Chapel Hill

Haibo Zhou

Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

We propose a two-stage outcome-dependent sampling design and inference procedure for studies that concern interval-censored failure time outcomes. This design enhances the study efficiency by allowing the selection probabilities of the second-stage sample, for which the expensive exposure variable is ascertained, to depend on the first-stage observed interval-censored failure time outcomes. In particular, the second-stage sample is enriched by selectively including subjects who are known or observed to experience the failure at an early or late time. We develop a sieve semiparametric maximum pseudo likelihood procedure that makes use of all available data from the proposed two-stage design. The resulting regression parameter estimator is shown to be consistent and asymptotically normal, and a consistent estimator for its asymptotic variance is derived. Simulation results demonstrate that the proposed design and inference procedure performs well in practical situations and is more efficient than the existing designs and methods. An application to a phase 3 HIV vaccine trial is provided.

Keywords

Bernstein polynomial; Biased sampling; Missing data; Proportional hazards model; Sieve estimation

1 Introduction

Interval-censored failure time data arise in many applications where the failure time is known or observed only to fall within some time interval instead of being exactly observed (e.g. Sun, 2006; Chen *et al.*, 2012). While the analysis of interval-censored data has attracted a great deal of attention, the existing work on interval-censored data is mainly with data obtained by simple random sampling, which may not be the optimal sampling strategy especially when the failure rate is low or the exposure measurements are expensive. The objective of this paper is to propose a cost-effective sampling design and inference procedure for interval-censored failure time data. An example that motivated our research arose from a phase 3 HIV vaccine trial where it was of interest to investigate the association between antibody responses to a vaccine and the incidence of HIV infection (e.g. Gilbert *et*

et al., 2005). In this study, 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. Since the antibody levels are usually measured by complex assays that are expensive to perform, cost-effective study designs that enable one to obtain a more informative sample of subjects at a fixed budget are desirable. In this paper, we propose a two-stage outcome-dependent sampling design for studies concerning interval-censored failure time outcomes and also develop a semiparametric likelihood-based inference procedure to analyze the observed data from this design.

Outcome-dependent sampling (ODS) design is a biased sampling scheme that allows the probability of obtaining the exposure measurement to depend on the observed value of the outcome. The fundamental idea of ODS design is to oversample subjects from the segments of the population that are thought to be more informative in terms of the exposure-outcome relationship. Case-control study is a simple and well-known example of ODS design with a binary outcome (e.g. Cornfield, 1951; Whittemore, 1997). References on the more general ODS design with a continuous outcome include Zhou *et al.* (2002), Chatterjee *et al.* (2003), Weaver and Zhou (2005), Song *et al.* (2009) and Zhou *et al.* (2011), among others. Case-cohort design developed by Prentice (1986) is a widely used cost-effective outcome-dependent sampling scheme for failure time data. Under the case-cohort design, the exposure measurements are obtained for a random sample of the study cohort, named subcohort, as well as all subjects who experience the failure, named cases, regardless of whether or not they belong to the subcohort (e.g. Self and Prentice, 1988; Chen and Lo, 1999; Kulich and Lin, 2004; Kang and Cai, 2009; Zeng and Lin, 2014; Sun *et al.*, 2017). Recently, Ding *et al.* (2014) and Yu *et al.* (2015) studied a general failure-time-dependent sampling scheme that incorporates the ideas of ODS design for continuous outcome and case-cohort design. Overview of the existing cost-effective sampling designs for failure time data can be found in Ding *et al.* (2017).

The existing cost-effective sampling designs for failure time data were primarily developed for traditional right-censored data where the failure time is either exactly observed or right-censored. Motivated by practical needs in epidemiological studies and disease prevention trials (e.g., the HIV vaccine trial mentioned above), the research on cost-effective designs for interval-censored failure time data has attracted a great deal of attention recently. For example, Gilbert *et al.* (2005), Li *et al.* (2008), Li and Nan (2011) and Zhou *et al.* (2017a) extended the case-cohort design to the case of interval-censored data. Gilbert *et al.* (2005) treated the midpoint of an interval-censored observation as the true failure time and then employed the case-cohort design for right-censored data. Li *et al.* (2008) and Li and Nan (2011) studied the case-cohort designs for grouped data and current status data which are important special cases of interval-censored data. More recently, Zhou *et al.* (2017a) developed the case-cohort design for failure time data subject to general interval-censoring. The case-cohort design samples all failure subjects and applies mainly to rare events. For studies that evaluate non-rare or not-so-rare events, it may not be affordable or feasible to sample all failure subjects. For this problem, Zhou *et al.* (2018) proposed a cost-effective design that selectively samples more informative failure subjects based on their failure times. They also developed an efficient inference procedure that reaps the gain from their design.

However, their proposed design and method does not utilize the information outside the selected sample which is often available and valuable in real studies.

In this paper, we propose a two-stage outcome-dependent sampling design for interval-censored failure time data along with an inference procedure that makes use of all available information. We refer to the proposed design as the two-stage interval-censoring ODS design. Under this design, the interval-censored failure time outcome is assumed to be observed for every member in the study cohort at the first stage, and at the second stage, the exposure variable is measured for a subset of the study cohort that is selected with a probability depending on the first-stage observed interval-censored failure time outcome. More specifically, the second-stage sample consists of a simple random sample (SRS) from the study cohort and two supplemental samples (S_1 and S_2) taken among the cohort members who are known or observed to experience the failure at an early or late time. An illustration of the proposed design and data structure is highlighted in Figures 1 and 2. The idea of oversampling from early and late failure subjects stems from the linear regression, where the least squares estimate of the regression parameter will have a smaller variance if more samples are taken from the tails of the covariate distribution. Intuitively, if the failure time T is positively associated with the covariate X , then high (low) T would be associated with high (low) X ; thus enriching the observed sample with subjects who have high or low T could potentially enhance the efficiency in evaluating the association between T and X .

We develop a semiparametric likelihood-based inference procedure that makes use of all available data from the proposed two-stage interval-censoring ODS design. The likelihood function in our case involves not only the regression parameter of interest but also two infinite-dimensional nuisance parameters, i.e., the exposure's distribution function and the cumulative baseline hazard function under the proportional hazards model. Estimating the regression parameter without specifying the nuisance functions is a challenging task. For this, we propose a two-step estimation procedure that leaves the nuisance functions completely unspecified. First, we address the exposure's distribution function using a pseudo likelihood method that is nonparametric and easy to implement. Second, we handle the cumulative baseline hazard function with the Bernstein-polynomials-based sieve method. Sieve methods enjoy both theoretical and computational advantages and have turned out to be a useful tool in solving various problems of interval-censored data (e.g. Huang and Rossini, 1997; Xue *et al.*, 2004; Zhang *et al.*, 2010; Zhou *et al.*, 2017b). We show that the proposed regression parameter estimator is consistent and asymptotically normal. We also provide a consistent variance estimator that is easy to compute.

The rest of the paper is organized as follows. Section 2 introduces the proposed two-stage interval-censoring ODS design and the data structure. Section 3 develops the sieve semiparametric maximum pseudo likelihood method that fits the proportional hazards model to the observed data. Section 3 also gives the asymptotic properties of the resulting estimator. Section 4 includes the results from simulation studies that investigate the performance of the proposed design and method in comparison to the existing designs and competing approaches. Section 5 illustrates the proposed design and method using data from a phase 3 HIV vaccine trial. Some concluding remarks are given in Section 6.

2 Two-stage Interval-censoring ODS Design

Consider a cohort study that concerns a failure time outcome and consists of N independent subjects. Suppose that only interval-censored failure time data are observed in the study and can be represented by

$$Y_i = \{U_i, V_i, \Delta_{1i} = I(T_i \leq U_i), \Delta_{2i} = I(U_i < T_i \leq V_i)\}, i = 1, \dots, N,$$

where T_i denotes the failure time, $\{U_i, V_i\}$ denote two random examination times, and $\{\Delta_{1i}, \Delta_{2i}\}$ indicate left-, interval- and right-censored observations, respectively. Let Z_i denote the p -dimensional covariate vector for subject i . Under the two-stage ODS design, the interval-censored failure time outcomes $\{Y_i, i = 1, \dots, N\}$ are observed for the full cohort at the first stage, and at the second stage, the covariate measurements are obtained for a subset of the study cohort selected with a probability depending on the first-stage observed outcomes. In particular, following the spirits of case-cohort and ODS designs (Prentice, 1986; Zhou *et al.*, 2002), at the second stage, we first take a simple random sample (SRS) of the study cohort and then supplement the SRS with subjects who are known or observed to experience the failure and have the observed time intervals falling within the two tails of the failure time's distribution. More specifically, let τ denote the length of study duration and $\{a_1, a_2\}$ two known constants satisfying $0 < a_1 < a_2 < \tau$. We partition the population into four mutually exclusive and exhaustive strata based on the interval-censored failure time outcome Y :

$$\begin{aligned} A_1 &= \{Y: \Delta_1 = 1, U < a_1\} \cup \{Y: \Delta_2 = 1, V < a_1\}, \\ A_2 &= \{Y: \Delta_2 = 1, U > a_2\}, \\ A_3 &= \{Y: \Delta_1 + \Delta_2 = 1\} \cap (A_1 \cup A_2)^c, \\ A_4 &= \{Y: \Delta_1 + \Delta_2 = 0\}, \end{aligned} \quad (1)$$

where $\cap$ and $\cup$ denote the intersection and union of sets, respectively, and B^c the complement of a set B . It is easy to see that $A_1 \cup A_2 \cup A_3$ includes all failures (i.e. subjects with $\Delta_1 + \Delta_2 = 1$) while A_4 consists of all non-failures (i.e. subjects with $\Delta_1 + \Delta_2 = 0$). Subjects with $Y \in A_1$ have the failure time falling in the lower tail $(0, a_1)$ while subjects with $Y \in A_2$ have the failure time belong to the upper tail (a_2, τ) . An illustration of the partitions $\{A_i, i = 1, \dots, 4\}$ is given in Figure 1. At the second stage, we obtain the covariate measurements Z for a simple random sample (SRS), indexed by I_0 , of size n_0 from the first-stage study cohort and a supplemental random sample (S_k), indexed by I_k , of size n_k from the part of the study cohort that belong to $A_k, k = 1, 2$ (note that $n_3 = n_4 = 0$). We refer to the data with Z observed as the validation sample and the data without Z observed as the nonvalidation sample. Let $\{n_{\bar{v}}, n_v = n_0 + n_1 + n_2\}$ and $\{I_{\bar{v}}, I_v = I_0 \cup I_1 \cup I_2\}$ denote the sizes and index sets of the nonvalidation and validation samples, respectively. Then the data structure under our two-stage interval-censoring ODS design has the form:

Nonvalidation sample: $Y_i, i \in I_{\bar{v}};$

Validation sample:

$$\begin{cases} (Y_i, Z_i), i \in I_0, \text{ simple random sample (SRS);} \\ (Y_i, Z_i | Y_i \in A_1), i \in I_1, \text{ supplemental sample } (S_1) \text{ from lower tail;} \\ (Y_i, Z_i | Y_i \in A_2), i \in I_2, \text{ supplemental sample } (S_2) \text{ from lower tail.} \end{cases} \quad (2)$$

Figure 2 presents an illustration of the data structure. In the next section, we develop a semiparametric likelihood-based method to make inference on the association between the failure time T and the covariate Z based on the observed data (2).

3 Sieve Semiparametric Maximum Pseudo Likelihood Method

Suppose that the failure time T follows the proportional hazards model:

$$\Lambda(t|Z) = \Lambda(t) \exp(\beta'Z), \quad (3)$$

where $\Lambda(t)$ is the unspecified cumulative baseline hazard function and β is the regression parameter of interest. Assume that the examination times (U, V) are independent of the covariate Z and also independent of T conditional on Z , and the joint distribution of (U, V, Z) does not depend on (β, Λ) . Then the likelihood function based on the observed data (2) can be written as:

$$\begin{aligned} L(\beta, \Lambda, G_Z) &= \prod_{i \in I_v} f(Y_i | Z_i; \beta, \Lambda) dG_Z(Z_i) \cdot \prod_{j \in I_{\bar{v}}} f_Y(Y_j) \\ &= \prod_{i \in I_v} f(Y_i | Z_i; \beta, \Lambda) dG_Z(Z_i) \cdot \prod_{j \in I_{\bar{v}}} \int f(Y_j | z; \beta, \Lambda) dG_Z(z) \end{aligned} \quad (4)$$

with

$$f(Y_i | Z_i; \beta, \Lambda) = (1 - S(U_i | Z_i))^{\Delta_{1i}} (S(U_i | Z_i) - S(V_i | Z_i))^{\Delta_{2i}} S(V_i | Z_i)^{1 - \Delta_{1i} - \Delta_{2i}}$$

where $S(t|Z) = \exp(-\Lambda(t) \exp(\beta'Z))$ is the survival function of T given $Z = z$, $f_Y(\cdot)$ is the marginal density of Y , and $G_Z(\cdot)$ is the marginal distribution of Z that does not depend on (β, Λ) . The likelihood function cannot be separated into two parts, where one part involves only the regression parameter β and the other involves only the infinite-dimensional parameters (Λ, G_Z) . Thus, to estimate the regression parameter β of primary interest, one has to deal with the nuisance functions (Λ, G_Z) . In the following, we develop a two-step estimation procedure for β that leaves the nuisance functions (Λ, G_Z) completely unspecified.

We first deal with the covariate distribution G_Z using a pseudo likelihood method. Specifically, we derive a consistent estimator of G_Z and then plug this estimator into the likelihood (4) to obtain a pseudo likelihood function that involves only (β, Λ) . Let N_k denote

the total number of subjects in the study cohort that belong to the stratum A_k , and $\{n_{\bar{v}_k}, n_{v_k}\}$ and $\{I_{\bar{v}_k}, I_{v_k}\}$ denote the sizes and index sets of the nonvalidation and validation samples that belong to A_k , respectively, where $k = 1, \dots, 4$. Then $N_k = n_{\bar{v}_k} + n_{v_k}$, $n_{\bar{v}} = \sum_{k=1}^4 n_{\bar{v}_k}$, $n_v = \sum_{k=1}^4 n_{v_k}$, $I_{\bar{v}} = \cup_{k=1}^4 I_{\bar{v}_k}$ and $I_v = \cup_{k=1}^4 I_{v_k}$. Furthermore, let n_{0k} denote the size of the SRS component of the validation sample that belongs to A_k , then we have $n_0 = \sum_{k=1}^4 n_{0k}$ and $n_{v_k} = n_k + n_{0k}$, where n_0 and n_k are the sizes of SRS and supplemental components of the validation sample, respectively. Using the law of total probability, the distribution G_Z can be written as

$$G_Z(z) = Pr(Z \leq z) = \sum_{k=1}^4 Pr(Y \in A_k) Pr(Z \leq z | Y \in A_k).$$

A consistent estimator of the distribution function G_Z can thus be obtained as:

$$\hat{G}_Z(z) = \sum_{k=1}^4 \frac{N_k}{N} \sum_{r \in I_{v_k}} \frac{I(Z_r \leq z)}{n_k + n_{0k}}. \quad (5)$$

Plugging $\hat{G}_Z$ into (4) and taking the logarithm, we have the pseudo log-likelihood function:

$$\begin{aligned} \hat{l}(\beta, \Lambda) &= \sum_{i \in I_v} \log f(Y_i | Z_i; \beta, \Lambda) + \sum_{j \in I_{\bar{v}}} \log \hat{f}_Y(Y_j; \beta, \Lambda) \\ &= \sum_{i \in I_v} \log f(Y_i | Z_i; \beta, \Lambda) + \sum_{j \in I_{\bar{v}}} \log \left\{ \int f(Y_j | z; \beta, \Lambda) d\hat{G}_Z(z) \right\} \\ &= \sum_{i \in I_v} \log f(Y_i | Z_i; \beta, \Lambda) + \sum_{j \in I_{\bar{v}}} \log \left\{ \sum_{k=1}^4 \frac{N_k}{N(n_k + n_{0k})} \sum_{r \in I_{v_k}} f(Y_j | Z_r; \beta, \Lambda) \right\} \end{aligned} \quad (6)$$

We now consider the estimation of β based on the pseudo log-likelihood (6), which involves only one nuisance function Λ . One common way of dealing with a nuisance function is the semiparametric maximum likelihood estimation (SPMLE) approach that assumes the function estimate only having masses at the observed data points. The SPMLE method, however, involves a large and growing number of parameters ($2N$ if there are no ties among the observed intervals) and often poses computation issues. As an alternative to SPMLE, sieve methods have become more and more popular in estimating the infinite-dimensional

nuisance parameter (e.g. Huang and Rossini, 1997; Xue *et al.*, 2004; Zhang *et al.*, 2010; Zhou *et al.*, 2017b). The basic idea of sieve methods is to approximate a large parameter space with a sequence of increasing subspaces (called sieves) such that the closure of the limiting subspace contains the original parameter space. Compared to SPMLE, sieve methods are computationally easier to work with as the sieve likelihood usually involves much fewer parameters than the semiparametric likelihood.

We propose to estimate $\theta = (\beta, \Lambda)$ using the Bernstein-polynomial-based sieve method. The reason that we use Bernstein polynomial is two-fold: first, compared with other bases like spline and piecewise linear functions, Bernstein polynomial is easier to work with and does not require the specification of interior knots; second, Bernstein polynomial can model the monotonicity and nonnegativity of the cumulative baseline hazard function with simple restrictions that can easily be removed through reparameterization. Specifically, we first define the parameter space as

$$\Theta = \{\theta = (\beta, \Lambda) \in \mathcal{B} \otimes \mathcal{M}\},$$

where $\mathcal{B} = \{\beta \in R^p: \|\beta\| \leq M\}$ with M being a positive constant, and $\mathcal{M}$ is the collection of all continuous nondecreasing and nonnegative functions over the interval $[\sigma, \tau]$, where σ and τ are known constants usually taken to be the lower and upper bounds of all observed times in practice. We define the Bernstein-polynomial-based sieve space as

$$\Theta_N = \{\theta_N = (\beta, \Lambda_N) \in \mathcal{B} \otimes \mathcal{M}_N\},$$

where $\mathcal{B}$ is defined above and

$$\mathcal{M}_N = \left\{ \Lambda_N(t) = \sum_{k=0}^m \phi_k B_k(t, m, \sigma, \tau): 0 \leq \phi_0 \leq \phi_1 \leq \dots \leq \phi_m \text{ and } \sum_{k=0}^m |\phi_k| \leq M_N \right\}$$

with $B_k(t, m, \sigma, \tau)$ being Bernstein basis polynomials of degree $m = \alpha(N^\nu)$ for some $\nu \in (0, 1)$,

$$B_k(t, m, \sigma, \tau) = \binom{m}{k} \left(\frac{t - \sigma}{\tau - \sigma} \right)^k \left(1 - \frac{t - \sigma}{\tau - \sigma} \right)^{m-k}, \quad k = 0, \dots, m,$$

and $M_N = O(N^a)$ for some $a > 0$ controlling the size of the sieve space. According to Lorentz (1986), any continuous $\Lambda(t)$ can be approximated by a Bernstein polynomial $\Lambda_N(t)$ with the coefficients $\phi_k = \Lambda(\sigma + (k/m)(\tau - \sigma))$ arbitrarily well as $N \rightarrow \infty$, and Θ_N can thus be used as a sieve space of Θ . We define the sieve semiparametric maximum pseudo likelihood estimator $\hat{\theta}_N = (\hat{\beta}_N, \hat{\Lambda}_N)$ of θ to be the value of θ that maximizes the sieve pseudo log-likelihood (6) over Θ_N . The boundedness and monotonicity constraints on the Bernstein coefficients ϕ_k 's in $\mathcal{M}_N$ ensure that $\hat{\Lambda}_N$ is a nonnegative and nondecreasing function, i.e. a valid cumulative baseline hazard function.

The asymptotic properties of the proposed estimator $\hat{\theta}_N$ are established in Theorems 1 and 2 below. Denote $F(u, v)$ the joint distribution function of the two random examination times (U, V) and define a distance on the parameter space $\Theta = \mathcal{B} \otimes \mathcal{M}$ as

$$d(\theta^1, \theta^2) = \left\{ \|\beta^1 - \beta^2\|^2 + \|\Lambda^1 - \Lambda^2\|_2^2 \right\}^{1/2},$$

for any $\theta^1 = (\beta^1, \Lambda^1) \in \Theta$ and $\theta^2 = (\beta^2, \Lambda^2) \in \Theta$, where $\|v\|$ denote the Euclidean norm for a vector v and $\|\Lambda^1 - \Lambda^2\|_2^2 = \int \left[(\Lambda^1(u) - \Lambda^2(u))^2 + (\Lambda^1(v) - \Lambda^2(v))^2 \right] dF(u, v)$. Let $\theta_0 = (\beta_0, \Lambda_0)$ denote the true value of θ and assume that $N_k/N \rightarrow \pi_k = Pr(Y \in A_k)$, $n_v/N \rightarrow \rho_v > 0$, $n_0/n_v \rightarrow \rho_0 > 0$ and $n_k/n_v \rightarrow \rho_k = 0$, $k = 1, \dots, 4$, as $N \rightarrow \infty$. The following theorems give the consistency and asymptotic normality of the proposed estimator $\hat{\theta}_N$ when $N \rightarrow \infty$. The regularity conditions needed for these theorems are given in the Appendix.

Theorem 1. Assume that Conditions (C1) - (C5) given in the Appendix hold, we have $d(\hat{\theta}_N, \theta_0) \rightarrow 0$ almost surely and $d(\hat{\theta}_N, \theta_0) = O_p(N^{-\min\{(1-\nu)/2, r\nu/2\}})$, where $\nu \in (0, 1)$ such that $m = \alpha(N^\nu)$ and r is defined in Condition (C3).

Theorem 2. Assume that Conditions (C1) - (C5) given in the Appendix hold. If $\nu > 1/2r$, we have

$$N^{1/2}(\hat{\beta}_N - \beta_0) \rightarrow N(0, \Sigma)$$

in distribution with the asymptotic covariance matrix given by

$$\Sigma = I^{-1}(\beta_0) + \sum_{l=1}^4 \frac{\pi_l^2}{\rho_l \rho_v + \pi_l \rho_0 \rho_v} I^{-1}(\beta_0) \int_{A_l} (\theta_0) I^{-1}(\beta_0),$$

where $I(\beta)$, defined in the Appendix, is the information for β if $G_Z(z)$ were known, and

$$\sum_l (\theta) = \text{var} \left\{ Z | Y \in A_l \right\} \left\{ \sum_{k=1}^4 [\pi_k (1 - \rho_0 \rho_v) - \rho_k \rho_v] E \left\{ Y | Y \in A_k \right\} \left[M_Z(Y; \theta) \right] \right\}$$

$$M_Z(Y; \theta) = \frac{\partial f(Y|Z; \theta) / \partial \theta}{f_Y(Y; \theta)} - \frac{\partial f_Y(Y; \theta) / \partial \theta}{[f_Y(Y; \theta)]^2} f(Y|Z; \theta)$$

with $f(Y|Z; \theta) / \partial \theta$ and $f_Y(Y; \theta) / \partial \theta$ defined in the Appendix.

In Theorem 2, the second term of the covariance matrix Σ is due to the variability of the estimated covariate distribution in the pseudo likelihood function. The proofs of Theorems 1 and 2 are outlined in the Appendix.

A consistent estimator of Σ can be obtained by treating the pseudo log-likelihood function $\hat{l}(\beta, \Lambda)$ as a function of the $(p+m+1)$ -dimensional parameter $\theta = (\beta, \phi)$ (Huang *et al.*, 2012).

In particular, the covariance matrix estimator $\hat{\Sigma}$ can be obtained from

$$\hat{l}^{-1}(\hat{\theta}_N) + \sum_{l=1}^4 \frac{(N_l/N)^2}{(n_l + n_{0l})/N} \hat{l}^{-1}(\hat{\theta}_N) \hat{\Sigma}_l(\hat{\theta}_N) \hat{l}^{-1}(\hat{\theta}_N),$$

where $\hat{l}(\theta)$ is the observed information matrix calculated from the pseudo likelihood and

$$\hat{\Sigma}_l(\theta) = \widehat{\text{var}}\{Z_i; i \in I_{v_l}\} \left\{ \sum_{k=1}^4 \frac{n_{\bar{v}_k}}{N} \widehat{M}_{Z_i, k}(\theta) \right\}$$

with

$$\widehat{M}_{Z_i, k}(\theta) = \sum_{j \in I_{\bar{v}_k}} \left\{ \frac{\partial f(Y_j | Z_i; \theta) / \partial \theta}{\hat{f}_Y(Y_j; \theta)} - \frac{\partial \hat{f}_Y(Y_j; \theta) / \partial \theta}{[\hat{f}_Y(Y_j; \theta)]^2} f(Y_j | Z_i; \theta) \right\} \Bigg| n_{\bar{v}_k}$$

and $\hat{f}_Y(Y_j; \theta)$ given in (6).

Several remarks on the implementation of the proposed inference procedure are given below. First, although there are some monotonicity and boundedness constraints on the Bernstein coefficients in the sieve space, these constraints can easily be removed via reparametrization. For example, regarding the constraint $0 \leq \phi_0 \leq \phi_1 \leq \dots \leq \phi_m$, one could reparametrize $\{\phi_0, \dots, \phi_m\}$ as the cumulative sums of $\{\exp(\phi_0^*), \dots, \exp(\phi_m^*)\}$. With respect to the constraint $\sum_{k=0}^m |\phi_k| \leq M_N$, since $M_N = O(N^a)$ can be taken reasonably large for fixed N , it is not necessary to consider this constraint in computation. Hence our estimation problem is actually an unconstrained finite-dimensional optimization problem. The Quasi-Newton algorithm built in *fminunc* in Matlab will be used for the numerical studies below. Another aspect to note about the implementation of the proposed method is how to determine the degree of Bernstein basis polynomial m . With respect to this, we suggest to choose m based on the AIC criterion, i.e. choose the m value that minimizes

$$AIC = -2\hat{l}(\hat{\theta}_N) + 2(p + m + 1). \quad (7)$$

More discussion on the determination of m will be given in the concluding remarks.

4 Simulation Studies

We conduct three simulation studies to evaluate the finite-sample performance of the proposed design and inference procedure. In the first study, we compare the ODS design with the commonly used SRS designs, and also compare the proposed estimator with the

inverse probability weighted estimator and Zhou *et al.* (2018)'s estimator. In the second study, we assess how the proposed method works under various sample sizes, particularly various size allocations of SRS and supplemental components in the validation sample. In the third study, we investigate how our method performs and compares with the competing ones if additional covariate information is available for the cohort.

In the first two studies, we simulate two covariates Z and W independently from $N(0, 1)$ and given (Z, W) , the failure time T is generated from the proportional hazards model $\Lambda(t|Z, W) = \Lambda(t) \exp(\beta_1 Z + \beta_2 W)$, where we set the cumulative baseline hazard function $\Lambda(t) = 0.1t$ and the regression parameter $\beta = (\beta_1, \beta_2) = (\log 2, 0.5)$. We assume that Z is the expensive covariate that is only observed for the validation sample and W is the cheap covariate available for the full cohort. Since Z and W are independent, the method described in Section 3 can be applied directly. Suppose that the length of study is τ . For each subject, we generate a sequence of examination times as cumulative sums of $U(0, \zeta)$ until τ . If the failure time T is smaller than the first examination time (i.e. left-censored), we set U to be the first examination time, V the second examination time and $(\delta_1, \delta_2) = (1, 0)$; if T is larger than the last examination time (i.e. right-censored), we take U to be the second to the last examination time, V to be the last examination time and $(\delta_1, \delta_2) = (0, 0)$; otherwise (i.e. interval-censored), (U, V) were set to be the two consecutive examination times that bracket T and $(\delta_1, \delta_2) = (0, 1)$. The parameters (τ, ζ) used in the data generation process are specified according to the desired proportion of failures (i.e. the proportion of subjects with $\delta_1 + \delta_2 = 1$), denoted by $\Pr(\text{failure})$, which is taken as 0.2 or 0.3. Regarding the cutpoints (a_1, a_2) used to define the partitions $\{A_i: i = 1, \dots, 4\}$, we consider two sets of values that correspond to (10, 90)-th or (20, 80)-th percentiles of all observation times (U, V) of failure subjects. For each scenario, a cohort of size $N = 3000$ is generated and the simulation results are obtained based on 1000 replicates.

In the first study, the validation sample is constructed with the sizes $(n_0, n_1, n_2) = (300, 25, 25)$. Table 1 presents the simulation results from five estimators: (i) the estimator based only on the SRS component of the validation sample, denoted by $\hat{\beta}_{SRS_{n_0}}$; (ii) the estimator based on a SRS sample of the same size as the validation sample, denoted by $\hat{\beta}_{SRS_{n_v}}$; (iii) the inverse probability weighted (IPW) estimator commonly used under biased sampling, denoted by $\hat{\beta}_{IPW}$; (iv) the estimator proposed by Zhou *et al.* (2018) that uses only information in the validation (ODS) sample, denoted by $\hat{\beta}_{ODS}$; (v) the proposed estimator that uses both the validation and nonvalidation samples, denoted by $\hat{\beta}_P$. The simulation results for β include "Bias" calculated as the average point estimate minus the true value, "SSD" the sample standard deviation of point estimates, "ESE" the average estimated standard errors, "CP" the coverage proportion of 95% confidence interval, and "RE" the relative efficiency compared to $\hat{\beta}_{SRS_{n_v}}$ calculated as $(SSD(\hat{\beta}_{SRS_{n_v}})/SSD(\hat{\beta}))^2$. The results in Table 1 show that for all scenarios considered: (i) the proposed estimator $\hat{\beta}_P$ yields negligible bias, the variance estimate reflects the true variability well, and the empirical

coverage proportion of 95% confidence interval is close to 0.95; (ii) the proposed design is more efficient than the commonly used SRS designs; for example, when the failure rate is 0.2 and the cutpoints are (20, 80)-th percentiles, $\hat{\beta}_p$ achieves 43% efficiency gain for β_1 compared to $\hat{\beta}_{SRS_{n_v}}$; (iii) under the ODS design, the proposed estimator is more efficient

than the IPW and Zhou *et al.* (2018)'s estimators; (iv) the efficiency gain of the proposed method is substantial for estimating β_2 , and this is expected since the proposed method can utilize information in the nonvalidation sample in contrast to the other methods.

In the second study, we consider different sample size allocations (n_0, n_1, n_2) , including (330, 10, 10), (300, 25, 25) and (270, 40, 40). Table 2 presents the results under these sample sizes when the cutpoints are (20; 80)-th percentiles. One can see from Table 2 that (i) when the total size of validation sample is fixed, as we allocate more sizes to the supplemental components of the validation sample, the efficiency of the proposed method improves; for example, as we change (n_0, n_1, n_2) from (330, 10, 10) to (300, 25, 25), the efficiency for estimating β_1 improves by $(0.127/0.114)^2 = 1.24$ when the failure rate is 0.2; (ii) as we increase the total size of validation sample while maintain the allocations of SRS and supplemental components of the validation sample, the efficiency of the proposed method improves; for example, as we change (n_0, n_1, n_2) from (300, 25, 25) to (600, 50, 50), the efficiency improves by $(0.114/0.080)^2 = 2.03$ when the failure rate is 0.2.

In the third study, we simulate the expensive covariate Z from $N(0, 1)$ and the cheap covariate $W = I(\log |Z| + e > 0)$ with $e \sim N(0, \sigma_e^2)$, i.e., W is weakly related to Z . The proposed method described in Section 3 can easily be extended to handle this situation. Specifically, if the cheap covariate W is assumed to be available for the full cohort, the likelihood function will involve the conditional distribution $G(z|w)$ of Z given W . Since W is discrete, $G(z|w)$ can simply be estimated with the conditional empirical distribution given $W = w$. Once $G(z|w)$ is estimated and plugged into the likelihood function, we can estimate β and Λ using the sieve approach similarly as in Section 3. For this simulation study, we set the failure rate as 0.2, the cutpoints (20, 80)-th percentiles, and the sample sizes $N = 3000$ and $(n_0, n_1, n_2) = (300, 25, 25)$. Figure 3 presents the simulation results, where the horizontal axis corresponds to the value of σ_e and the vertical axis gives the sample standard deviation $SSD(\hat{\beta})$. One can see from Figure 3 that for all values of σ_e considered, the proposed method is more efficient than the competing ones.

5 An Application to a Phase 3 HIV Vaccine Trial

We illustrate the proposed method using data from the world's first phase 3 HIV-1 vaccine efficacy trial – VAX004. This trial was conducted in North America and the Netherlands in 5403 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. It was designed as a randomized, double-blinded, placebo-controlled trial and completed in 2003 (Harro *et al.*, 2004). Within this randomized 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 subject, the time to HIV-1 infection was known only 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 assess correlations between antibody responses to rgp120 and the incidence of HIV-1 infection. Gilbert *et al.* (2005) designed a case-cohort study for this purpose. Their sample consisted of all 241 infected subjects and 167, a fraction of 5%, uninfected subjects, selected from vaccine recipients by Bernoulli simple random sampling at enrollment. Eight antibody variables were measured at available pre-infection samples at the month 0, 0.5, 1, 1.5, 6, 6.5, 12, 12.5, 18, 18.5, 24, 24.5, 30, 30.5 study visits, only for the selected subjects. Lack of appropriate methods dealing with biased-sampled interval-censored data, Gilbert *et al.* (2005) conducted the statistical analysis with the HIV-1 infection date estimated as the midpoint of the last negative and first positive test dates.

Here we illustrate our design and method by analyzing the cohort of vaccine recipients at-risk for HIV-1 infection at the month 6 study visit. We focus on the functional assay MN neutralization titer, among all antibody variables, and consider its measurements up to the month 6 study visit as the main covariate of interest. Specifically, in our analysis, we use the area under the curve of MN neutralization titer values over time from month 0 to 6 as the main covariate. We consider the proportional hazards model (3) adjusting for three baseline demographic variables: age, sex and behavioral risk score. The first-stage cohort in our analysis consists of 3381 vaccine recipients at-risk for HIV-1 infection at the month 6 study visit. The second-stage ODS sample includes a SRS sample of 166 subjects from the cohort plus 50 early (before month 18) and late (after month 24) infected subjects.

We compared the proposed method with two other approaches: (i) Zhou *et al.* (2018)'s method that utilizes only the second-stage ODS sample; (ii) the inverse probability weighted (IPW) method commonly used for handling biased sample (e.g. Zhou *et al.*, 2017a). In terms of choosing the degree of Bernstein polynomial m , we suggest using AIC defined as $-2l(\hat{\theta}) + 2(p + m + 1)$, where $l(\theta)$ is the log-likelihood function and p is the number of regression parameters. Note that for the same dataset, the three methods may end up choosing different m values based on AIC, since they have different likelihood functions and yield different parameter estimates. The estimation results, including regression coefficient estimates, standard errors and p-values for testing the regression coefficients being zero, are presented in Table 3. The proposed method implies that higher MN neutralization titer value is significantly associated with lower risk of HIV-1 infection, with the regression coefficient estimate being -0.6157 , standard error 0.0556 and p-value close to zero. This is consistent with the findings of Gilbert *et al.* (2005). One can also see from Table 3 that the proposed method generally yields smaller standard errors than Zhou *et al.* (2018)'s and IPW methods; and it detects the significant predictor "age" that is not suggested by the other methods.

6 Concluding Remarks

Cost-effective sampling designs for studies concerning interval-censored failure time data are in practical need but rarely studied. In this paper, we proposed a two-stage outcome-

dependent sampling design for interval-censored failure time data, called two-stage interval-censoring ODS design, where we enhance the study efficiency by allowing the selection probability of the second-stage sample to depend on the first-stage observed interval-censored failure time outcomes. We developed a matching semiparametric likelihood-based inference procedure that would reap the benefits gained from the proposed design. Our design together with the inference procedure provides investigators, who deal with interval-censored failure time data, an effective statistical tool that can reduce the study cost and improve the study efficiency at a fixed budget level.

We make a few remarks on the proposed design and inference procedure. First, to implement the proposed design, one needs to determine the cutpoints and the allocations of SRS and supplemental components. Based on our numerical experience, the proposed method works reasonably well by choosing the cutpoints as the 10-th to 35-th lower and upper percentiles of all observation times and by allocating the SRS and supplemental components with a range of ratios from 3:1 to 10:1. One may expect to gain more statistical efficiency by assigning larger sizes to supplemental components or even completely removing the SRS. However, in practice, the study investigators usually prefer to keep a fairly large SRS component that is representative of the whole population (like in the case-cohort study). Our recommendation on the design parameters is based on our numerical experience and inputs from practitioners. Deriving a theoretically optimal two-stage design is not an easy task. It is not within the scope of this paper and will be pursued in future research. Second, to implement the proposed inference procedure, one needs to determine the degree of Bernstein polynomials $m = \alpha(N^\nu)$. Note from Theorem 1 that the optimal rate $n/[2(1+r)]$ is achieved if $(1-\nu)/2 = r\nu/2$, i.e., $\nu = 1/(1+r)$. Hence, theoretically, $m = \alpha(N^{1/(1+r)})$ is an optimal choice. Nevertheless, for a practical problem with a finite sample size, it is hard to determine the value of m (if exists) that is theoretically optimal. Here are two reasons: first, the smoothness parameter r is unknown; second, even if r can be reasonably estimated, m still cannot be determined from the formula $m = \alpha(N^{1/(1+r)})$. Thus, in practice, we recommend to use AIC to select m . Lastly, the proposed pseudo likelihood procedure may not be the most efficient method. We are working towards the direction of semiparametrically efficient estimation, though this seems to be challenging both theoretically and computationally.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank the Editor, Associate Editor and reviewers for their helpful comments and suggestions that have improved the paper. The authors also thank the Global Solutions in Infectious Diseases (GSID) and Dr. Peter Gilbert for providing data from the phase 3 HIV vaccine trial VAX004. This research was partially supported by grants from the National Institutes of Health (R01ES021900, P01CA142538 and P30ES010126). Qingning Zhou's work was supported, in part, by funds provided by the University of North Carolina at Charlotte.

Appendix:

Proofs of Theorems 1 and 2

In the appendix, we sketch the proofs of Theorems 1 and 2. Let $O = \{Y = \{U, V, \delta_1 = I(T < U), \delta_2 = I(U < T - V)\}, RZ, R\}$ denote a single observation, where U and V are two random examination times, Z is the p -dimensional covariate vector and R is the indicator of an observation being in the validation sample. The following regularity conditions are needed for proving the theorems:

- (C1) There exists $\eta > 0$ such that $P(V - U \leq \eta) = 1$. The union of the supports of U and V is contained in the interval $[\sigma, \tau]$, where $0 < \sigma < \tau < \infty$.
- (C2) The distribution of Z , denoted by $G_Z(z)$, has a bounded support and is not concentrated on any proper subspace of R^p .
- (C3) For $r = 1$ or 2 , the function $\Lambda_0(t) \in \mathcal{M}$ is continuously differentiable up to order r in $[\sigma, \tau]$ with the first derivative being strictly positive, and satisfies $\alpha^{-1} < \Lambda_0(\sigma) < \Lambda_0(\tau) < \alpha$ for some positive constant α . Also β_0 is an interior point of $\mathcal{B}$, a compact subset of R^p . $\mathcal{M}$ and $\mathcal{B}$ are defined in Section 3.
- (C4) The conditional density $g(u, v|z)$ of (u, v) given z has bounded partial derivatives with respect to u and v , and the bounds of these partial derivatives do not depend on (u, v, z) .
- (C5) $E\{\text{var}(Z|U)\}$ and $E\{\text{var}(Z|V)\}$ are positive definite.

These conditions are commonly used in the studies of interval-censored data (e.g. Huang and Wellner, 1997; Huang and Rossini, 1997; Zhang *et al.*, 2010). In addition, similarly as in Zeng and Lin (2014), one can show that the conclusions of Theorems 1&2 hold under the proposed sampling scheme if they hold under independent sampling. Specifically, Zeng and Lin (2014) in their Appendix established the following result under general two-phase cohort studies based on Le Cam's third lemma: the consistency and asymptotic normality of MLE hold under the sampling mechanism satisfying their condition (C.6) if they hold under independent sampling. It is easy to verify that our two-stage ODS scheme satisfies the condition (C.6) in Zeng and Lin (2014) with

$$p(1; y) = I(y \in A_1)[p_0 + (1 - p_0)p_1] + I(y \in A_2)[p_0 + (1 - p_0)p_2] + I(y \in A_3)p_0 + I(y \in A_4)p_0,$$

where $y = \{u, v, \delta_1, \delta_2\}$ is the outcome, $\{A_1, A_2, A_3, A_4\}$ are the four strata defined based on the outcome given in (1), p_0 is the sampling fraction of the SRS component, and p_1 and p_2 are the sampling fraction of the supplemental components from the two tail strata A_1 and A_2 respectively. Thus, we assume in the following that the observations $\{O_i, i = 1, \dots, N\}$ are independent and identically distributed.

Note that the proofs of our theorems differ from those of Zhou *et al.* (2017b, 2018) in several aspects. First, our likelihood function is not exact, since an estimate of the covariate distribution rather than the true one is used. Thus, we have to deal with the difference

between our approximate likelihood and the exact likelihood that assumes the covariate distribution to be known. For establishing consistency and rate of convergence, the proofs of our theorems follow the similar ideas as those in Zhou *et al.* (2017b, 2018), except that we need to additionally establish the closeness of the approximate and exact likelihoods. For establishing asymptotic normality and deriving the asymptotic covariance matrix, our approach is quite different from those in Zhou *et al.* (2017b, 2018), since we have to account for the additional variability induced by the estimated covariate distribution.

Before proving Theorems 1&2, we first define the class of functions $\mathcal{L}_N = \{l(\theta, O): \theta \in \Theta_N\}$, where $l(\theta, O)$ is the log-likelihood function based on a single observation O given by

$$\begin{aligned} l(\theta, O) &= R \log f(Y|Z; \theta) + (1 - R) \log f_Y(Y) \\ &= R \log f(Y|Z; \theta) + (1 - R) \log \int f(Y|z; \theta) dG_Z(z), \end{aligned}$$

where the covariate distribution $G_Z(z)$ is assumed to be known and

$$f(Y|Z; \theta) = (1 - S(U|Z))^{\Delta_1} 1_{(S(U|Z) - S(V|Z))}^{\Delta_2} 2_{S(V|Z)}^{1 - \Delta_1 - \Delta_2}$$

with $S(t|Z) = \exp(-\Lambda(t)\theta'Z)$. Let P_N denote the empirical measure. For any $\epsilon > 0$, we define the covering number $N(\epsilon, \mathcal{L}_N, L_1(P_N))$ as the smallest value of k for which there exists $\{\theta^{(1)}, \dots, \theta^{(k)}\} \in \Theta_N$ such that

$$\min_{j \in \{1, \dots, k\}} \frac{1}{N} \sum_{i=1}^N |l(\theta, O_i) - l(\theta^{(j)}, O_i)| < \epsilon$$

for all $\theta \in \Theta_N$. If no such k exists, define $N(\epsilon, \mathcal{L}_N, L_1(P_N)) = \infty$.

Proof of Theorem 1

We now prove the strong consistency of $\hat{\theta}_N$. Based on Lemma 1 in the Supplementary Materials, the covering number of $\mathcal{L}_N$ satisfies

$$N(\epsilon, \mathcal{L}_N, L_1(P_N)) \leq KM_N^{(m+1)} \epsilon^{-(p+m+1)}.$$

Then by Lemma 2 in the Supplementary Materials, we have

$$\sup_{\theta \in \Theta_N} |P_N l(\theta, O) - Pl(\theta, O)| \rightarrow 0 \quad \text{almost surely.} \quad (\text{A.1})$$

Furthermore, define

$$\begin{aligned}
p_N \hat{l}(\theta, O) &= \frac{1}{N} \sum_{i=1}^N \left\{ R_i \log f(Y_i | Z_i; \theta) + (1 - R_i) \log \int f(Y_i | z; \theta) d\hat{G}_Z(z) \right\} \\
&= \frac{1}{N} \sum_{i=1}^N \left[R_i \log f(Y_i | Z_i; \theta) + (1 - R_i) \log \left\{ \sum_{k=1}^4 \frac{N_k}{N(n_k + n_{0k})} \sum_{r \in I_{V_k}} f(Y_i | Z_r; \theta) \right\} \right].
\end{aligned}$$

Then it is easy to show that

$$\sup_{\theta \in \Theta_N} |P_N \hat{l}(\theta, O) - P_N l(\theta, O)| \rightarrow 0 \quad \text{almost surely.} \quad (\text{A.2})$$

Let $M(\theta, O) = -l(\theta, O)$ and $P_N \hat{M}(\theta, O) = -P_N \hat{l}(\theta, O)$. Define $K_\epsilon = \{\theta: d(\theta, \theta_0) \leq \epsilon, \theta \in \Theta_N\}$ for $\epsilon > 0$ and

$$\zeta_{1N} = \sup_{\theta \in \Theta_N} |P_N M(\theta, O) - PM(\theta, O)|,$$

$$\zeta_{2N} = \sup_{\theta \in \Theta_N} |P_N \hat{M}(\theta, O) - P_N M(\theta, O)|,$$

$$\zeta_{3N} = P_N \hat{M}(\theta_0, O) - PM(\theta_0, O).$$

Then we obtain

$$\inf_{K_\epsilon} PM(\theta, O) \leq \zeta_{1N} + \zeta_{2N} + \inf_{K_\epsilon} P_N \hat{M}(\theta, O). \quad (\text{A.3})$$

If $\hat{\theta}_N \in K_\epsilon$, we have

$$\inf_{K_\epsilon} P_N \hat{M}(\theta, O) = P_N \hat{M}(\hat{\theta}_N, O) \leq P_N \hat{M}(\hat{\theta}_0, O) = \zeta_{3N} + PM(\theta_0, O). \quad (\text{A.4})$$

Define $\delta_\epsilon = \inf_{K_\epsilon} PM(\theta, O) - PM(\theta_0, O)$. Then under Conditions (C1) - (C5), using the same arguments as those in Zhang *et al.* (2010, p. 352), we can prove $\delta_\epsilon > 0$. It follows from (A.3) and (A.4) that

$$\inf_{K_\epsilon} PM(\theta, O) \leq \zeta_{1N} + \zeta_{2N} + \zeta_{3N} + PM(\theta_0, O) = \zeta_N + PM(\theta_0, O)$$

with $\zeta_N = \zeta_{1N} + \zeta_{2N} + \zeta_{3N}$, and hence $\zeta_N \leq \delta_\epsilon$. This gives $\{\hat{\theta}_N \in K_\epsilon\} \subseteq \{\zeta_N \leq \delta_\epsilon\}$, and by (A.1), (A.2) and the strong law of large numbers, we have $\zeta_N \rightarrow 0$ almost surely. Therefore,

$\cup_{k=1}^{\infty} \cap_{N=k}^{\infty} \{\hat{\theta}_N \in K_\epsilon\} \subseteq \cup_{k=1}^{\infty} \cap_{N=k}^{\infty} \{\zeta_N \geq \delta_\epsilon\}$, which proves that $d(\hat{\theta}_N, \theta_0) \rightarrow 0$ almost surely.

Now we will derive the rate of convergence by using Theorem 3.4.1 of van der Vaart and Wellner (1996). Below let $\tilde{K}$ denote a universal positive constant that may differ from place to place. First note from Theorem 1.6.2 of Lorentz (1986) that there exists a Bernstein polynomial Λ_{N0} such that $\|\Lambda_{N0} - \Lambda_0\|_\infty = O(m^{-r/2}) = O(N^{-rv/2})$. Define $\theta_{N0} = (\beta_0, \Lambda_{N0})$, then $d(\theta_{N0}, \theta_0) = O(N^{-rv/2})$. For any $\eta > 0$; define the class of functions

$\mathcal{F}_\eta = \{l(\theta, O) - l(\theta_{N0}, O) : \theta \in \Theta_N, \eta/2 < d(\theta, \theta_{N0}) \leq \eta\}$. One can easily show that

$P(l(\theta_0, O) - l(\theta_{N0}, O)) \leq \tilde{K}d(\theta_0, \theta_{N0}) \leq \tilde{K}N^{-rv/2}$. Also under Condition (C1)-(C5), using the

same arguments as those in Zhang *et al.* (2010, p. 352), we obtain

$P(l(\theta_0, O) - l(\theta, O)) \geq \tilde{K}d^2(\theta_0, \theta)$. Therefore, for large N , we have

$P(l(\theta, O) - l(\theta_{N0}, O)) = P(l(\theta, O) - l(\theta_0, O)) + P(l(\theta_0, O) - l(\theta_{N0}, O)) \leq -\tilde{K}\eta^2 + \tilde{K}N^{-rv/2} = -\tilde{K}\eta^2$

, for any $l(\theta, O) - l(\theta_{N0}, O) \in \mathcal{F}_\eta$.

Following the calculations in Shen and Wong (1994, p. 597), we have that for $0 < \epsilon < \eta$,

$\log N_{[]}(\epsilon, \mathcal{F}_\eta, L_2(P)) \leq \tilde{K}(m+1)\log(\eta/\epsilon)$. Some algebraic manipulations give

$P(l(\theta, O) - l(\theta_{N0}, O))^2 \leq \tilde{K}\eta^2$ for any $l(\theta, O) - l(\theta_{N0}, O) \in \mathcal{F}_\eta$. Also under Conditions (C1) -

(C4), $\mathcal{F}_\eta$ is uniformly bounded. Then by Lemma 3.4.2 of van der Vaart and Wellner (1996),

we obtain

$$E_P \|N^{1/2}(P_N - P)\|_{\mathcal{F}_\eta} \leq \tilde{K}J_{[]}(\eta, \mathcal{F}_\eta, L_2(P)) \left(1 + \frac{J_{[]}(\eta, \mathcal{F}_\eta, L_2(P))}{\eta^2 N^{1/2}} \right)$$

with

$$J_{[]}(\eta, \mathcal{F}_\eta, L_2(P)) = \int_0^\eta \left\{ 1 + \log N_{[]}(\epsilon, \mathcal{F}_\eta, L_2(P)) \right\}^{1/2} d\epsilon \leq \tilde{K}(m+1)^{1/2} \eta.$$

Then we have

$$\begin{aligned} & \sup_{\eta/2 < d(\theta, \theta_{N0}) \leq \eta, \theta \in \Theta_N} N^{1/2} [(P_N \hat{l}(\theta, O) - Pl(\theta, O)) - (P_N \hat{l}(\theta_{N0}, O) - Pl(\theta_{N0}, O))] \\ &= \sup_{\eta/2 < d(\theta, \theta_{N0}) \leq \eta, \theta \in \Theta_N} N^{1/2} [\{ (P_N \hat{l}(\theta, O) - P_N l(\theta, O)) - (P_N \hat{l}(\theta_{N0}, O) - P_N l(\theta_{N0}, O)) \} \\ & \quad + \{ (P_N l(\theta, O) - Pl(\theta, O)) - (P_N l(\theta_{N0}, O) - Pl(\theta_{N0}, O)) \}] \\ &\leq 2N^{1/2} \sup_{\theta \in \Theta_N} |P_N \hat{l}(\theta, O) - P_N l(\theta, O)| + E_P \|N^{1/2}(P_N - P)\|_{\mathcal{F}_\eta} \\ &\leq \tilde{K}\{(m+1)^{1/2}\eta + (m+1)N^{-1/2}\}. \end{aligned}$$

Define $\phi_N(\eta) = (m+1)^{1/2}\eta + (m+1)N^{-1/2}$. It is obvious that $\phi_N(\eta)/\eta$ is decreasing in η . Let $r_N = N^{\min\{(1-\nu)/2, r\sqrt{2}\}}$, then $r_N^2 \phi_N(1/r_N) = r_N(N^\nu + 1)^{1/2} + r_N^2(N^\nu + 1)N^{-1/2} \leq \tilde{K}N^{1/2}$.

Note that $d(\hat{\theta}_N, \theta_{N0}) \leq d(\hat{\theta}_N, \theta_0) + d(\theta_N, \theta_{N0}) \rightarrow 0$ in probability. It then follows from Theorem 3.4.1 of van der Vaart and Wellner (1996) that $r_N d(\hat{\theta}_N, \theta_{N0}) = O_p(1)$. Furthermore, by $d(\theta_{N0}, \theta_0) = O(N^{-r\sqrt{2}})$, we have $r_N d(\hat{\theta}_N, \theta_0) \leq r_N d(\hat{\theta}_N, \theta_{N0}) + r_N d(\theta_{N0}, \theta_0) = O_p(1)$ which completes the proof.

Proof of Theorem 2

We now sketch the proof of the asymptotic normality of $\hat{\beta}_N$. Let $l_{\beta}(\theta, O)$ denote the score for β given by

$$l_{\beta}(\theta, O) = R \frac{f_{\beta}(Y|Z; \theta)}{f(Y|Z; \theta)} + (1-R) \frac{\int f_{\beta}(Y|z; \theta) dG_Z(z)}{\int f(Y|z; \theta) dG_Z(z)},$$

where

$$f_{\beta}(Y|Z; \theta) = \left(1 - S_{\beta}(U|Z)\right)^{\Delta_1} \left(S_{\beta}(U|Z) - S_{\beta}(V|Z)\right)^{\Delta_2} S_{\beta}(V|Z)^{1 - \Delta_1 - \Delta_2}$$

and

$$S_{\beta}(t|Z) = \exp(-\Lambda(t)e^{\beta'Z}) / (-\Lambda(t)e^{\beta'Z})Z.$$

Consider a parametric smooth submodel with parameter $\theta_{(s)} = (\beta, \Lambda_{(s)})$, where $\Lambda_{(0)} = \Lambda$ and

$$\left. \frac{\partial \Lambda_{(s)}}{\partial s} \right|_{s=0} = h.$$

Let $\mathcal{H} \subseteq \mathcal{L}_2(P)$ denote the class of functions h defined by this equation. The score operator for Λ is

$$\begin{aligned} l_{\Lambda}(\theta, O)[h] &= \left. \frac{\partial l(\theta_{(s)}, O)}{\partial s} \right|_{s=0} \\ &= R \frac{f_{\Lambda}(Y|Z; \theta)[h]}{f(Y|Z; \theta)} + (1-R) \frac{\int f_{\Lambda}(Y|z; \theta)[h] dG_Z(z)}{\int f(Y|z; \theta) dG_Z(z)}, \end{aligned}$$

where

$$f_{\Lambda}(Y|Z; \theta)[h] = (1 - S_{\Lambda}(U|Z)[h])^{\Delta_1} (S_{\Lambda}(U|Z)[h] - S_{\Lambda}(V|Z)[h])^{\Delta_2} S_{\Lambda}(V|Z)[h]^{1 - \Delta_1 - \Delta_2}$$

and

$$S_{\Lambda}(t|Z)[h] = \exp(-\Lambda(t)e^{\beta'Z})(-e^{\beta'Z})_{h(t)}.$$

According to Bickel *et al.* (1993), the efficient score for β is

$$l^*(\theta, O) = l_{\beta}(\theta, O) - l_{\Lambda}(\theta, O)[h_0],$$

where $h_0 \in \mathcal{H}^p$ minimizes $P\|l_{\beta}(\theta, O) - l_{\Lambda}(\theta, O)[h]\|^2$ and is called the least favorable direction.

Then the information for β is

$$I(\beta) = P[l^*(\theta, O) \otimes^2] = P[l_{\beta}(\theta, O) - l_{\Lambda}(\theta, O)[h_0] \otimes^2],$$

where $v \otimes^2 = vv'$ for a column vector $v \in R^p$. Under Conditions (C1)-(C5), the existence of the least favorable direction and the positive definiteness of the information can be similarly established as Theorem 4.1 in Huang and Wellner (1997).

Since $\hat{\theta}_N$ maximizes $P_N \hat{l}(\theta, O)$ which is obtained by replacing G_Z in $P_N l(\theta, O)$ with its consistent estimator $\hat{G}_Z$, $\hat{\theta}_N$ is the solution to the functional equation $P_N \hat{l}^*(\theta, O) = 0$, where $P_N \hat{l}^*(\theta, O)$ is obtained by replacing G_Z in $P_N l^*(\theta, O)$ with $\hat{G}_Z$. First note that $\hat{G}_Z$ is a $\sqrt{N}$ -consistent estimator of G_Z . Similarly as the proofs of Theorem 2 in Zhang *et al.* (2010), one can establish that

$$N^{1/2}(\hat{\beta}_N - \beta_0) = N^{1/2}I^{-1}(\beta_0)P_N \hat{l}^*(\theta_0, O) + o_p(1).$$

Following the proofs of Theorem 2 in Weaver and Zhou (2005), one can further establish that

$$N^{1/2}(\hat{\beta}_N - \beta_0) \rightarrow N(0, \Sigma)$$

in distribution with the asymptotic covariance matrix given by

$$\Sigma = I^{-1}(\beta_0) + \sum_{l=1}^4 \frac{\pi_l^2}{\rho_l \rho_v + \pi_l \rho_0 \rho_v} I^{-1}(\beta_0) \Sigma_l(\theta_0) I^{-1}(\beta_0).$$

Here $I(\beta)$ is the information for β defined above, and

$$\Sigma_I(\theta) = \text{var} \left\{ Z|Y \in A_{[I]} \right\} \left\{ \sum_{k=1}^4 [\pi_k(1 - \rho_{0v}) - \rho_k \rho_v] E \left\{ Y|Y \in A_k \right\} M_Z(Y; \theta) \right\},$$

where

$$M_Z(Y; \theta) = \frac{\partial f(Y|Z; \theta) / \partial \theta}{f_Y(Y; \theta)} - \frac{\partial f_Y(Y; \theta) / \partial \theta}{[f_Y(Y; \theta)]^2} f(Y|Z; \theta),$$

with

$$\frac{\partial f(Y|Z; \theta)}{\partial \theta} \equiv f_{\beta}(Y|Z; \theta) - f_{\Lambda}(Y|Z; \theta)[h_0] \text{ and } \frac{\partial f_Y(Y; \theta)}{\partial \theta} \equiv \int \frac{\partial f(Y|z; \theta)}{\partial \theta} dG_Z(z).$$

References

- Bickel PJ, Klaassen CA, Ritov Y, and Wellner JA (1993). Efficient and adaptive estimation for semiparametric models Johns Hopkins University Press: Baltimore.
- Chatterjee N, Chen Y-H, and Breslow NE (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Am. Statist. Ass.*, 98(461), 158–168.
- Chen D-G, Sun J, and Peace KE (2012). Interval-Censored Time-to-Event Data: Methods and Applications CRC Press.
- Chen K and Lo S-H (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika*, 86(4), 755–764.
- Cornfield J (1951). A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast, and cervix. *J. Natn. Cancer Inst.*, 11(6), 1269–1275.
- Ding J, Zhou H, Liu Y, Cai J, and Longnecker MP (2014). Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biostatistics*, 15(4), 636–650. [PubMed: 24812419]
- Ding J, Lu T-S, Cai J, and Zhou H (2017). Recent progresses in outcome-dependent sampling with failure time data. *Lifetime Data Anal.*, 23(1), 57–82. [PubMed: 26759313]
- Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, Sinangil F, Burke D, and Berman PW (2005). 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(5), 666–677. [PubMed: 15688279]
- Harro CD, Judson FN, Gorse GJ, Mayer KH, Kostman JR, Brown SJ, Koblin B, Marmor M, Bartholow BN, Popovic V, et al. (2004). Recruitment and baseline epidemiologic profile of participants in the first phase 3 HIV vaccine efficacy trial. *J. of Acquir. Immune De c. Syndr.*, 37(3), 1385–1392.
- Huang J and Rossini A (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Statist. Ass.*, 92(439), 960–967.
- Huang J and Wellner JA (1997). Interval censored survival data: a review of recent progress In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer.
- Huang J, Zhang Y, and Hua L (2012). Consistent variance estimation in semiparametric models with application to interval-censored data. *Interval-Censored Time-to-Event Data: Methods and Applications*, eds. Chen DG, Sun J and Peace KE, pages 233–268.
- Kang S and Cai J (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, 96(4), 887–901. [PubMed: 23946547]

- Kulich M and Lin D (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Am. Statist. Ass.*, 99(467), 832–844.
- Li Z and Nan B (2011). Relative risk regression for current status data in case-cohort studies. *Canad. J. Statist.*, 39(4), 557–577.
- Li Z, Gilbert P, and Nan B (2008). Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics*, 64(4), 1247–1255. [PubMed: 19032178]
- Lorentz GG (1986). *Bernstein Polynomials* New York: Chelsea Publishing Co.
- Prentice RL (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1), 1–11.
- Self SG and Prentice RL (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, 16(1), 64–81.
- Shen X and Wong WH (1994). Convergence rate of sieve estimates. *Ann. Statist.*, 22(2), 580–615.
- Song R, Zhou H, and Kosorok MR (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, 96(1), 221–228. [PubMed: 20107493]
- Sun J (2006). *The Statistical Analysis of Interval-Censored Failure Time Data* New York: Springer.
- Sun Y, Qian X, Shou Q, and Gilbert PB (2017). Analysis of two-phase sampling data with semiparametric additive hazards models. *Lifetime Data Anal.*, 23(3), 377–399. [PubMed: 26995733]
- van der Vaart AW and Wellner JA (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics* New York: Springer.
- Weaver MA and Zhou H (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *J. Am. Statist. Ass.*, 100(470), 459–469.
- Whittemore AS (1997). Multistage sampling designs and estimating equations. *J. R. Statist. Soc. B.*, 59(3), 589–602.
- Xue H, Lam K, and Li G (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *J. Am. Statist. Assoc.*, 99(466), 346–356.
- Yu J, Liu Y, Sandler DP, and Zhou H (2015). Statistical inference for the additive hazards model under outcome-dependent sampling. *Canad. J. Statist.*, 43(3), 436–453.
- Zeng D and Lin DY (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Am. Statist. Ass.*, 109(505), 371–383.
- Zhang Y, Hua L, and Huang J (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scand. J. Statist.*, 37(2), 338–354.
- Zhou H, Weaver M, Qin J, Longnecker M, and Wang M (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2), 413–421. [PubMed: 12071415]
- Zhou H, Song R, Wu Y, and Qin J (2011). Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. *Biometrics*, 67(1), 194–202. [PubMed: 20560938]
- Zhou Q, Zhou H, and Cai J (2017a). Case-cohort studies with interval-censored failure time data. *Biometrika*, 104(1), 17–29. [PubMed: 28943643]
- Zhou Q, Hu T, and Sun J (2017b). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *J. Am. Statist. Ass.*, 112(518), 664–672.
- Zhou Q, Cai J, and Zhou H (2018). Outcome-dependent sampling with interval-censored failure time data. *Biometrics*, 74(1), 58–67. [PubMed: 28771664]

Figure 1:
An illustration of the partitions $\{A_i: i=1, \dots, 4\}$ under the two-stage interval-censoring ODS design

Figure 2:
An illustration of the data structure under the two-stage interval-censoring ODS design

Figure 3:

Comparison of different methods on the regression parameter estimation under the model $\Lambda(t|Z, W) = \Lambda(t) \exp(\beta_1 Z + \beta_2 W)$, where $\Lambda(t) = 0.1t$, the expensive covariate Z is generated from $N(0, 1)$ and the cheap covariate $W = I(\log |Z| + e > 0)$ with $e \sim N(0, \sigma_e^2)$.

Table 1:

Simulation results for the estimation of β when the full cohort size is $N = 3000$ and the validation sample sizes are $(n_0, n_1, n_2) = (300, 25, 25)$

Pr(failure)	cutpoints		$\beta_1 = \log 2$					$\beta_2 = 0.5$				
			Bias	SSD	ESE	CP	RE	Bias	SSD	ESE	CP	RE
0.2	(20%, 80%)	$\hat{\beta}_{SRS_{n_0}}$	0.015	0.148	0.141	0.94	0.85	0.014	0.141	0.138	0.94	0.77
		$\hat{\beta}_{SRS_{n_v}}$	0.010	0.137	0.130	0.94	1.00	0.004	0.124	0.126	0.96	1.00
		$\hat{\beta}_{IPW}$	0.014	0.131	0.124	0.94	1.10	0.011	0.125	0.122	0.94	0.97
		$\hat{\beta}_{ODS}$	-0.006	0.117	0.123	0.94	1.37	-0.004	0.110	0.116	0.95	1.25
		$\hat{\beta}_P$	0.009	0.115	0.110	0.94	1.43	0.000	0.046	0.046	0.95	7.07
	(10%, 90%)	$\hat{\beta}_{SRS_{n_0}}$	0.010	0.142	0.141	0.95	0.78	0.010	0.146	0.137	0.94	0.77
		$\hat{\beta}_{SRS_{n_v}}$	0.009	0.126	0.130	0.96	1.00	0.003	0.128	0.126	0.95	1.00
		$\hat{\beta}_{IPW}$	0.010	0.131	0.130	0.94	0.92	0.012	0.134	0.127	0.94	0.91
		$\hat{\beta}_{ODS}$	-0.010	0.115	0.125	0.96	1.21	-0.005	0.119	0.116	0.94	1.16
		$\hat{\beta}_P$	0.002	0.112	0.109	0.94	1.27	0.002	0.045	0.046	0.96	8.06
0.3	(20%, 80%)	$\hat{\beta}_{SRS_{n_0}}$	0.019	0.120	0.117	0.95	0.80	0.008	0.114	0.113	0.95	0.88
		$\hat{\beta}_{SRS_{n_v}}$	0.009	0.108	0.108	0.95	1.00	0.007	0.107	0.104	0.95	1.00
		$\hat{\beta}_{IPW}$	0.015	0.108	0.105	0.95	1.00	0.009	0.104	0.102	0.95	1.07
		$\hat{\beta}_{ODS}$	-0.010	0.100	0.102	0.95	1.15	-0.010	0.095	0.099	0.95	1.26
		$\hat{\beta}_P$	0.011	0.100	0.097	0.95	1.16	0.003	0.041	0.040	0.95	6.84
	(10%, 90%)	$\hat{\beta}_{SRS_{n_0}}$	0.009	0.121	0.116	0.95	0.73	0.008	0.113	0.112	0.96	0.88
		$\hat{\beta}_{SRS_{n_v}}$	0.007	0.103	0.107	0.96	1.00	0.009	0.106	0.104	0.95	1.00
		$\hat{\beta}_{IPW}$	0.010	0.113	0.108	0.94	0.84	0.008	0.106	0.106	0.95	1.00
		$\hat{\beta}_{ODS}$	-0.015	0.102	0.106	0.94	1.03	-0.009	0.098	0.102	0.95	1.16
		$\hat{\beta}_P$	0.005	0.098	0.096	0.95	1.11	0.002	0.041	0.040	0.94	6.82

$\hat{\beta}_{SRS_{n_0}}$, the estimator based only on the SRS component of the validation sample; $\hat{\beta}_{SRS_{n_v}}$ the estimator based on a SRS sample of the same size as the validation sample; $\hat{\beta}_{IPW}$, the inverse probability weighted estimator; $\hat{\beta}_{ODS}$, the estimator developed by Zhou *et al.* (2018) that uses only the validation (ODS) sample; $\hat{\beta}_P$, the proposed estimator that uses both the validation and nonvalidation samples. The model is given by $\Lambda(t|Z, W) = \Lambda(t) \exp(\beta_1 Z + \beta_2 W)$, where $\Lambda(t) = 0.1t$, and the covariates Z (expensive) and W (cheap) are generated independently from $N(0, 1)$.

Table 2:

Simulation results for the estimation of β based on the proposed method when the full cohort size is $N = 3000$ and the cutpoints are (20, 80)-th percentiles

(n_0, n_1, n_2)	Pr(failure)	$\beta_1 = \log 2$				$\beta_2 = 0.5$			
		Bias	SSD	ESE	CP	Bias	SSD	ESE	CP
(330; 10; 10)	0.2	0.004	0.127	0.119	0.95	0.001	0.048	0.047	0.94
	0.3	0.005	0.104	0.101	0.93	0.002	0.041	0.040	0.95
(300; 25; 25)	0.2	0.002	0.114	0.109	0.94	0.003	0.046	0.046	0.95
	0.3	0.003	0.097	0.096	0.95	0.001	0.040	0.040	0.95
(270; 40; 40)	0.2	0.004	0.110	0.102	0.93	0.003	0.047	0.046	0.95
	0.3	0.003	0.096	0.093	0.93	0.003	0.040	0.040	0.95
(600; 50; 50)	0.2	0.002	0.080	0.077	0.95	0.001	0.047	0.045	0.94
	0.3	0.001	0.070	0.068	0.95	0.001	0.039	0.038	0.95

Table 3:

Analysis results for the HIV vaccine trial data

Variable	Proposed method			Zhou et al. (2018, Biometrics)			IPW method		
	$\hat{\beta}$	SE	P-value	$\hat{\beta}$	SE	P-value	$\hat{\beta}$	SE	P-value
MN Neutralization Titer	-0.6157	0.0556	0.0000	-0.3026	0.0870	0.0005	-0.2273	0.1306	0.0818
Behavioral Risk Score	0.3772	0.0965	0.0001	0.3035	0.1492	0.0419	0.5961	0.2386	0.0125
Age	-0.0287	0.0134	0.0320	-0.0259	0.0373	0.4879	0.0156	0.0413	0.7054
Sex	-0.9596	2.6406	0.7163	-0.7452	1.3195	0.5723	-0.7793	2.0419	0.7027