Published in final edited form as:
Biometrics. 2018 March ; 74(1): 58–67. doi:10.1111/biom.12744.

Outcome-Dependent Sampling with Interval-Censored Failure Time Data

Qingning Zhou, **Jianwen Cai**, and **Haibo Zhou**[*]
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Summary

Epidemiologic studies and disease prevention trials often seek to relate an exposure variable to a failure time that suffers from interval-censoring. When the failure rate is low and the time intervals are wide, a large cohort is often required so as to yield reliable precision on the exposure-failure-time relationship. However, large cohort studies with simple random sampling could be prohibitive for investigators with a limited budget, especially when the exposure variables are expensive to obtain. Alternative cost-effective sampling designs and inference procedures are therefore desirable. We propose an outcome-dependent sampling (ODS) design with interval-censored failure time data, where we enrich the observed sample by selectively including certain more informative failure subjects. We develop a novel sieve semiparametric maximum empirical likelihood approach for fitting the proportional hazards model to data from the proposed interval-censoring ODS design. This approach employs the empirical likelihood and sieve methods to deal with the infinite-dimensional nuisance parameters, which greatly reduces the dimensionality of the estimation problem and eases the computation difficulty. The consistency and asymptotic normality of the resulting regression parameter estimator are established. The results from our extensive simulation study show that the proposed design and method works well for practical situations and is more efficient than the alternative designs and competing approaches. An example from the Atherosclerosis Risk in Communities (ARIC) study is provided for illustration.

Keywords

Biased sampling; Empirical likelihood; Interval-censoring; Semiparametric inference; Sieve estimation

1. Introduction

In many epidemiologic studies and disease prevention trials, the outcome of interest is a failure time that suffers from interval-censoring, i.e., the failure time cannot be exactly observed but only an interval that it belongs to is known or observed (e.g. Sun, 2006; Chen et al., 2012). One example of interval-censored failure time data arises from HIV preventive

[*] zhou@bios.unc.edu.

7. Supplementary Materials
The Web Appendix referenced in Section 3.3, Web Table 1 referenced in Section 4, and codes for the proposed method are available at the *Biometrics* website on Wiley Online Library.

vaccine trials where investigators are interested in assessing the association between antibody responses to a vaccine and the incidence of HIV infection (e.g. Gilbert et al., 2005). In this case, since the study subjects are tested for HIV infection only at discrete clinic visits instead of being continuously monitored, the time to HIV infection is known only to fall between two consecutive visits rather than being exactly observed and thus only interval-censored data on the infection time are available. When the failure rate is low and the observation time intervals are wide, such as in the HIV vaccine trials mentioned above, a large cohort is often required so as to yield reliable precision on the exposure-failure-time relationship. Compounding to this issue, measurements of the exposure variable of interest are often expensive or difficult to obtain, such as the antibody levels measured by complex assays in the HIV vaccine trials above. As a consequence, large cohort studies with simple random sampling could be prohibitively expensive to conduct for investigators with a limited budget. Alternative cost-effective sampling designs and inference procedures with interval-censored failure time data are therefore desirable, and this motivates the research in this paper.

Outcome-dependent sampling (ODS) is a cost-effective sampling scheme that enhances the efficiency and reduces the cost of a study by allowing the probability of acquiring the exposure measurement to depend on the observed value of the outcome. The case-control study with a binary outcome is a simple and well-known example of ODS design and it has been extensively studied and used over the past decades (e.g. Cornfield, 1951; Whittemore, 1997). In recent years, the more general ODS design with a continuous outcome has been an important research area (e.g. Zhou et al., 2002; Chatterjee et al., 2003; Weaver and Zhou, 2005). The fundamental idea of such a design is to oversample observations from the segments of the population, usually the two tails of the response variable’s distribution, that are believed to be more informative regarding the exposure-response relationship. Recent references on ODS design with a continous outcome include Zhou et al. (2007), Song et al. (2009) and Zhou et al. (2011), among others. The case-cohort design is a well-known biased-sampling scheme for censored failure time data. Under this design, measurements of the exposure are obtained for a random sample of the study cohort, called subcohort, and for all subjects who experience the failure regardless of whether or not they are in the subcohort (e.g. Prentice, 1986; Self and Prentice, 1988; Chen and Lo, 1999; Lu and Tsiatis, 2006; Kong and Cai, 2009; Zeng and Lin, 2014). When the failure is non-rare, the generalized case-cohort design has been proposed where besides a subcohort, the exposure measurements are assembled only on a subset of the failure subjects instead of all failure subjects (e.g. Cai and Zeng, 2007; Kang and Cai, 2009). Reaping the benefits of both ODS and case-cohort designs, Ding et al. (2014) and Yu et al. (2015) considered a general failure-time dependent sampling design where a simple random sample of the cohort is enriched by selectively including certain more informative failure subjects. An overview of failure-time dependent sampling designs can be found in Ding et al. (2017).

We note that the existing cost-effective sampling designs for failure time data were primarily developed for traditional censored data where the failure time is either exactly observed or right-censored. We found only a few papers that discussed biased-sampling designs for interval-censored failure time data. Gilbert et al. (2005) considered a biased-sampling design for a phase 3 HIV-1 preventive vaccine trial where the outcome of interest is the time to HIV infection that suffers from interval-censoring. In particular, they defined the infection time as the midpoint of the dates between the last negative and first positive tests and then employed the case-cohort design for traditional censored data. Li et al. (2008) considered the same HIV vaccine trial as in Gilbert et al. (2005) and they simplified the structure of interval-censored data by assuming that the test dates are fixed and the same for all study subjects and then extended the case-cohort design to fit in the resulting data. Li and Nan (2011) studied the case-cohort design with current status data, a special case of interval-censored data, which arise when each study subject is examined only once for the occurrence of the failure and thus the failure time is either left- or right-censored at the only examination. Recently, Zhou et al. (2017) developed the case-cohort design and an inference procedure for failure time data subject to general interval-censoring. The case-cohort design samples all the failure subjects and applies mainly to rare events. In this paper, we propose an outcome-dependent sampling (ODS) design with general interval-censored failure time data, that applies primarily to non-rare or not-so-rare events where it may not be feasible to sample all the failure subjects. Under the proposed interval-censoring ODS design, we enrich a simple random sample by selectively including certain more informative failure subjects. Specifically, we supplement a simple random sample of the study cohort with certain subjects who are known to experience the failure (i.e. the observed interval containing the failure time has a finite right endpoint) and who are believed to be more informative in terms of the exposure-failure-time relationship (i.e. the observed interval belongs to the two tails of the failure time's distribution). The idea of oversampling from the tails stems from the intuition that if the outcome Y is positively associated with the exposure X, then high (low) Y would be associated with high (low) X; enriching the observed sample with subjects who have high or low Y could potentially enhance the efficiency in evaluating the association between Y and X. This intuition can easily be justified in simple linear regression.

We develop a semiparametric likelihood-based procedure for fitting the proportional hazards model to data from the proposed interval-censoring ODS design. Due to the complicated data structure and the fact that the failure time is never exactly observed, the analysis of interval-censored data is in general much more challenging than that of right-censored data both theoretically and computationally. For example, for regression analysis of interval-censored data, one usually needs to deal with or estimate the finite-dimensional regression parameter and the infinite-dimensional nuisance parameter simultaneously as no tools like the partial likelihood commonly used for right-censored data is available anymore. In particular, regression analysis of interval-censored data, obtained by simple random sampling, under the proportional hazards model remains a popular research topic over the past three decades. Among others, Finkelstein (1986) considered the parametric maximum likelihood estimation with a discrete baseline hazard assumption; Huang (1996) and Zeng et al. (2016) studied the fully semiparametric maximum likelihood estimation for current status data and mixed-case interval-censored data, respectively; Satten (1996) proposed a marginal likelihood approach which avoids estimating the baseline hazard but is still computationally intensive; Pan (2000) suggested a multiple imputation approach which is semiparametric but did not provide theoretical justification; Lin et al. (2015) and Wang et al. (2016) developed efficient algorithms for computing the maximum likelihood estimates via two-stage Poisson data augmentations from Bayesian and Frequentist perspectives, respectively; Zhang et al. (2010) and Zhou et al. (2016) proposed sieve semiparametric maximum likelihood methods and proved the asympotic normality and efficiency of the regression parameter estimators. As discussed by Zhang et al. (2010) and Zhou et al. (2016), the sieve method enjoys both theoretical and computational advantages compared to the alternative methods. Under the proposed interval-censoring ODS design, the likelihood function with the observed data involves two sets of infinite-dimensional nuisance parameters, i.e. the cumulative baseline hazard function and the distributions of examination times and covariates. Following Zhou et al. (2016), we employ a Bernstein-polynomial-based sieve method to deal with the cumulative baseline hazard function. For handling the distributions of examination times and covariates, we adopt the empirical likelihood method considered by Vardi (1985) and Qin (1993) for biasedsampling problems. Both the sieve and empirical likelihood methods yield great dimension reduction on the estimation problem and thus significantly ease the computational difficulty.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed interval-censoring ODS design and describe the likelihood function. In Section 3, we develop a sieve semiparametric maximum empirical likelihood estimation appr oach, where we employ the empirical likelihood and sieve methods to deal with the infinite-dimensional nuisance parameters. We also establish the asymptotic properties of the resulting estimator. In Section 4, we evaluate the performance of the proposed design and estimator through an extensive simulation study. In Section 5, an illustrative example from the ARIC study is provided. Final remarks are given in Section 6.

2. Interval-Censoring ODS design and Likelihood Function

Let T denote the failure time of interest and Z a p-dimensional covariate vector that may affect T. Suppose that the failure time is subject to interval-censoring and the observation can be represented by

$$Y=\{U, V, \Delta_1=I(T \leq U), \Delta_2=I(U<T \leq V)\},$$

where U and V are two random examination times, and (Δ_1, Δ_2, $1 - \Delta_1 - \Delta_2$) indicate left-, interval- and right-censored observations, respectively.

Now we describe our ODS design with interval-censored failure time data. Let τ denote the length of study and a_1 and a_2 two known constants satisfying $0 < a_1 < a_2 < \tau$. The fundamental idea of ODS design is to oversample observations that are believed to be more informative regarding the exposure-response relationship. Following this idea, we oversample subjects who experience the failure (i.e. $\Delta_1 + \Delta_2 = 1$) and who have the failure time falling within either the tail $(0, a_1)$ or (a_2, τ). Specifically, we partition the failures $A = \{Y: \Delta_1 + \Delta_2 = 1\}$ into three mutually exclusive and exhaustive strata: A_k, $k = 1, 2, 3$, defined as

$$\begin{aligned} &A_1=\{Y:\{\Delta_1=1,U<a_1\}\text{ or }\{\Delta_2=1,V<a_1\}\},\\ &A_2=\{Y:\Delta_2=1,U>a_2\},\\ &A_3=A\cap(A_1\cup A_2)^c, \end{aligned} \qquad (1)$$

where ∩ and ∪ denote the intersection and union of sets, respectively, and B^c the complement of a set B. Figure 1 provides an illustration of the partitions A_k, $k = 1, 2, 3$. One can see that 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). Our ODS sample consists of a simple random sample (SRS) of size n_0 and two supplemental samples of sizes n_1 and n_2 from A_1 and A_2, respectively. Let $n = n_0 + n_1 + n_2$ denote the size of the ODS sample, and I_v, I_0 and I_k the index set of the ODS sample, the SRS sample, and the supplemental sample from A_k, respectively. Then the data structure can be represented by

$$\begin{aligned} &\text{The SRS sample: } (Y_i,Z_i),\ \ i\in I_0;\\ &\text{The supplemental samples: } \begin{cases} (Y_i,Z_i|Y_i\in A_1),\ \ i\in I_1,\ \ \text{from the lower tail;}\\ (Y_i,Z_i|Y_i\in A_2),\ \ i\in I_2,\ \ \text{from the upper tail.}\end{cases} \end{aligned} \qquad (2)$$

where $Y_i = \{U_i, V_i, \Delta_{1i} = I(T_i \le U_i), \Delta_{2i} = I(U_i < T_i \le V_i)\}$ for $i \in I_V = I_0 \cup I_1 \cup I_2$. We remark that the proposed design and method will work for the following two scenarios: (i) a two-phase design where the first-phase data on Y are observed for a well-documented parent cohort and the second-phase data on Z are collected for those sampled into I_0, I_1 and I_2 from the cohort; (ii) a design where the information on the parent cohort is unknown, e.g., the subjects in each of I_0, I_1 and I_2 are recruited from a clinic and the recruitment will stop after a target number of subjects is met. In Scenario (ii), the sampling proportions are unknown and one would not have any information on the underlying population other than those in I_0, I_1 and I_2. We assume that independent Bernoulli sampling is used, and the SRS sample is selected first and then the supplemental samples are chosen.

Suppose that the failure time T follows the proportional hazards model with the conditional cumulative hazard function on Z given by

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

where $\Lambda(t)$ is the unspecified cumulative baseline hazard function and β is the p-dimensional regression parameter of primary interest. We assume that T is conditionally independent of the examination times (U, V) given Z and the joint distribution of (U, V, Z) does not involve the parameters (β, Λ). The likelihood function can then be written as

$$
\begin{aligned}
L(\beta,\Lambda,Q)=&\left\{\prod_{i\in I_0} f(Y_i,Z_i)\right\}\cdot\left\{\prod_{k=1}^{2}\prod_{i\in I_k} f(Y_i,Z_i|Y_i\in A_k)\right\}\\
=&\prod_{i\in I_0}\left\{(1-S(U_i|Z))^{\Delta_{1i}}(S(U_i|Z_i)-S(V_i|Z_i))^{\Delta_{2i}}S(V_i|Z_i)^{1-\Delta_{1i}-\Delta_{2i}}q(U_i,V_i,Z_i)\right\}\\
&\cdot\prod_{i\in I_1}\left\{\frac{(1-S(U_i|Z_i))^{\Delta_{1i}}(S(U_i|Z_i)-S(V_i|Z_i))^{\Delta_{2i}}q(U_i,V_i,Z_i)}{P(Y_i\in A_1)}\right\}\\
&\cdot\prod_{i\in I_2}\left\{\frac{(S(U_i|Z_i)-S(V_i|Z_i))q(U_i,V_i,Z_i)}{P(Y_i\in A_2)}\right\},
\end{aligned}
\tag{4}
$$

where $S(t|z) = \exp\{-\Lambda(t)e^{\beta' z}\}$ is the survival function of T given $Z = z$, $Q(\cdot)$ and $q(\cdot)$ denote the joint distribution function and density function of (U, V, Z), respectively, which do not depend on (β, Λ), and $P(Y\in A_k)$ represents the probability that an interval-censored observation $Y = \{U, V, \Delta_1, \Delta_2\}$ belongs to A_k, $k = 1, 2$, given by

$$
\begin{aligned}
&P(Y\in A_1)=P(\Delta_1=1,U<a_1)+P(\Delta_2=1,V<a_1)\\
&=\int\{I(u<a_1)(1-S(u|z))+I(v<a_1)(S(u|z)-S(v|z))\}\,dQ(u,v,z),\\
&P(Y\in A_2)=P(\Delta_2=1,U>a_2)=\int\{I(u>a_2)(S(u|z)-S(v|z))\}\,dQ(u,v,z).
\end{aligned}
$$

The nonparametric components (Λ, Q) cannot be separated from the above likelihood function. Thus, to estimate the regression parameter β, one has to deal with the infinite-dimensional nuisance parameters (Λ, Q). To handle this challenging task, we develop a sieve semiparametric maximum empirical likelihood approach without specifying (Λ, Q).

3. Sieve Semiparametric Maximum Empirical Likelihood Approach

First note that based on the observed data from the interval-censoring ODS design

$$O_i=\{Y_i=\{U_i,V_i,\Delta_{1i}=I(T_i\le U_i),\Delta_{2i}=I(U_i<T_i\le V_i)\},Z_i\}\,,\ i\in I_v,$$

the log-likelihood function can be written as

$$l(\beta,\Lambda,Q)=\log L(\beta,\Lambda,Q)=\sum_{i\in I_v}l_1(O_i;\beta,\Lambda)+\sum_{i\in I_v}\log q(U_i,V_i,Z_i)-\sum_{k=1}^{2}n_k\log\pi_k, \tag{5}$$

where $L(\beta, \Lambda, Q)$ is given by (4),

$$l_1(O_i;\beta,\Lambda)=\Delta_{1i}\log(1-S(U_i|Z_i))+\Delta_{2i}\log(S(U_i|Z_i)-S(V_i|Z_i))+(1-\Delta_{1i}-\Delta_{2i})\log S(V_i|Z_i),$$

and $\pi_k = \int G_k(u, v, z; \beta, \Lambda) dQ(u, v, z)$ with G_k, $k = 1, 2$, defined as

$$G_1(u, v, z; \beta, \Lambda) = I(u < a_1)(1 - S(u|z)) + I(v < a_1)(S(u|z) - S(v|z)),$$
$$G_2(u, v, z; \beta, \Lambda) = I(u > a_2)(S(u|z) - S(v|z)).$$

In the above, $S(t|z) = \exp\{-\Lambda(t)e^{\beta' z}\}$ is the survival function of T given $Z = z$.

Maximizing $l(\beta, \Lambda, Q)$ with respect to β without specifying (Λ, Q) is not straightforward as one has to handle the infinite-dimensional nuisance parameters (Λ, Q). For this, we propose a novel two-step procedure by first employing the empirical likelihood method to deal with Q and then using the sieve method to address Λ. This approach greatly reduces the dimensionality of the estimation problem and relieves the computational burden. In the following, we describe the proposed two-step procedure in details and also establish the asymptotic properties of the resulting estimator.

3.1 Empirical Likelihood Method

We first employ the empirical likelihood method to profile out Q for fixed (β, Λ) in the log-likelihood function (5) (e.g. Vardi, 1985; Qin, 1993; Zhou et al., 2002; Ding et al., 2014). To find a distribution function Q that maximizes $l(\beta, \Lambda, Q)$, it is easy to see that we can restrict our search to the class of discrete distribution functions which have positive jumps only at the observed data points $\{(U_i, V_i, Z_i)\ i \in I_v\}$. Let $p_i = q(U_i, V_i, Z_i) = dQ(U_i, V_i, Z_i)$, $i \in I_v$. Then for fixed (β, Λ), the log-likelihood function (5) can be written as

$$l(\beta, \Lambda, \{p_i\}) = \sum_{i \in I_v} l_1(O_i; \beta, \Lambda) + \sum_{i \in I_v} \log p_i - \sum_{k=1}^{2} n_k \log\left\{\sum_{i \in I_v} G_k(U_i, V_i, Z_i; \beta, \Lambda) p_i\right\}. \quad (6)$$

We want to search for $\{\hat{p}_i\}$ that maximize (6) under the constraints $\{\Sigma_{i \in I\text{<sub>}v\text{</sub>}}\, p_i = 1$, $p_i \geq 0$, $i \in I_v\}$. To solve this constrained optimization problem, we use the Lagrange multiplier method by considering the following Lagrange function

$$H(\beta, \Lambda, \{p_i\}, \rho) = \sum_{i \in I_v} l_1(O_i; \beta, \Lambda) + \sum_{i \in I_v} \log p_i - \sum_{k=1}^{2} n_k \log\left\{\sum_{i \in I_v} G_k(U_i, V_i, Z_i; \beta, \Lambda) p_i\right\} + \rho\left\{1 - \sum_{i \in I_v} p_i\right\},$$

where ρ is the Lagrange multiplier. Taking the derivative of H with respect to p_i, we obtain

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \sum_{k=1}^{2} n_k \frac{G_k(U_i, V_i, Z_i; \beta, \Lambda)}{\sum_{i \in I_v} G_k(U_i, V_i, Z_i; \beta, \Lambda) p_i} - \rho = 0.$$

Solving this equation, we have

$$\hat{\rho}=n_0 \quad \text{and} \quad \hat{p}_i=\left[n_0\left\{1+\sum_{k=1}^{2}\frac{n_k}{n_0\pi_k}G_k(U_i,V_i,Z_i;\beta,\Lambda)\right\}\right]^{-1},\ i\in I_v.$$

Plugging $\{\hat{p}_i\}$ back into $l(\beta, \Lambda, \{p_i\})$ in (6), we have the resulting profile likelihood function

$$l(\theta)=\sum_{i\in I_v}l_1(O_i;\beta,\Lambda)-\sum_{i\in I_v}\log\left[n_0\left\{1+\sum_{k=1}^{2}\frac{n_k}{n_0\pi_k}G_k(U_i,V_i,Z_i;\beta,\Lambda)\right\}\right]-\sum_{k=1}^{2}n_k\log\pi_k, \tag{7}$$

where $\theta = (\xi, \Lambda)$ and $\xi = (\beta', \pi_1, \pi_2)'$. After applying the empirical likelihood method, we have greatly reduced the dimensionality associated with Q and only need to deal with (7).

3.2 Sieve Method

We now consider the estimation of unknown parameters $\theta = (\xi, \Lambda)$ based on the profile likelihood function $l(\theta)$ in (7). Let $\Theta = \{\theta = (\xi, \Lambda) \in \mathscr{B} \otimes \mathscr{M}\}$ denote the parameter space of θ. Here $\mathscr{B} = \{\xi = (\beta', \pi_1, \pi_2)' \in R^{p+2} : \|\beta\| \le M, \pi_k \in [c, d], k = 1, 2\}$ with p being the dimension of β, M a positive constant and $c < d$ two constants in (0, 1), and $\mathscr{M}$ is the collection of all continuous nondecreasing and nonnegative functions over the interval $[\sigma, \tau]$, where σ and τ are known constants usually taken as the lower and upper bounds of all examination times in practice.

Maximizing the profile likelihood function (7) with respect to ξ is still not straightforward as one still has to deal with the infinite-dimensional nuisance parameter Λ. Note that only the values of Λ at the examination times $\{U_i, V_i : i = 1, \ldots, n\}$ matter in (7), one may follow the conventional approach by taking the nonparametric maximum likelihood estimator of Λ as a right-continuous nondecreasing step function with jumps only at the examination times and then maximizing (7) with respect to ξ and the jump sizes (Huang, 1996). However, it is apparent that such fully semiparametric estimation method could involve a large number of parameters if there are no ties among $\{U_i, V_i : i = 1, \ldots, n\}$. To ease the computation difficulty, by following Zhang et al. (2010) and Zhou et al. (2016), we propose to employ the sieve estimation method. Specifically, we define the sieve space as

$$\Theta_n=\{\theta_n=(\xi,\Lambda_n)\in\mathscr{B}\otimes\mathscr{M}_n\},$$

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

$$\mathscr{M}_n=\left\{\Lambda_n(t)=\sum_{k=0}^{m}\phi_k B_k(t,m,\sigma,\tau):0\le\phi_0\le\phi_1\le\ldots\le\phi_m \text{ and } \sum_{k=0}^{m}|\phi_k|\le M_n\right\}$$

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

$$B_k(t,m,\sigma,\tau)=\begin{pmatrix} m \\ k \end{pmatrix}\left(\frac{t-\sigma}{\tau-\sigma}\right)^{k}\left(1-\frac{t-\sigma}{\tau-\sigma}\right)^{m-k},\ k=0,\ldots,m,$$

and $M_n = O(n^a)$ for some $a > 0$ controlling the size of the sieve space. The constraints on the Bernstein coefficients ϕ_k's in $\mathfrak{M}_n$ are imposed to guarantee that the estimate of the cumulative baseline hazard function $\Lambda(t)$ is nonnegative and nondecreasing. In fact, one can show that any $\Lambda(t)$ can be approximated by a Bernstein polynomial $\Lambda_n(t)$ with the coefficients $\phi_k = \Lambda(\sigma + (k/m)(\tau - \sigma))$ arbitrarily well as $n \to \infty$, i.e., the sieve space Θ_n approximates the parameter space Θ arbitrarily well as $n \to \infty$ (Lorentz, 1986; Shen, 1997; Wang and Ghosh, 2012). We define the sieve semiparametric maximum empirical likelihood estimator $\hat{\theta}_n=(\hat{\xi}_n,\hat{\Lambda}_n)=(\hat{\beta}_n,\hat{\pi}_{1n},\hat{\pi}_{2n},\sum_{k=0}^{m}\hat{\phi}_{kn}B_k(\cdot,m,\sigma,\tau))$ of θ to be the value of θ that maximizes the sieve log-likelihood function $l_n(\theta)$ over Θ_n, where

$$\begin{aligned} l_n(\theta)=\sum_{i=1}^{n}[&\Delta_{1i}\log(1\\ &-S_n(U_i|Z_i))\\ &+\Delta_{2i}\log(S_n(U_i|Z_i)\\ &-S_n(V_i|Z_i))\\ &+(1-\Delta_{1i}-\Delta_{2i})\log S_n(V_i|Z_i)\\ &-\log\left\{n_0\left(1+\sum_{k=1}^{2}\frac{n_k}{n_0\pi_k}G_k(U_i,V_i,Z_i;\beta,\Lambda_n)\right)\right\}\\ &-\sum_{k=1}^{2}\frac{n_k}{n}\log\pi_k\Bigg]. \end{aligned} \qquad (8)$$

In the above, $S_n(t|z) = \exp\{-\Lambda_n(t)e^{\beta' z}\}$ is the survival function of T given $Z = z$ and

$$\begin{aligned} G_1(u,v,z;\beta,\Lambda_n)&=I(u<a_1)(1-S_n(u|z))+I(v<a_1)(S_n(u|z)-S_n(v|z)),\\ G_2(u,v,z;\beta,\Lambda_n)&=I(u>a_2)(S_n(u|z)-S_n(v|z)). \end{aligned}$$

Compared to the fully semiparametric estimation method, the sieve method significantly reduces the dimensionality of the optimization problem and relieves the computation burden as the number of Bernstein bases needed to reasonably approximate the unknown function Λ grows much slower as the sample size increases. Bernstein polynomial basis has several advantages compared to other bases such as piecewise linear function and spline. First, it can model the monotonicity and nonnegativity of the cumulative baseline hazard function with simple restrictions that can easily be removed through reparameterization. Second, Bernstein polynomial is easier to work with as it does not require the specification of interior knots.

3.3 Asymptotic Properties

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

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

for any $\theta^1 = (\xi^1, \Lambda^1) \in \Theta$ and $\theta^2 = (\xi^2, \Lambda^2) \in \Theta$, where $\|v\|$ denote the Euclidean norm for a vector v and $\|\Lambda^1-\Lambda^2\|_2^2=\int\{(\Lambda^1(u)-\Lambda^2(u))^2+(\Lambda^1(v)-\Lambda^2(v))^2\}dG(u,v)$. Let $\theta_0 = (\xi_0, \Lambda_0) = (\beta_0, \pi_{10}, \pi_{20}, \Lambda_0)$ denote the true value of θ and assume that $n_0/n \rightarrow \rho_0 > 0$ and $n_k/n \rightarrow \rho_k \geq 0$, $k = 1, 2$, 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 Web Appendix.

Theorem 1—*Assume that Conditions (C1) – (C4) given in the* Web Appendix *hold. Then we have that* $d(\hat{\theta}_n, \theta_0) \rightarrow 0$ *almost surely and* $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{(1-v)/2,\ vr/2\}})$*, where* $v \in (0, 1)$ *such that* $m = o(n^v)$ *and r is defined in Condition (C3).*

Theorem 2—*Assume that Conditions (C1) – (C4) given in the* Web Appendix *hold. If* $v > 1/2r$*, we have* $\sqrt{n}(\hat{\xi}_n-\xi_0)\rightarrow_d N(0,\sum)$*, where* $\Sigma = \Gamma^{-1}\Psi\Gamma^{-1}$ *with* $\Gamma=\sum_{k=0}^{2}\rho_k J_k(\xi_0)$ *and* $\Psi=\sum_{k=0}^{2}\rho_k var_k(h_k(\xi_0,\Lambda_0;O))$*.*

The proofs of Theorems 1 and 2 will be sketched in the Web Appendix. $J_k(\xi)$ and $h_k(\xi, \Lambda; O)$ in Theorem 2 are the information and efficient score of ξ corresponding to the k-th stratum, $k = 0, 1,2$ ($k = 0$ corresponds to the whole population), and they will be further discussed in the Web Appendix. Following Huang et al. (2012), we can obtain a consistent variance estimator of $\hat{\xi}_n$ by treating the log-likelihood function $l_n(\theta)$ in (8) as if it is a function of the $(p + m + 3)$-dimensional parameter $\theta = (\xi_{(p+2)\times1}, \phi_{(m+1)\times1})$ and then replacing the large-sample quantities in Σ given above with the corresponding small-sample quantities.

We now make a few remarks on the implementation of the proposed estimation procedure. First, it should be noted that there are some restrictions on the parameters due to boundedness and monotonicity. On the other hand, they can easily be removed through reparameterization. For example, we could reparameterize the parameters π_k as $1/(1+\exp(-\pi_k^*))$, $k = 1, 2$, and $\{\phi_0, \ldots, \phi_m\}$ as the cumulative sums of $\{\exp(\phi_0^*),\ldots,\exp(\phi_m^*)\}$. This reparameterization is a simple one-to-one transformation and does not complicate the computation. Regarding the restriction $\sum_{k=0}^{m}|\phi_k|\leq M_n$, since $M_n = O(n^a)$ is imposed mainly for technical purposes and can be chosen reasonably large for fixed sample size in practice, we do not need to consider this restriction in computation. Thus, to obtain the proposed estimator $\hat{\theta}_n$, many existing unconstrained optimization methods can be used and for the numerical studies in Sections 4 and 5, we employ the

Nelder-Mead simplex algorithm built in *fminsearch* in Matlab. Also for the implementation of the proposed estimation procedure, one needs to determine the degree of Bernstein polynomials m which controls the smoothness of the sieve approximation. For this, we suggest to consider several different values of m and choose the one that minimizes

$$AIC=-2l_n(\hat{\theta}_n)+2(p+m+3). \quad (9)$$

4. A Simulation Study

We carry out an extensive simulation study to evaluate the finite-sample performance of the proposed interval-censoring ODS design and estimator, including the comparison of our ODS design with the SRS and generalized case-cohort designs and the comparison of our estimator with other naive or adapted estimators. Specifically, we generated the covariate $Z \sim N(0, 1)$ and the failure time T from the proportional hazards model given Z:

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

with the cumulative baseline hazard $\Lambda(t) = 0.1t$ and regression parameter $\beta = 0$ or log2.

To generate the interval-censored observation $Y = \{U, V, \Delta_1 = I(T \le U), \Delta_2 = I(U < T \le V)\}$, we mimicked medical or epidemiologic follow-up studies. Suppose that a subject was scheduled to be examined at a sequence of time points in $[0, \tau]$ generated as cumulative sums of uniform random variables on $[0, \delta]$ until τ, where τ is the length of study and $0 < \delta < \tau$. At each of these time points, it was assumed that the subject could miss the scheduled examination with probability ζ, independent of the examination results at other time points. If T was smaller than the first examination time (i.e. left-censored), we defined U as the first examination time, V the second examination time and $(\Delta_1, \Delta_2) = (1, 0)$; if T was larger than the last examination time (i.e. right-censored), we defined U as the second to the last examination time, V the last examination time and $(\Delta_1, \Delta_2) = (0, 0)$; otherwise, U and V were defined as the two consecutive examination times bracketing T and $(\Delta_1, \Delta_2) = (0, 1)$. In the simulation study, we adjusted the values of τ, δ and ζ according to the desired proportion of failures (0.1, 0.2 or 0.3). Here the proportion of failures, denoted by Pr(failure), refers to the proportion of subjects who experience the failure.

The ODS sample consists of a SRS sample of size n_0 and two supplemental samples of sizes n_1 and n_2 from the lower tail A_1 and upper tail A_2, respectively. For the cutpoints (a_1, a_2) that define A_1 and A_2, we considered the (10, 90)- or (20, 80)-th percentiles. In Table 1, five estimators of β were compared under $(n_0, n_1, n_2) = (470, 40, 40)$: (i) the sieve maximum likelihood estimator based only on the SRS portion of the ODS sample, denoted by $\hat{\beta}_{SRS}$<sub>n_0</sub>; (ii) the sieve maximum likelihood estimator based on a SRS sample of the same size as the ODS sample, denoted by $\hat{\beta}_{SRS}$<sub>n</sub>; (iii) the sieve weighted estimator based on the generalized case-cohort sample that consists of a subcohort of size n_0 and a SRS of size $n_1 + n_2$ selected from the remaining cases (i.e. failure subjects), denoted by $\hat{\beta}_{GCC}$; (iv) the inverse probability weighted estimator (Breslow and Wellner, 2007) based on the ODS sample, denoted by $\hat{\beta}_{IPW}$; (v) the proposed estimator, denoted by $\hat{\beta}_P$. The degree of Bernstein polynomial used in the sieve estimation was taken as $m = 3$. The simulation results include "Bias" calculated as the average of point estimates minus the true value, "SSD" the sample standard deviation, "ESE" the average of estimated standard errors, "CP" the empirical coverage proportion of 95% confidence interval and "RE" the sample relative efficiency with respect to $\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$ calculated as $[\mathrm{SSD}(\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}})/\mathrm{SSD}(\hat{\beta})]^2$. The results were based on 1000 replicates.

From Table 1, one can see that for all situations considered: (i) the proposed estimator under the interval-censoring ODS design is virtually unbiased; (ii) the standard error estimates are close to the empirical standard deviations; (iii) the empirical coverage proportions are close to 95%, which indicates that the normal approximation to the distribution of the proposed estimator is reasonable; (iv) the proposed ODS design ($\hat{\beta}_P$) is more efficient than the alternative SRS designs ($\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$ and $\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$); for example, when the failure rate is 0.1, the cutpoints are (10, 90)-th percentiles and $\beta = \log 2$, it achieves 132% efficiency gain compared to $\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$; (v) the proposed estimator is more efficient than the estimator based on the generalized case-cohort sample; for example, when the failure rate is 0.1, the cutpoints are (10, 90)-th percentiles and $\beta = \log 2$, the relative efficiency of $\hat{\beta}_P$ compared to $\hat{\beta}_{GCC}$ is $(0.137/0.103)^2 = 1.77$; (vi) the proposed estimator $\hat{\beta}_P$ is more efficient than the inverse probability weighted estimator $\hat{\beta}_{IPW}$ that is routinely used to accommodate sampling bias; for example, when the failure rate is 0.1, the cutpoints are (10, 90)-th percentiles and $\beta = \log 2$, the relative efficiency of $\hat{\beta}_P$ compared to $\hat{\beta}_{IPW}$ is $(0.148/0.103)^2 = 2.06$. In practice, sampling without replacement is often used to select random samples. Therefore, we conducted some additional simulations to examine the performance of our proposed estimator in the situation when sampling without replacement is used. We considered the same setup and parameter values as those for Table 1 and the results are presented in Web Table 1. The results show that the proposed estimator under the sampling without replacement situation performs similarly to that under independent Bernoulli sampling.

To evaluate the performances of the proposed ODS design and estimator under different sizes of n_0, n_1 and n_2, we conducted additional simulations with $(n_0, n_1, n_2) = (530, 10, 10)$, (500, 25, 25) and (1000, 50, 50) and presented the results in Table 2. The simulation setups in Table 2 are the same as those in Table 1 except for the sizes of n_0, n_1 and n_2. Also the cutpoints, (10, 90)-th percentiles, are used in Table 2. One can see from Table 2 that (i) for a fixed overall ODS sample size $n = n_0 + n_1 + n_2$, as we allocate more samples to the tails, the efficiency of the proposed estimator $\hat{\beta}_P$ improves; for example, when the failure rate is 0.1, the cutpoints are (10, 90)-th percentiles and $\beta = \log 2$, as we change (n_0, n_1, n_2) from (530, 10, 10) to (500, 25, 25) or to (470, 40, 40), the efficiency improves by $(0.124/0.116)^2 = 1.14$ or $(0.124/0.107)^2 = 1.34$; (ii) as we increase the overall ODS sample size, the efficiency of $\hat{\beta}_P$ improves as expected; for example, when the failure rate is 0.1, the cutpoints are (10, 90)-th percentiles and $\beta = \log 2$, as we increase (n_0, n_1, n_2) from (500, 25, 25) to (1000, 50, 50), the efficiency improves by $(0.116/0.081)^2 = 2.05$.

5. Analysis for Diabetes from the ARIC Study

In this section, we illustrate the proposed interval-censoring ODS design and inference procedure by analyzing a dataset on incident diabetes from the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators, 1989). The ARIC study is a longitudinal epidemiologic observational study conducted in four US field centers (Forsyth County, NC (Center-F), Jackson, MS (Center-J), Minneapolis Suburbs, MN (Center-M) and Washington County, MD (Center-W)). The study began in 1987 and each field center recruited a cohort sample of approximately 4000 men and women aged 45–64 from their community. Forsyth County, Minneapolis Suburbs, and Washington County include white participants, and Forsyth County and Jackson Center include African American participants. Each participant received an extensive examination at recruitment, including medical, social, and demographic data, and was scheduled to be re-examined on average of every three years with the first examination (baseline) occurring in 1987–89, the second in 1990–92, the third in 1993–95 and the fourth in 1996–98. Since the incidence of diabetes can be determined only between two consecutive examinations, the observed data were subject to interval-censoring.

We illustrate the proposed interval-censoring ODS design and inference procedure by assessing the effect of high-density lipoprotein (HDL) cholesterol level on the risk of diabetes after adjusting for confounding variables and other risk factors in white men younger than 55 years. In particular, we constructed the ODS sample as follows. The cohort of interest consists of 2110 white men younger than 55 years and 244 were observed to have developed diabetes during the study. We took a simple random sample of size $n_0 = 520$ from the cohort and selected two supplemental samples of sizes $n_1 = n_2 = 15$ from the strata A_1 and A_2 defined in (1), where $a_1 = 1092$ (days) and $a_2 = 2127$ (days) are approximate (25, 75)-th percentiles of the cohort, respectively. Thus, the ODS sample had total $n = 550$ subjects. We considered the following proportional hazards model

$$\Lambda(t|Z)=\Lambda(t)\exp(Z^{'}\beta),$$

where the vector of covariates Z included HDL cholesterol level, total cholesterol level, body mass index (BMI), age, smoking status, and indicators for field centers (Center-M was chosen as reference). We compared three estimators: (i) the proposed estimator; (ii) the inverse probability weighted (IPW) estimator; (iii) the sieve maximum likelihood estimator based on only the SRS portion of the ODS sample. Regarding the choice of the degree of Bernstein polynomial used in the sieve estimation, we considered the integers $m = 3$ to 8 and the AIC criterion (9) suggested to choose $m = 3$ for all three estimators. From the results in Table 3, one can see that (i) the proposed and IPW methods indicate that higher HDL cholesterol level is significantly associated with lower risk of diabetes in white men younger than 55 years; (ii) the proposed method yielded smaller standard error and more significant result compared to the other methods; in particular, the regression coefficient estimate for HDL cholesterol level based on the proposed method is −0.0272, the standard error estimate is 0.0126, and the P-value is 0.0311; (iii) all three methods suggest that lower BMI level is significantly associated with lower risk of diabetes in white men younger than 55 years.

6. Discussion

We proposed an innovative and cost-effective sampling design with interval-censored failure time outcome, i.e., the interval-censoring ODS design, which enables investigators to make more efficient use of their study budget by selectively collecting more informative failure subjects. For analyzing data from the proposed interval-censoring ODS design, we developed an efficient and robust sieve semiparametric maximum empirical likelihood method. As shown in the simulation study, the proposed design and method is more efficient than the SRS and generalized case-cohort designs as well as the IPW method.

We provide some remarks on the practical use of the proposed ODS design. To implement this design, one needs to determine the cutpoints (a_1, a_2) and the allocations of the SRS and supplemental samples. For the cutpoints, we suggest to take the lower and upper k-th percentiles of all examination times and recommend k between 10 and 35. For the allocations of the SRS and supplemental samples, it seems from the simulation results that the more supplemental subjects are sampled from the tails, the more efficient the proposed method could be. However, one also needs to keep the SRS sample large enough so as to maintain the adequate representativeness of the ODS sample on the whole population. Due to these considerations and our simulation experiences, we recommend the ratio of SRS size and total supplemental sample size to be at least 5 : 1. Another aspect of the ODS design is the selection of the SRS and supplemental samples. The asymptotic properties of our proposed estimator are derived based on independent Bernoulli sampling. In practice, sampling without replacement is often used. The asymptotic properties can be derived under sampling without replacement, but the derivation would be much more tedious. Based on the literature for case-cohort designs (e.g. Breslow and Wellner, 2007), we expect the difference in asymptotic variance to be very small if there is any. We conducted additional simulations to examine the effect of using sampling without replacement on our proposed estimator. The results show that our proposed estimator performs well even though sampling without replacement is used for selecting the ODS samples in the practical situations we considered.

We also make some comments on the proposed inference procedure. First, regarding the number of Bernstein basis polynomials m, in theory, it depends on the sample size n with $m = o(n^{\nu})$. However, we found that m does not need to be very large for the results to be satisfying. In practice, based on our numerical experiences, we recommend to consider the values of m to be from 3 to 8 and choose the one that minimizes the AIC. On the other hand, we focused on the proportional hazards model in this paper for its good interpretation and wide application. In fact, the proposed method could easily be extended to other semiparametric models, such as the proportional odds model and transformation model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Editor, the Associate Editor and the referee for their valuable comments which have led to great improvement of the paper. This work was partially supported by grants from the National Institutes of Health (R01 ES 021900 and P01 CA 142538). The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts. The authors thank the staff and participants of the ARIC study for their important contributions.

References

Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scandinavian Journal of Statistics. 2007; 34:86–102.

Cai J, Zeng D. Power calculation for case–cohort studies with nonrare events. Biometrics. 2007; 63:1288–1295. [PubMed: 17608788]

Chatterjee N, Chen YH, Breslow NE. A pseudoscore estimator for regression problems with two-phase sampling. Journal of the American Statistical Association. 2003; 98:158–168.

Chen, D-G., Sun, J., Peace, KE. Interval-Censored Time-to-Event Data: Methods and Applications. CRC Press; 2012.

Chen K, Lo SH. Case-cohort and case-control analysis with Cox's model. Biometrika. 1999; 86:755–764.

Cornfield J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast, and cervix. Journal of the National Cancer Institute. 1951; 11:1269–1275. [PubMed: 14861651]

Ding J, Lu TS, Cai J, Zhou H. Recent progresses in outcome-dependent sampling with failure time data. Lifetime Data Analysis. 2017; 23:57–82. [PubMed: 26759313]

Ding J, Zhou H, Liu Y, Cai J, Longnecker MP. Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. Biostatistics. 2014; 15:636–650. [PubMed: 24812419]

Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics. 1986; 42:845–854. [PubMed: 3814726]

Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, 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. Journal of Infectious Diseases. 2005; 191:666–677. [PubMed: 15688279]

Huang J. Efficient estimation for the proportional hazards model with interval censoring. Annals of Statistics. 1996; 24:540–568.

Huang, J., Zhang, Y., Hua, L. Consistent variance estimation in semiparametric models with application to interval-censored data. In: Chen, DG.Sun, J., Peace, KE., editors. Interval-Censored Time-to-Event Data: Methods and Applications. 2012. p. 233-268.

Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika. 2009; 96:887–901. [PubMed: 23946547]

Kong L, Cai J. Case-cohort analysis with accelerated failure time model. Biometrics. 2009; 65:135–142. [PubMed: 18537948]

Li Z, Gilbert P, Nan B. Weighted likelihood method for grouped survival data in case–cohort studies with application to HIV vaccine trials. Biometrics. 2008; 64:1247–1255. [PubMed: 19032178]

Li Z, Nan B. Relative risk regression for current status data in case-cohort studies. Canadian Journal of Statistics. 2011; 39:557–577.

Lin X, Cai B, Wang L, Zhang Z. A Bayesian proportional hazards model for general interval-censored data. Lifetime Data Analysis. 2015; 21:470–490. [PubMed: 25098226]

Lorentz, GG. Bernstein Polynomials. New York: Chelsea Publishing Co; 1986.

Lu W, Tsiatis AA. Semiparametric transformation models for the case-cohort study. Biometrika. 2006; 93:207–214.

Pan W. A multiple imputation approach to Cox regression with interval-censored data. Biometrics. 2000; 56:199–203. [PubMed: 10783796]

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Qin J. Empirical likelihood in biased sample problems. Annals of Statistics. 1993; 21:1182–1196.

Satten GA. Rank-based inference in the proportional hazards model for intervalcensored data. Biometrika. 1996; 83:355–370.

Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Annals of Statistics. 1988; 16:64–81.

Shen X. On methods of sieves and penalization. Annals of Statistics. 1997; 25:2555–2591.

Song R, Zhou H, Kosorok MR. A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. Biometrika. 2009; 96:221–228. [PubMed: 20107493]

Sun, J. The Statistical Analysis of Interval-Censored Failure Time Data. New York: Springer; 2006.

The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. American Journal of Epidemiology. 1989; 129:687–702. [PubMed: 2646917]

Vardi Y. Empirical distributions in selection bias models. Annals of Statistics. 1985; 13:178–203.

Wang J, Ghosh SK. Shape restricted nonparametric regression with bernstein polynomials. Computational Statistics and Data Analysis. 2012; 56:2729–2741.

Wang L, McMahan CS, Hudgens MG, Qureshi ZP. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. Biometrics. 2016; 72:222–231. [PubMed: 26393917]

Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. Journal of the American Statistical Association. 2005; 100:459–469.

Whittemore AS. Multistage sampling designs and estimating equations. Journal of the Royal Statistical Society, Series B. 1997; 59:589–602.

Yu J, Liu Y, Sandler DP, Zhou H. Statistical inference for the additive hazards model under outcome-dependent sampling. Canadian Journal of Statistics. 2015; 43:436–453. [PubMed: 26379363]

Zeng D, Lin DY. Efficient estimation of semiparametric transformation models for two-phase cohort studies. Journal of the American Statistical Association. 2014; 109:371–383. [PubMed: 24659837]

Zeng D, Mao L, Lin D. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. Biometrika. 2016; 103:253–271. [PubMed: 27279656]

Zhang Y, Hua L, Huang J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. Scandinavian Journal of Statistics. 2010; 37:338–354.

Zhou H, Chen J, Rissanen TH, Korrick SA, Hu H, Salonen JT, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology. 2007; 18:461–468. [PubMed: 17568219]

Zhou H, Song R, Wu Y, Qin J. Statistical inference for a two-stage outcome-dependent sampling design with a continuous outcome. Biometrics. 2011; 67:194–202. [PubMed: 20560938]

Zhou H, Weaver M, Qin J, Longnecker M, Wang M. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2002; 58:413–421. [PubMed: 12071415]

Zhou Q, Hu T, Sun J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. Journal of the American Statistical Association. 2016; doi: 10.1080/01621459.2016.1158113

Zhou Q, Zhou H, Cai J. Case-cohort studies with interval-censored failure time data. Biometrika. 2017; 104:17–29. [PubMed: 28943643]

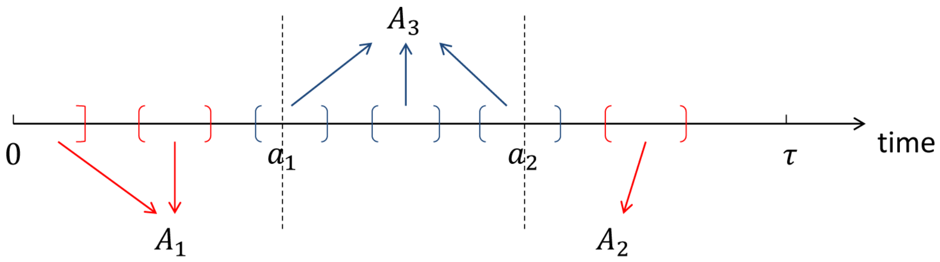

Figure 1.

An illustration of the partitions $\{A_k, k = 1, 2, 3\}$ of failures under the interval-censoring ODS design

Table 1

Simulation results for the estimation of β when $(n_0, n_1, n_2) = (470, 40, 40)$

			$\beta = 0$					$\beta = \log 2$				
Pr(failure)	**cutpoints**		**Bias**	**SSD**	**ESE**	**CP**	**RE**	**Bias**	**SSD**	**ESE**	**CP**	**RE**
0.1	(20%, 80%)	$\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$	−0.000	0.150	0.145	0.94	0.84	0.000	0.163	0.149	0.93	0.89
		$\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$	−0.003	0.137	0.134	0.95	1.00	−0.004	0.154	0.138	0.93	1.00
		$\hat{\beta}_{GCC}$	−0.001	0.111	0.111	0.95	1.53	−0.002	0.132	0.131	0.95	1.36
		$\hat{\beta}_{IPW}$	0.000	0.127	0.121	0.94	1.18	−0.005	0.152	0.139	0.93	1.03
		$\hat{\beta}_{P}$	0.004	0.100	0.098	0.94	1.89	−0.008	0.106	0.106	0.94	2.11
	(10%, 90%)	$\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$	−0.001	0.148	0.145	0.95	0.86	0.001	0.160	0.149	0.94	0.97
		$\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$	0.004	0.138	0.133	0.94	1.00	−0.002	0.158	0.137	0.93	1.00
		$\hat{\beta}_{GCC}$	−0.001	0.115	0.112	0.95	1.42	−0.003	0.137	0.132	0.94	1.32
		$\hat{\beta}_{IPW}$	−0.003	0.136	0.130	0.94	1.02	0.002	0.148	0.148	0.94	1.13
		$\hat{\beta}_{P}$	0.001	0.105	0.100	0.94	1.71	−0.004	0.103	0.105	0.93	2.32
0.2	(20%, 80%)	$\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$	−0.000	0.106	0.103	0.94	0.81	0.006	0.110	0.107	0.94	0.81
		$\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$	−0.002	0.096	0.095	0.95	1.00	0.002	0.099	0.099	0.95	1.00
		$\hat{\beta}_{GCC}$	−0.008	0.108	0.107	0.95	0.79	0.010	0.113	0.113	0.94	0.77
		$\hat{\beta}_{IPW}$	0.001	0.095	0.092	0.94	1.00	0.004	0.098	0.097	0.94	1.01
		$\hat{\beta}_{P}$	0.001	0.084	0.083	0.95	1.28	−0.010	0.086	0.086	0.94	1.33
	(10%, 90%)	$\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$	0.002	0.103	0.103	0.95	0.92	0.003	0.106	0.107	0.95	0.92
		$\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$	0.003	0.098	0.095	0.94	1.00	0.002	0.102	0.099	0.95	1.00
		$\hat{\beta}_{GCC}$	0.000	0.111	0.106	0.94	0.79	0.008	0.114	0.113	0.94	0.79
		$\hat{\beta}_{IPW}$	0.001	0.098	0.097	0.94	1.00	0.004	0.100	0.101	0.95	1.04
		$\hat{\beta}_{P}$	0.003	0.084	0.084	0.94	1.37	−0.004	0.083	0.086	0.95	1.50
0.3	(20%, 80%)	$\hat{\beta}_{SRS\text{<sub>}n_0\text{</sub>}}$	0.001	0.087	0.084	0.95	0.85	0.007	0.095	0.091	0.93	0.87
		$\hat{\beta}_{SRS\text{<sub>}n\text{</sub>}}$	0.001	0.080	0.078	0.94	1.00	0.001	0.088	0.084	0.95	1.00

			$\beta = 0$					$\beta = \log 2$				
Pr(failure)	**cutpoints**		**Bias**	**SSD**	**ESE**	**CP**	**RE**	**Bias**	**SSD**	**ESE**	**CP**	**RE**
		$\hat{\beta}_{GCC}$	−0.003	0.099	0.101	0.94	0.65	0.005	0.112	0.109	0.94	0.62
		$\hat{\beta}_{IPW}$	0.001	0.081	0.078	0.94	0.97	0.008	0.086	0.084	0.94	1.06
		$\hat{\beta}_{P}$	0.002	0.076	0.072	0.93	1.10	−0.016	0.082	0.076	0.93	1.15
	(10%, 90%)	$\hat{\beta}_{SRS}$<sub>n_0</sub>	−0.003	0.087	0.084	0.94	0.78	0.002	0.089	0.091	0.96	0.95
		$\hat{\beta}_{SRS}$<sub>n</sub>	0.003	0.077	0.078	0.96	1.00	0.004	0.086	0.083	0.94	1.00
		$\hat{\beta}_{GCC}$	−0.004	0.100	0.101	0.96	0.59	0.006	0.105	0.108	0.95	0.67
		$\hat{\beta}_{IPW}$	−0.004	0.082	0.081	0.95	0.89	0.004	0.085	0.087	0.95	1.03
		$\hat{\beta}_{P}$	−0.004	0.077	0.076	0.94	1.00	−0.008	0.077	0.078	0.95	1.24

$\hat{\beta}_{SRS}$<sub>n_0</sub>, the sieve MLE based only on the SRS portion of the ODS sample; $\hat{\beta}_{SRS}$<sub>n</sub>, the sieve MLE based on a SRS sample of the same size as the ODS sample; $\hat{\beta}_{GCC}$, the estimator based on the generalized case-cohort sample; $\hat{\beta}_{IPW}$, the inverse probability weighted estimator based on the ODS sample; $\hat{\beta}_{P}$, the proposed estimator based on the ODS sample.

Table 2

Simulation results for the estimation of β using the proposed method under different sample sizes

		$\beta = 0$				$\beta = \log 2$			
(n_0, n_1, n_2)	**Pr(failure)**	**Bias**	**SSD**	**ESE**	**CP**	**Bias**	**SSD**	**ESE**	**CP**
(530, 10, 10)	0.1	−0.002	0.124	0.121	0.94	−0.001	0.124	0.124	0.95
	0.2	0.000	0.087	0.090	0.95	−0.012	0.092	0.096	0.94
	0.3	−0.007	0.076	0.075	0.94	−0.008	0.082	0.084	0.94
(500, 25, 25)	0.1	−0.003	0.104	0.104	0.95	−0.006	0.116	0.113	0.95
	0.2	0.001	0.085	0.086	0.95	−0.002	0.089	0.091	0.94
	0.3	−0.005	0.076	0.075	0.94	−0.010	0.084	0.078	0.93
(470, 40, 40)	0.1	0.003	0.097	0.098	0.95	−0.001	0.107	0.102	0.94
	0.2	0.005	0.085	0.084	0.95	−0.006	0.088	0.088	0.95
	0.3	−0.005	0.078	0.075	0.93	−0.004	0.079	0.078	0.95
(1000, 50, 50)	0.1	0.001	0.072	0.072	0.96	−0.006	0.081	0.078	0.94
	0.2	−0.001	0.063	0.061	0.94	−0.005	0.066	0.063	0.94
	0.3	0.001	0.054	0.053	0.94	−0.009	0.059	0.058	0.92

Table 3

Analysis results for diabetes data from the ARIC study

	Proposed method			IPW method			SRS portion only		
Variable	$\hat{\beta}$	SE	P-value	$\hat{\beta}$	SE	P-value	$\hat{\beta}$	SE	P-value
HDL Cholesterol	−0.0272	0.0126	0.0311	−0.0271	0.0133	0.0426	−0.0228	0.0150	0.1273
Total Cholesterol	0.0003	0.0031	0.9364	0.0016	0.0032	0.6137	0.0029	0.0037	0.4360
BMI	0.1145	0.0273	0.0000	0.1094	0.0372	0.0032	0.1244	0.0328	0.0001
Age	−0.0544	0.0217	0.0120	−0.0324	0.0407	0.4257	−0.0089	0.0546	0.8703
Current smoking	−0.0649	0.2563	0.8001	−0.0862	0.1990	0.6648	−0.2679	0.3478	0.4411
Center-F	−0.0429	0.3150	0.8917	−0.1214	0.1897	0.5224	−0.1619	0.3770	0.6676
Center-W	0.0245	0.2730	0.9285	−0.2077	0.2166	0.3376	−0.1475	0.3240	0.6488