View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Carolina Digital Repository

HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Biometrika. 2017 March ; 104(1): 17–29. doi:10.1093/biomet/asw067.

Case-cohort studies with interval-censored failure time data

Q. ZHOU, **H. ZHOU**, and **J. CAI**

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Summary

The case-cohort design has been widely used as a means of cost reduction in assembling or measuring expensive covariates in large cohort studies. The existing literature on the case-cohort design is mainly focused on right-censored data. In practice, however, the failure time is often subject to interval-censoring; it is known only to fall within some random time interval. In this paper, we consider the case-cohort study design for interval-censored failure time and develop a sieve semiparametric likelihood approach for analyzing data from this design under the proportional hazards model. We construct the likelihood function using inverse probability weighting and build the sieves with Bernstein polynomials. The consistency and asymptotic normality of the resulting regression parameter estimator are established and a weighted bootstrap procedure is considered for variance estimation. Simulations show that the proposed method works well for practical situations, and an application to real data is provided.

Keywords

Case-cohort design; Interval-censoring; Missing covariates; Proportional hazards model; Sieve method; Weighted likelihood

1. Introduction

In epidemiologic cohort studies, the outcomes of interest are often times to failure events, such as cancer, heart disease and HIV infection, which are relatively rare even after a long period of follow-up; the study cohorts are usually chosen very large so as to yield reliable information about the effect of exposure variables on these rare failure times. In many cases, the exposure variables of interest are difficult or expensive to collect or measure. With limited funds, it could be prohibitive to obtain these variables for all subjects in a large cohort. Prentice (1986) proposed the case-cohort design where the expensive exposure variables are obtained only for a random sample, named the subcohort, from the study cohort, as well as for subjects who have experienced the failure event during the follow-up period. Extensive research has been done on this design. Under the proportional hazards model, Prentice (1986) and Self & Prentice (1988) proposed pseudolikelihood approaches; Chen & Lo (1999) and Chen (2001) developed estimating equation methods; Marti &

Supplementary material

Supplementary material available at *Biometrika* online includes the two lemmas used in the proof of Theorem 1 and their proofs, and the Matlab code for the proposed inference procedure.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Chavance (2011) and Keogh & White (2013) proposed multiple imputation approaches; Scheike & Martinussen (2004) and Zeng & Lin (2014) considered maximum likelihood estimation; and Kang & Cai (2009) and Kim et al. (2013) developed weighted estimating equation approaches for case-cohort studies with multiple outcomes. Other related cost-effective sampling schemes include outcome-dependent sampling designs (Zhou et al., 2002; Ding et al., 2014). All of these designs and methods are primarily focused on right-censored data where the failure time of interest is either exactly observed or is right-censored. In practice, however, the occurrences of some failure events, such as HIV infection and diabetes, are not accompanied by any symptoms and their determinations rely on laboratory tests or physician diagnosis; the exact times to these failure events are not available.

In this paper, we consider the case-cohort study design for interval-censored failure time data, which arise when the failure time of interest is observed or known only to belong to a random time interval (Sun, 2006). Areas that often produce such data include epidemiologic studies, biomedical follow-up studies, demographic studies and social sciences, where the study subjects are only examined for the occurrence of the failure event at discrete visits instead of being continuously monitored. One example is the Atherosclerosis Risk in Communities study, a longitudinal epidemiologic cohort study, where the participants were scheduled to be examined for health status every three years on average. In this study, the occurrence of a disease such as diabetes was known only between two consecutive examinations, so only interval-censored data on time to the disease were available. Interval-censoring is a general type of censoring that includes left- and right-censoring as special cases. If a participant had developed the disease at the first follow-up examination U, we would have a left-censored observation denoted by $(0, U]$; if a participant had not yet developed the disease at the last follow-up examination V, we would obtain a right-censored observation denoted by $(V, +\infty)$; otherwise, the observation would be a finite time interval with both endpoints in $(0, +\infty)$. Here we consider the interval-censored case-cohort design in which the expensive exposure variables are obtained only for a subcohort that is a simple random sample of the study cohort and for subjects who are known to have experienced the failure event, i.e., who have the right endpoint of the observed interval finite.

To the best of our knowledge, there is no method to date in the literature that deals with the general interval-censored case-cohort design described above, although several papers discuss related issues. Gilbert et al. (2005) considered the case-cohort design for a HIV vaccine trial where they treated the midpoint of the finite observed interval as the exact HIV infection time and then employed Self & Prentice (1988)'s method developed for right-censored case-cohort data to do the analysis. Li et al. (2008) presented a special interval-censored case-cohort design by assuming that the inspection time intervals are fixed and the same for all study subjects and the number of time intervals does not change with the sample size. Li & Nan (2011) considered fitting the relative risk regression model to the case-cohort sampled current status data, a special case of interval-censored data that arise when each study subject is examined only once for the occurrence of the failure event and thus the failure time is either left- or right-censored at the only examination. In this paper, we consider the case-cohort study design for general interval-censored failure time and develop a novel semiparametric method for fitting the proportional hazards model to data arising from this design.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Many authors have studied regression analysis of interval-censored data, obtained by simple random sampling, under the proportional hazards model. Among others, Finkelstein (1986) considered the maximum likelihood estimation with a discrete 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 function but remains computationally intensive; Satten et al. (1998) developed a rank-based procedure using imputed failure times, where a parametric baseline hazard is assumed; 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) represented the cumulative baseline hazard function as a monotone spline and then developed methods from Bayesian and frequentist perspectives, respectively, via two-stage Poisson data augmentations; Zhang et al. (2010) proposed a spline-based sieve semiparametric maximum likelihood method and proved that the resulting regression parameter estimator is asymptotically normal and efficient. Zhang et al. (2010) also provided a motivation of the sieve method, reasoning about the choice of basis functions, a theoretical framework and rigorous proofs based on empirical process theory. Besides having attractive asymptotic properties under various scenarios (e.g. Huang & Rossini, 1997; Shen, 1998; Xue et al., 2004), the sieve method is easy to implement and computationally fast as, for example, it usually involves much fewer parameters than a fully semiparametric method. In this paper, we focus on fitting the proportional hazards model to interval-censored data from the case-cohort design. We employ inverse probability weighting to construct the likelihood function and then, following the idea of Zhang et al. (2010), we develop a Bernstein-polynomial-based sieve likelihood estimation method. We also present a weighted bootstrap procedure for variance estimation.

2. Data, model and likelihood

Suppose that there are n independent subjects in a cohort study. Let T_i denote the failure time of subject i and Z_i a p-dimensional vector of covariates that may affect T_i. Suppose that the failure time is subject to interval-censoring and the full cohort data are denoted by

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

where U_i and V_i are two random examination times, and $(\Delta_{1i},\, 1-\Delta_{1i}-\Delta_{2i})$ indicate left- and right-censored observations, respectively.

Under our interval-censored case-cohort design, the covariates are obtained only for subjects from the subcohort as well as those who are known to have experienced the failure event, i.e., $\Delta_{1i}=1$ or $\Delta_{2i}=1$. Let ξ_i indicate that the covariate Z_i is obtained, $i=1,\ldots,n$. Then the observed data under our interval-censored case-cohort design can be represented by

$$O_i^{\xi}=\{U_i,\, V_i,\, \Delta_{1i}=I(T_i \le U_i),\, \Delta_{2i}=I(U_i<T_i \le V_i),\, \xi_i Z_i,\, \xi_i\},\ \ i=1,\ldots,n.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For the selection of the subcohort, we consider independent Bernoulli sampling with selection probability $q \in (0, 1)$. Thus, under our design, the probability that we observe the covariate Z_i is

$$\Pr(\xi_i=1) \equiv \pi_q(\Delta_{1i}, \Delta_{2i})=\Delta_{1i}+\Delta_{2i}+(1-\Delta_{1i}-\Delta_{2i})q, \ \ i=1, \ldots, n.$$

Since the covariates under our design can be considered as missing at random, we employ inverse probability weighting to construct the likelihood function. In particular, suppose that the failure time follows the proportional hazards model, under which the conditional cumulative hazard function of T_i given Z_i has the form

$$\Lambda(t|Z_i)=\Lambda(t)\exp(Z_i'\beta), \quad (1)$$

where β is a p-dimensional regression parameter and $\Lambda(t)$ is an unspecified cumulative baseline hazard function. Assume that T_i is conditionally independent of the examination times (U_i, V_i) given Z_i and the joint distribution of (U_i, V_i, Z_i) does not involve the parameters (β, Λ). Then that inverse probability weighted log-likelihood function has the form

$$l_n^w(\beta, \Lambda)=\sum_{i=1}^n l^w(\beta, \Lambda; O_i^\xi)=\sum_{i=1}^n w_i l(\beta, \Lambda; O_i)$$
$$=\sum_{i=1}^n w_i \left\{\Delta_{1i}\log\left[1-\exp\{-\Lambda(U_i)\exp(Z_i'\beta)\}\right]+\Delta_{2i}\log\left[\exp\{-\Lambda(U_i)\exp(Z_i'\beta)\} -\exp\{-\Lambda(V_i)\exp(Z_i'\beta)\}\right]-(1-\Delta_{1i}-\Delta_{2i})\Lambda(V_i \quad (2)$$

where the weight w_i is

$$w_i=\frac{\xi_i}{\pi_q(\Delta_{1i}, \Delta_{2i})}=\frac{\xi_i}{\Delta_{1i}+\Delta_{2i}+(1-\Delta_{1i}-\Delta_{2i})q}.$$

3. Sieve estimation and inference

Now we consider the estimation of $\theta = (\beta, \Lambda)$. Let

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

denote the parameter space of θ, where $\mathscr{B} = \{\beta \in R^p, \|\beta\| \leq M\}$, M is a positive constant, and $\mathscr{M}$ is the collection of all continuous nonnegative and nondecreasing functions over the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

interval $[\sigma, \tau]$. As defined in Condition (C1) in the Appendix, σ and τ are known constants usually taken in practice to be the lower and upper bounds of all observation times.

To estimate θ, it is natural to maximize the weighted log-likelihood (2). However, this is not easy, as l_n^w involves both the finite-dimensional regression parameter β and the infinite-dimensional nuisance parameter Λ. Since only the values of Λ at the examination times $\{U_i, V_i : i = 1,\ldots, n\}$ matter in the log-likelihood l_n^w, 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 l_n^w with respect to β and the jump sizes (Huang, 1996). However, such a fully semiparametric estimation method could involve a large number of parameters ($p + 2n$) if there are no ties among $\{U_i, V_i : i = 1,\ldots, n\}$. To ease the computation burden, by following the idea of Zhang et al. (2010), we propose a sieve estimation approach via Bernstein polynomials. In particular, we define the sieve space as

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

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

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

with $B_k(t, m, \sigma, \tau)$ 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 Θ_n. Because the cumulative baseline hazard function $\Lambda(t)$ is nonnegative and nondecreasing, it is desirable to restrict its estimate to be nonnegative and nondecreasing and we impose these constraints on the ϕ_k. One can show that $\Lambda(t)$ can be approximated by the Bernstein polynomial $\Lambda_n(t)$ with the coefficients $\phi_k = \Lambda(\sigma + (k/m)(\tau - \sigma))$ arbitrarily well as $n \to \infty$, that is, the sieve space Θ_n approximates the parameter space Θ arbitrarily well as $n \to \infty$ (Feller, 1971; Lorentz, 1986; Shen, 1997; Wang & Ghosh, 2012). We define the sieve likelihood estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ of θ to be the value of θ that maximizes the weighted log-likelihood function l_n^w over Θ_n. Compared to the fully semiparametric method, the sieve method significantly reduces the dimensionality of the optimization problem and relieves the computation burden.

We now establish the asymptotic properties of the proposed estimator $\hat{\theta}_n$. Let $O^\xi = \{U, V, \Delta_1, \Delta_2, \xi Z, \xi\}$ denote a single observation under our interval-censored case-cohort design and $G(u, v)$ the joint distribution function of the two random examination times (U, V). We assume both $G(u, v)$ and $g(u, v \mid z)$ to be unknown, where $g(u, v \mid z)$ is the conditional

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

density of (U, V) given $Z = z$ defined in Condition (C4) in the Appendix. For any $\theta^1 = (\beta^1, \Lambda^1)$ and $\theta^2 = (\beta^2, \Lambda^2)$ in the parameter space $\Theta = \mathcal{B} \otimes \mathcal{M}$, define a distance:

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

where $\|v\|$ denotes 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 $\theta_0 = (\beta_0, \Lambda_0)$ denote the true value of θ. The following theorems give the consistency and asymptotic normality of the proposed estimator $\hat{\theta}_n$ when $n \to \infty$. The proofs of these theorems and the regularity conditions needed for them are given in the Appendix.

Theorem 1

Assume that Conditions (C1) – (C5) given in the Appendix *hold. Then* $d(\hat{\theta}_n, \theta_0) \to 0$ *almost surely and* $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{(1-\nu)/2, \nu r/2\}})$, *where* $\nu \in (0, 1)$ *such that* $m = o(n^\nu)$ *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*

$$n^{1/2}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0) n^{-1/2} \sum_{i=1}^{n} w_i l^*(\beta_0, \Lambda_0; O_i) + o_p(1) \to N(0, \sum)$$

in distribution, where

$$\sum = I^{-1}(\beta_0) + I^{-1}(\beta_0) E\left\{ \frac{1 - \pi_q(\Delta_1, \Delta_2)}{\pi_q(\Delta_1, \Delta_2)} \{l^*(\beta_0, \Lambda_0; O)\}^{\otimes 2} \right\} I^{-1}(\beta_0).$$

with $v^{\otimes 2} = vv'$ *for a vector v, and* $I(\beta)$ *and* $l^*(\beta, \Lambda; O)$ *being the information and efficient score for* β *based on* $O = \{U, V, \Delta_1, \Delta_2, Z\}$, *respectively, which will be discussed in the* Appendix.

Note that $I(\beta)$ does not have an explicit expression, since its determination involves an integral equation which has no closed-form solution in general (Huang & Wellner, 1997). Thus, for variance estimation of $\hat{\beta}_n$, we suggest to employ the weighted bootstrap procedure of Ma & Kosorok (2005), which is easy to implement and works reasonably well in our setting. Let $\{u_1, \ldots, u_n\}$ denote n independent realizations of a bounded positive random variable u satisfying $E(u) = 1$ and $\text{var}(u) = \varepsilon_0 < +\infty$. Define the new weights $w_i^* = u_i w_i \ (i = 1, \ldots, n)$. Let $\hat{\theta}_n^* = (\hat{\beta}_n^*, \hat{\Lambda}_n^*)$ be the sieve estimator that maximizes the new weighted log-likelihood function $l_n^{w^*}$ over Θ_n, where $l_n^{w^*}$ is obtained by replacing w_i with w_i^* in l_n^w. If we generate B samples of $\{u_1, \ldots, u_n\}$ and obtain the corresponding $\hat{\beta}_n^*$, then the sample variance of these $\hat{\beta}_n^*$'s rescaled by ε_0 can be used to estimate the variance of $\hat{\beta}_n$. The

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

weighted bootstrap variance estimator is consistent under the assumptions of Theorem 2. In fact, this result can easily be seen from Theorem 2 of Ma & Kosorok (2005) and as commented by Ma & Kosorok (2005) right after their Theorem 2: once the asymptotic properties of the semiparametric M-estimators are established, the weighted bootstrap can be verified almost automatically. More details can be found in Ma & Kosorok (2005).

There are restrictions on the parameters due to nonnegativity and monotonicity, but they can be easily removed by reparameterization. For example, we may reparameterize the parameters $\{\phi_0,\ldots,\phi_m\}$ as the cumulative sums of $\{\exp(\phi_0^*),\ldots,\exp(\phi_m^*)\}$. Regarding the restriction $\Sigma_{0\le k\le m}|\phi_k|\le M_n$, since $M_n = O(n^a)$ is defined mainly for technical reasons and can be chosen reasonably large for fixed sample size in practice, we need not consider this restriction in computation. To obtain the proposed estimator $\hat{\theta}_n$, many existing optimization methods can be used, including the Nelder–Mead simplex algorithm and the Newton–Raphson method. For the numerical studies in Sections 4 and 5, the Nelder–Mead simplex algorithm in *fminsearch* in Matlab was used. One also needs to specify the degree of Bernstein polynomials m, which controls the smoothness of the approximation. For this, we suggest to consider several different values of m and choose the one that minimizes

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

More guidelines and discussion on the choice of m will be given below. The Matlab code that implements the proposed inference procedure is available in the Supplementary Material.

4. A simulation study

In this section, we perform a simulation study to evaluate the finite-sample performance of the proposed method. We assumed that the covariate Z had the standard normal distribution and that given Z, the failure time T followed the proportional hazards model (1) with the cumulative baseline hazard function $\Lambda(t) = 0.2t^2$. We considered $\beta = 0$ or log 2.

To generate interval-censored data $\{U_i, V_i, \Delta_{1i} = I(T_i \le U_i), \Delta_{2i} = I(U_i < T_i \le V_i) : i = 1,\ldots, n\}$, we mimicked biomedical follow-up studies. In particular, we assumed that each study subject was scheduled to be examined at k different follow-up time points within the interval $[0, \tau]$ in addition to the baseline exam at time 0. More specifically, to mimic the Atherosclerosis Risk in Communities study, we chose k equally spaced time points over the interval $[0, \tau]$ denoted by $e_1,\ldots, e_k$. For each subject, the k scheduled follow-up time points were generated as e_i plus a uniform random variable on $[-\tau/\{3(k+1)\}, \tau/\{3(k+1)\}]$, $i = 1, \ldots, k$. At each of these time points, it was assumed that a subject could miss the scheduled examination with probability ζ, independent of the examination results at other time points. For subject i, if the failure event had already occurred at the first follow-up examination, we defined U_i to be the first follow-up examination time, $V_i = \tau$ and $(\Delta_{1i}, \Delta_{2i}) = (1, 0)$; if the failure event had not yet occurred at the last follow-up examination, we defined V_i to be the last follow-up examination time, $U_i = 0$ and $(\Delta_{1i}, \Delta_{2i}) = (0, 0)$; otherwise, we defined U_i and V_i to be the two consecutive follow-up examination times bracketing T_i and $(\Delta_{1i}, \Delta_{2i}) = (0,$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1). We used $k = 8$ and $\zeta = 0.2$, and determined the length of study τ according to the desired proportion of events, i.e., subjects with $\Delta_1 = 1$ or $\Delta_2 = 1$. Regarding the proportion of events, we considered 0.05 or 0.15.

To generate the subcohort, we employed independent Bernoulli sampling with selection probability $q = 0.2$. For the variance estimation of the proposed estimator $\hat{\beta}_n$, we used the weighted bootstrap procedure described in Section 3 and generated the random sample $\{u_1, \ldots, u_n\}$ from the exponential distribution.

Table 1 presents the simulation results for the estimation of β based on the proposed method, denoted by $\hat{\beta}_{prop}$, when the cohort size is $n = 500$, 1000 or 2000. These results were obtained from 1000 replicates and the variance estimate was calculated based on 200 bootstrap samples. For comparison, we also provided in Table 1 the estimation results using the sieve maximum likelihood method based on: (i) the subcohort only, denoted by $\hat{\beta}_{sub}$; (ii) a simple random sample of the cohort which has the same size as the case-cohort sample, denoted by $\hat{\beta}_{srs}$. For the degree of Bernstein polynomials, we used $m = 3$ for all methods and situations considered.

Table 1 shows that the proposed estimator is virtually unbiased. The variance estimates based on the weighted bootstrap procedure are close to the corresponding empirical variances and yield reasonable coverages. In addition, under all situations considered, the proposed estimator is more efficient than the estimators based on subcohort only or a simple random sample of the same size as the case-cohort sample. Especially, when the cohort size is 500 or 1000 and the proportion of events is 0.05, the subcohort-based and simple-random-sample-based estimators yield larger biases and inflated variances while the proposed estimator still has good performance. We also conducted simulations with $\Lambda(t) = 0.1t$, $k = 6$, $\zeta = 0.3$, $q = 0.25$ and $m = 4$ or 5 as well as other methods for generating interval-censored data and obtained similar results. In particular, the results seem to be fairly robust to the choice of m.

5. An application

In this section, we illustrate the proposed method using data from the Atherosclerosis Risk in Communities study, a longitudinal epidemiologic observational study consisting of men and women aged 45–64 at baseline, recruited from four US field centers (Forsyth County, NC (Center-F), Jackson, MS (Center-J), Minneapolis Suburbs, MN (Center-M) and Washington County, MD (Center-W)). Forsyth County, Minneapolis Suburbs, and Washington County include white participants, and Forsyth County and Jackson Center include African American participants. The study began in 1987 and the participants received an extensive examination, including medical, social and demographic data. These participants were scheduled to be re-examined on average of every three years with the first exam occurring in 1987–89, the second in 1990–92, the third in 1993–95 and the fourth in 1996–98. There were participants that missed some scheduled re-visits and thus had less than four follow-up examinations. For each participant, the occurrence of a disease such as diabetes can be observed only between two consecutive examinations and therefore only interval-censored failure time data were available. We illustrate the proposed method by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

investigating the effect of high-density lipoprotein cholesterol level on the risk of diabetes after adjusting for confounding variables and other risk factors in white women younger than 55 years based on data from an interval-censored case-cohort sample. Specifically, we constructed the interval-censored case-cohort sample in the following way. The cohort of interest consists of 2799 white women younger than 55 years and 202 were observed to have developed diabetes during the study. We selected a simple random sample of the cohort by Bernoulli sampling and set the selection probability equal to $q = 0.1$. The subcohort had 272 subjects and the final case-cohort sample had 451 subjects. We considered the proportional hazards model

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

where the vector of covariates Z included high-density lipoprotein cholesterol level, total cholesterol level, body mass index, age, smoking status, and indicators for field centers where Center-M was chosen as reference. We fitted this model using the proposed method and presented the results in Table 2. For comparison, we also provide in Table 2 the analysis results based on the subcohort only. Regarding the degree of Bernstein polynomials, we chose $m = 3$ for both analyses according to the AIC criterion described in Section 3. One can see from Table 2 that the proposed method based on the case-cohort sample yielded smaller standard errors and more significant results compared to the method based on the subcohort only. In particular, the results suggest that higher high-density lipoprotein cholesterol, lower total cholesterol and lower body mass index levels are significantly associated with lower risk of diabetes in white women younger than 55 years.

6. Concluding remarks

There are some practical considerations for the implementation of the proposed design and method. First, under our design, the subcohort is a simple random sample of the cohort selected by independent Bernoulli sampling. When the subcohort is selected by sampling without replacement, our method should work, though more complicated arguments would be needed to develop the asymptotic results (Saegusa & Wellner, 2013). Moreover, when some covariates are available for all cohort members, a stratified case-cohort design based on those covariates could be considered to improve the study efficiency and adapting our method to such design should be straightforward. Second, regarding the degree of Bernstein polynomials m, there does not seem to be a single true value. According to the simulation studies, the results seem to be fairly robust to the choice of m. In practice, we suggest to consider several different values such as $m = 3$ to 8 and base the selection on the AIC criterion. Although similar strategies are commonly used in the literature (e.g. Wang et al., 2016), further study on AIC and other model selection criteria or methods in this setting would be appreciated. Third, assessing the goodness-of-fit of the proportional hazards model is an important practical issue. Ren & He (2011) and Wang et al. (2006) considered this problem for univariate and correlated interval-censored data, respectively, obtained by simple random sampling. Extensions of these methods to the case-cohort design warrant future research. Lastly, as suggested by the Associate Editor, the missing data problem may

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

arise when the covariates are not obtainable for some subjects in the case-cohort sample. Accommodating such situation would be practically useful and merits further investigation. Another interesting future research direction, suggested by a referee, is to consider cost-effective sampling designs for more general types of censored or truncated data (e.g. Turnbull, 1976; Huber et al., 2009).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Editor, Associate Editor and two referees for their valuable comments which have led to significant improvement of the paper. This work was partially supported by grants from the National Institutes of Health. 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 Atherosclerosis Risk in Communities study for their important contributions.

References

Chen K. Generalized case–cohort sampling. J R Statist Soc B. 2001; 63:791–809.

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

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]

Feller, W. An Introduction to Probability Theory and Its Applications, Volume II. John Wiley; 1971.

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. J Infect Dis. 2005; 191:666–677. [PubMed: 15688279]

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

Huang J, Rossini A. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. J Am Statist Assoc. 1997; 92:960–967.

Huang, J., Wellner, JA. Interval censored survival data: a review of recent progress. Proceedings of the First Seattle Symposium in Biostatistics; Springer; 1997.

Huber C, Solev V, Vonta F. Interval censored and truncated data: Rate of convergence of NPMLE of the density. J Statist Plann Inference. 2009; 139:1734–1749.

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

Keogh RH, White IR. Using full-cohort data in nested case–control and case–cohort studies by multiple imputation. Statist Med. 2013; 32:4021–4043.

Kim S, Cai J, Lu W. More efficient estimators for case-cohort studies. Biometrika. 2013; 100:695–708. [PubMed: 24634519]

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. Canad J Statist. 2011; 39:557–577.

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

Ma S, Kosorok MR. Robust semiparametric M-estimation and the weighted bootstrap. J Multivar Anal. 2005; 96:190–217.

Marti H, Chavance M. Multiple imputation analysis of case–cohort studies. Statist Med. 2011; 30:1595–1607.

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.

Ren JJ, He B. Estimation and goodness-of-fit for the Cox model with various types of censored data. J Statist Plann Inference. 2011; 141:961–971.

Saegusa T, Wellner JA. Weighted likelihood estimation under two-phase sampling. Ann Statist. 2013; 41:269–295.

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

Satten GA, Datta S, Williamson JM. Inference based on imputed failure times for the proportional hazards model with interval-censored data. J Am Statist Assoc. 1998; 93:318–327.

Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under case–cohort sampling. Scand J Statist. 2004; 31:283–293.

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

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

Shen X. Propotional odds regression and sieve maximum likelihood estimation. Biometrika. 1998; 85:165–177.

Shen X, Wong WH. Convergence rate of sieve estimates. Ann Statist. 1994; 22:580–615.

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

Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. J R Statist Soc B. 1976; 38:290–295.

van der Vaart, AW., Wellner, JA. Weak Convergence and Empirical Processes: With Applications to Statistics. New York: Springer; 1996.

Wang J, Ghosh SK. Shape restricted nonparametric regression with bernstein polynomials. Comput Statist Data Anal. 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]

Wang L, Sun L, Sun J. A goodness-of-fit test for the marginal Cox model for correlated interval-censored failure time data. Biom J. 2006; 48:1020–1028. [PubMed: 17240659]

Xue H, Lam K, Li G. Sieve maximum likelihood estimator for semiparametric regression models with current status data. J Am Statist Assoc. 2004; 99:346–356.

Zeng D, Lin DY. Efficient estimation of semiparametric transformation models for two-phase cohort studies. J Am Statist Assoc. 2014; 109:371–383.

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. Scand J Statist. 2010; 37:338–354.

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]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Appendix

Proofs of Theorems 1 and 2

In this appendix, we provide the proofs of Theorems 1 and 2. Denote the observation on a single subject under our interval-censored case-cohort design by $O^{\xi} = \{U, V, \Delta_1 = I(T \le U), \Delta_2 = I(U < T \le V), \xi Z, \xi\}$, where U and V are two random examination times, $(\Delta_1, 1 - \Delta_1 - \Delta_2)$ indicate left- and right-censored observations, respectively, and ξ indicates the covariate Z being observed with $\Pr(\xi = 1) \equiv \pi_q(\Delta_1, \Delta_2) = \Delta_1 + \Delta_2 + (1 - \Delta_1 - \Delta_2)q$. Before proving the theorems, we first describe the regularity conditions needed as follows:

- **(C1)** There exists $\eta > 0$ such that $P(V - U \ge \eta) = 1$. 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 has a bounded support and is not concentrated on any proper subspace of R^p. Also $E\{\mathrm{var}(Z|U)\}$ and $E\{\mathrm{var}(Z|V)\}$ are positive definite.
- **(C3)** For $r = 1$ or 2, the function $\Lambda_0 \in \mathcal{M}$ 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 $\mathcal{B} \subset R^p$.
- **(C4)** The conditional density $g(u, v \mid z)$ 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)** $0 < q \le \pi_q(\Delta_1, \Delta_2) \le 1$, where q is a known constant.

Note that Conditions (C1) – (C4) are commonly used in the studies of interval-censored data (Huang & Rossini, 1997; Zhang et al., 2010) and are usually satisfied in practice. In the following, we will prove Theorems 1 and 2 under these conditions by employing the empirical process theory and some nonparametric methods or techniques. For the proofs, define $Pf = \int f(y)dP(y)$, the expectation of $f(Y)$ taken under the distribution P, and $P_n f = n^{-1}\sum_{i=1}^{n} f(Y_i)$, the expectation of $f(Y)$ under the empirical measure P_n.

Proof of Theorem 1

We first prove the strong consistency of $\hat{\theta}_n$. Let $l^w(\theta, O^{\xi})$ denote the weighted log-likelihood function based on a given single observation O^{ξ} and consider the class of functions $\mathcal{L}_n = \{l^w(\theta, O^{\xi}) = w\, l(\theta, O) : \theta \in \Theta_n\}$ where the functions are random variables on the probability space indexed by θ. Then based on Lemma 1 given in the Supplementary Material, the covering number of $\mathcal{L}_n$ satisfies

$$N\{\varepsilon, \mathcal{L}_n, L_1(P_n)\} \le K M_n^{(m+1)} \varepsilon^{-(p+m+1)}.$$

Furthermore, by Lemma 2 given in the Supplementary Material, we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\sup_{\theta\in\Theta_n}|P_n l^w(\theta,O^\xi)-Pl^w(\theta,O^\xi)|\ \to 0 \qquad \text{(A.1)}$$

almost surely. Let $M(\theta, O^\xi) = -l^w(\theta, O^\xi)$, and define $K_\varepsilon = \{\theta : d(\theta, \theta_0) \geq \varepsilon, \theta \in \Theta_n\}$ for $\varepsilon > 0$ and

$$\zeta_{1n}=\sup_{\theta\in\Theta_n}|P_nM(\theta,O^\xi)-PM(\theta,O^\xi)|,\ \ \zeta_{2n}=P_nM(\theta_0,O^\xi)-PM(\theta_0,O^\xi).$$

Then

$$\inf_{K_\varepsilon}PM(\theta,O^\xi)=\inf_{K_\varepsilon}\left\{PM(\theta,O^\xi)-P_nM(\theta,O^\xi)+P_nM(\theta,O^\xi)\right\}\leq\zeta_{1n}+\inf_{K_\varepsilon}P_nM(\theta,O^\xi).$$

(A.2)

If $\hat{\theta}_n \in K_\varepsilon$, then we have

$$\inf_{K_\varepsilon}P_nM(\theta,O^\xi)=P_nM(\hat{\theta}_n,O^\xi)\leq P_nM(\theta_0,O^\xi)=\zeta_{2n}+PM(\theta_0,O^\xi). \qquad \text{(A.3)}$$

Define $\delta_\varepsilon = \inf_{K_\varepsilon} PM(\theta_0, O^\xi) - PM(\theta_0, O^\xi)$. Then under Condition (C2), using the same arguments as those in Zhang et al. (2010, p. 352), we can prove $\delta_\varepsilon > 0$. It follows from (A.2) and (A.3) that

$$\inf_{K_\varepsilon}PM(\theta,O^\xi)\leq\zeta_{1n}+\zeta_{2n}+PM(\theta_0,O^\xi)=\zeta_n+PM(\theta_0,O^\xi)$$

with $\zeta_n = \zeta_{1n} + \zeta_{2n}$, and hence $\zeta_n \geq \delta_\varepsilon$. This gives $\{\hat{\theta}_n \in K_\varepsilon\} \subseteq \{\zeta_n \geq \delta_\varepsilon\}$, and by (A.1) and the strong law of large numbers, we have both $\zeta_{1n} \to 0$ and $\zeta_{2n} \to 0$ almost surely. Therefore, $\cup_{k=1}^{\infty}\cap_{n=k}^{\infty}\{\hat{\theta}_n \in K_\varepsilon\} \subseteq \cup_{k=1}^{\infty}\cap_{n=k}^{\infty}\{\zeta_n \geq \delta_\varepsilon\}$, which proves that $d(\hat{\theta}_n, \theta_0) \to 0$ almost surely.

Now we will show the convergence rate of $\hat{\theta}_n$ by using Theorem 3.4.1 of van der Vaart & Wellner (1996). Below we use $\tilde{K}$ to denote a universal positive constant which 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 $\|\Lambda_{n0} - \Lambda_0\|_\infty = O(m^{-r/2})$. Define $\theta_{n0} = (\beta_0, \Lambda_{n0})$. Then we have $d(\theta_{n0}, \theta_0) = O(n^{-rv/2})$. For any $\eta > 0$, define the class of functions $\mathcal{F}_\eta = \{l^w(\theta, O^\xi) - l^w(\theta_{n0}, O^\xi) : \theta \in \Theta_n, \eta/2 < d(\theta, \theta_{n0}) \leq \eta\}$ for a given single observation O^ξ, where the functions are random variables on the probability space indexed by θ. One can easily show

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that $P(l^w(\theta_0, O^\xi) - l^w(\theta_{n0}, O^\xi)) \le \tilde{K}d(\theta_0, \theta_{n0}) \le \tilde{K}n^{-rv/2}$. Also under Condition (C2), using the same arguments as those in Zhang et al. (2010, p. 352), we obtain $P(l^w(\theta_0, O^\xi) - l^w(\theta, O^\xi)) \le \tilde{K}d^2(\theta_0, \theta)$. Thus, for large n, we have $P(l^w(\theta, O^\xi) - l^w(\theta_{n0}, O^\xi)) = P(l^w(\theta, O^\xi) - l^w(\theta_0, O^\xi)) + P(l^w(\theta_0, O^\xi) - l^w(\theta_{n0}, O^\xi)) \le \tilde{K}\eta^2 + \tilde{K}n^{-rv/2} = \tilde{K}\eta^2$, for any $l^w(\theta, O^\xi) - l^w(\theta_{n0}, O^\xi) \in \mathcal{F}_\eta$.

Following the calculations in Shen & Wong (1994, p. 597), we can establish that for $0 < \varepsilon < \eta$, $\log N_{[]}(\varepsilon, \mathcal{F}_\eta, L_2(P)) \le \tilde{K}N \log(\eta/\varepsilon)$ with $N = m + 1$. Moreover, some algebraic manipulations yield that $P(l^w(\theta, O^\xi) - l^w(\theta_{n0}, O^\xi))^2 \le \tilde{K}\eta^2$ for any $l^w(\theta, O^\xi) - l^w(\theta_{n0}, O^\xi) \in \mathcal{F}_\eta$. Under Conditions (C1) – (C5), it is easy to see that $\mathcal{F}_\eta$ is uniformly bounded. Therefore, by Lemma 3.4.2 of van der Vaart & Wellner (1996), we obtain

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

where $J_{[]}\{\eta, \mathcal{F}_\eta, L_2(P)\} = \int_0^\eta [1 + \log N_{[]}\{\varepsilon, \mathcal{F}_\eta, L_2(P)\}]^{1/2} d\varepsilon \le \tilde{K}N^{1/2}\eta$. This yields $\phi_n(\eta) = N^{1/2}\eta + N/n^{1/2}$. It is easy to see that $\phi_n(\eta)/\eta$ is decreasing in η, and $r_n^2 \phi_n(1/r_n) = r_N N^{1/2} + r_n^2 N/n^{1/2} \le \tilde{K}n^{1/2}$; where $r_n = N^{-1/2}n^{1/2} = n^{(1-v)/2}$.

Finally note that $P_n\{l^w(\hat{\theta}_n, O^\xi) - l^w(\theta_{n0}, O^\xi) \ge 0$ and $d(\hat{\theta}_n, \theta_{n0}) \le d(\hat{\theta}_n, \theta_0) + d(\theta_0, \theta_{n0}) \to 0$ in probability. Thus by applying Theorem 3.4.1 of van der Vaart & Wellner (1996), we have $n^{(1-v)/2}d(\theta_n, \theta_{n0}) = O_p(1)$. This together with $d(\theta_{n0}, \theta_0) = O(n^{-rv/2})$ yields that $d(\hat{\theta}_n, \theta_0) = O_p(n^{-(1-v)/2} + n^{-rv/2})$ and the proof is completed.

Proof of Theorem 2

Now we will prove the asymptotic normality of $\hat{\beta}_n$. Note that $w = \xi/\pi_q(\Delta_1, \Delta_2)$ is bounded and does not depend on the parameters (β, Λ), and $E\{w|O\} = 1$. Following the proof of Theorem 2 in Zhang et al. (2010), one can obtain that

$$n^{1/2}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0) n^{-1/2} \sum_{i=1}^n w_i l^*(\beta_0, \Lambda_0; O_i) + o_p(1),$$

where $l^*(\beta, \lambda; O)$ and $I(\beta)$, the efficient score and information for β based on $O = \{U, V, \Delta_1, \Delta_2, Z\}$, are defined as in Zhang et al. (2010, p. 344) with our parameters (β, Λ) corresponding to theirs $(\theta, \exp(\phi))$. Note that

$$\begin{aligned}
\mathrm{var}\{wl^*(\beta_0, \Lambda_0; O)\} &= \mathrm{var}\{E\{wl^*(\beta_0, \Lambda_0; O)|O\}\} + E\{\mathrm{var}\{wl^*(\beta_0, \Lambda_0; O)|O\}\} \\
&= \mathrm{var}\{l^*(\beta_0, \Lambda_0; O)\} + E\left\{\mathrm{var}(\xi|O) \frac{\{l^*(\beta_0, \Lambda_0; O)\}^{\otimes 2}}{\pi_q^2(\Delta_1, \Delta_2)}\right\} \\
&= I(\beta_0) + E\left\{\frac{1 - \pi_q(\Delta_1, \Delta_2)}{\pi_q(\Delta_1, \Delta_2)} \{l^*(\beta_0, \Lambda_0; O)\}^{\otimes 2}\right\}.
\end{aligned}$$

Thus, we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$n^{1/2}(\hat{\beta}_n - \beta_0) \to N(0, \sum)$$

in distribution, where

$$\sum = I^{-1}(\beta_0) + I^{-1}(\beta_0) E \left\{ \frac{1 - \pi_q(\Delta_1, \Delta_2)}{\pi_q(\Delta_1, \Delta_2)} \{ l^*(\beta_0, \Lambda_0; O) \}^{\otimes 2} \right\} I^{-1}(\beta_0).$$

This completes the proof of Theorem 2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Simulation results for comparing different methods

			$\beta = 0$					$\beta = \log 2$				
n	**p(event)**		**Bias**	**SD**	**SE**	**CP**	**RE**	**Bias**	**SD**	**SE**	**CP**	**RE**
500	0.05	$\hat{\beta}_{sub}$	0.3	55.2	38.9	83	0.2	6.9	71.7	44.6	83	0.2
		$\hat{\beta}_{srs}$	−1.2	48.1	37.4	85	0.2	3.6	52.1	40.9	89	0.3
		$\hat{\beta}_{prop}$	0.3	23.2	22.7	95	1.0	4.0	27.7	25.2	92	1.0
	0.15	$\hat{\beta}_{sub}$	1.1	27.2	26.3	94	0.3	2.0	29.7	28.6	93	0.4
		$\hat{\beta}_{srs}$	−0.4	22.1	20.8	95	0.5	1.9	22.6	22.2	95	0.6
		$\hat{\beta}_{prop}$	−0.0	15.3	15.3	95	1.0	1.8	17.8	16.6	92	1.0
1000	0.05	$\hat{\beta}_{sub}$	0.0	33.3	30.2	91	0.2	1.0	35.7	31.5	90	0.3
		$\hat{\beta}_{srs}$	−0.2	30.9	27.3	91	0.3	1.2	31.2	29.2	92	0.3
		$\hat{\beta}_{prop}$	0.5	16.2	15.8	95	1.0	1.0	17.8	17.7	94	1.0
	0.15	$\hat{\beta}_{sub}$	0.8	19.4	18.4	94	0.3	1.3	20.3	19.5	94	0.3
		$\hat{\beta}_{srs}$	0.2	14.7	14.6	95	0.5	1.2	15.1	15.2	96	0.6
		$\hat{\beta}_{prop}$	0.5	10.3	10.7	96	1.0	1.1	11.5	11.8	95	1.0
2000	0.05	$\hat{\beta}_{sub}$	0.2	21.7	21.8	94	0.2	0.3	24.4	22.6	93	0.3
		$\hat{\beta}_{srs}$	0.4	20.9	19.8	93	0.3	0.1	21.6	20.6	93	0.4
		$\hat{\beta}_{prop}$	−0.1	10.5	11.2	96	1.0	0.3	12.8	12.6	95	1.0
	0.15	$\hat{\beta}_{sub}$	0.5	13.3	13.0	94	0.3	1.5	13.9	13.7	94	0.4
		$\hat{\beta}_{srs}$	0.5	10.8	10.3	95	0.5	0.7	10.7	10.8	95	0.6
		$\hat{\beta}_{prop}$	0.5	7.3	7.5	95	1.0	1.0	8.6	8.4	93	1.0

n, cohort size; p(event), proportion of events; Bias, 100 * (mean($\hat{\beta}$) − β); SD, 100 * sample standard deviation; SE, 100 * average of standard error estimates obtained from the weighted bootstrap procedure; CP, empirical coverage percentage of the 95% confidence interval; RE, relative efficiency calculated as $SD^2(\hat{\beta}_{prop})/SD^2(\hat{\beta}_{sub})$ and $SD^2(\hat{\beta}_{prop})/SD^2(\hat{\beta}_{srs})$, respectively. $\hat{\beta}_{prop}$, proposed method; $\hat{\beta}_{sub}$, method with subcohort only; $\hat{\beta}_{srs}$, method with a simple random sample of the same size as the case-cohort sample.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Analysis results for diabetes data from the ARIC study

	Proposed method			Subcohort only		
Variable	$\hat{\beta}$	**SE**	**P-value**	$\hat{\beta}$	**SE**	**P-value**
High-density Lipoprotein Cholesterol	−0.028	0.006	0.000	−0.026	0.011	0.012
Total Cholesterol	0.005	0.002	0.021	0.002	0.003	0.504
Body Mass Index	0.115	0.024	0.000	0.118	0.038	0.002
Current Smoking	−0.305	0.314	0.331	−0.149	0.609	0.807
Age	0.006	0.061	0.923	0.136	0.113	0.228
Center-F	−0.195	0.284	0.492	−0.337	0.524	0.520
Center-W	0.094	0.269	0.728	−0.037	0.465	0.936

$\hat{\beta}$, estimate of β; SE, standard error estimate; P-value, p-value for testing $H_0 : \beta = 0$.