Published in final edited form as:

Lifetime Data Anal. 2017 July ; 23(3): 377–399. doi:10.1007/s10985-016-9363-2.

Analysis of two-phase sampling data with semiparametric additive hazards models

Yanqing Sun¹, Xiyuan Qian², Qiong Shou³, and Peter B. Gilbert⁴

¹Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, 28223, USA

²Department of Mathematics, East China University of Science and Technology, Shanghai, China

³Merck China & Co., Inc., Beijing, China

⁴University of Washington and Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA

Abstract

Under the case-cohort design introduced by Prentice (1986), the covariate histories are ascertained only for the subjects who experience the event of interest (i.e., the cases) during the follow-up period and for a relatively small random sample from the original cohort (i.e., the subcohort). The case-cohort design has been widely used in clinical and epidemiological studies to assess the effects of covariates on failure times. Most statistical methods developed for the case-cohort design use the proportional hazards model, and few methods allow for time-varying regression coefficients. In addition, most methods disregard data from subjects outside of the subcohort, which can result in inefficient inference. Addressing these issues, this paper proposes an estimation procedure for the semiparametric additive hazards model with case-cohort/two-phase sampling data, allowing the covariates of interest to be missing for cases as well as for non-cases. A more flexible form of the additive model is considered that allows the effects of some covariates to be time varying while specifying the effects of others to be constant. An augmented inverse probability weighted estimation procedure is proposed. The proposed method allows utilizing the auxiliary information that correlates with the phase-two covariates to improve efficiency. The asymptotic properties of the proposed estimators are established. An extensive simulation study shows that the augmented inverse probability weighted estimation is more efficient than the widely adopted inverse probability weighted complete-case estimation method. The method is applied to analyze data from a preventive HIV vaccine efficacy trial.

Keywords

Asymptotics; augmented inverse probability weighted estimation; auxiliary variables; double robustness; efficiency; estimating equations; HIV vaccine efficacy trial; inverse probability weighted complete-case; parametric regression; time-varying effects

Corresponding Author: Yanqing Sun, yasun@uncc.edu.

Supplementary Materials

The proofs of Theorem 1 to 3 in this article are available at the online Supplementary Materials at the website of *Lifetime Data Analysis*.

1 Introduction

In many medical and epidemiological studies, the covariates are not observed for all study subjects. Under the case-cohort design introduced by Prentice (1986), the covariate histories are ascertained only for the subjects who experience the event of interest during the follow-up period (the cases) and for a relatively small random sample (the subcohort) from the original cohort. The case-cohort design has been widely used in clinical and epidemiological studies to assess the effects of possibly time-dependent covariates on a failure time. This design is especially useful in large studies with infrequent occurrence of the failure event, for which the assembly of covariate histories from all cohort members may be prohibitively expensive. The case-cohort data are a biased sample from the study population and thus applying standard methods for randomly sampled data may result in biased estimation.

There is an extensive literature in the analysis of case-cohort data. Most statistical methods for case-cohort studies are based on modifications of the full data partial likelihood score function for the Cox proportional hazards model, which weight the contributions of cases and subcohort members by the inverses of true or estimated sampling probabilities. We refer to Prentice (1986), Self and Prentice (1988), Kalbfleisch and Lawless (1988), Lin and Ying (1993), Barlow (1994), Chen and Lo (1999), Borgan et al. (2000), Chen (2001), Kulich and Lin (2004), and Samuelsen, Ånestad and Skrdal (2007), among others. An alternative approach, which relaxes the assumption that the hazard function conditional on covariates of interest included in the model equals the hazard function conditional on these covariates plus the phase-one categorical variable used for phase-two stratified sampling, is based on weighted likelihood (Breslow and Wellner, 2007; Li, Gilbert, and Nan, 2008; Breslow et al., 2009a; Breslow et al., 2009b).

The Cox model assumes that the hazard functions associated with different covariate values are proportional over time. This assumption may be too restrictive and the Cox model does not always fit data well in practice. Many alternative models have been studied, including the proportional odds model (cf. Murphy, Rossini and van der Vaart (1997)), the accelerated failure time model (cf. Jin et al. (2003)), the linear transformation model (cf. Cheng, Wei and Ying (1995)), and the additive hazards model (cf. Aalen (1980)). The analysis of case-cohort data with the additive hazards model with constant covariate effects was studied by Kulich and Lin (2000) by modifying the pseudo-score equation of Lin and Ying (1994). Chen (2001) proposed a weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. By weighting the full cohort estimating function, Kong and Cai (2009) developed statistical methods for analyzing case-cohort data with a semiparametric accelerated failure time model. Kang, Cai and Chambless (2013) recently proposed an estimation method for case-cohort data with the simple additive model of Lin and Ying (1994) that allows only constant covariate effects. Nan and Wellner (2013) established the asymptotic theory for a general semiparametric Z-estimation approach for case-cohort studies that included the Cox model and the additive hazards model of Lin and Ying (1994). Many of the above-referenced case-cohort methods accommodate two-phase sampling, where the phase-one data are measured in all subjects (including the failure time, censoring time, and covariates) and the phase-two data are

measured from a stratified random sample. A discrete stratification variable based on the phase one data is specified (which usually includes case status such that the sampling is outcome-dependent), and within each stratum subjects are selected into phase two. Some authors have restricted the term “two-phase sampling” to without replacement stratified sampling (Breslow et al., 2009a, 2009b, Saegusa and Wellner, 2013), whereas others have broadened the definition to include either Bernoulli sampling or without replacement sampling (Breslow and Lumley, 2013); here we adopt the broader definition. Most theoretical results for two-phase sampling methods have assumed Bernoulli sampling, which facilitates easier proofs of asymptotic properties because the sampling indicators are iid. Saegusa and Wellner (2013) is an exception that tackled the challenge of dependent sampling indicators in proving theoretical results under two-phase stratified without replacement sampling. In this paper we develop the results under Bernoulli two-phase sampling, which matches the sampling design of the real data application described below.

The current research is motivated by an HIV vaccine efficacy trial known as the Vax-Gen 004 trial (Flynn et al., 2005) conducted by VaxGen Inc. from 1998 to 2003. 5403 HIV uninfected volunteers at high risk for acquiring HIV infection were randomized to receive the vaccine AIDSVAX (3, 598) or placebo (1, 805). Subjects were monitored for 3 years for the primary study endpoint of HIV infection. One of the immune responses measured at the Month 12.5 visit in the trial, CD4i_HIV2_IC50, is a labor-intensive assay developed by George Shaw (cf., Gottschalk and Dunn, 2005), which measures neutralizing antibodies that specifically target CD4-induced antibody epitopes. CD4i_HIV2_IC50 is a phase-two covariate measured for 52 of the 95 subjects who became HIV-1 positive during the trial (cases) and for 39 of those (1073) who remained HIV-1 negative (non-cases). Anti-CD4i antibodies have been hypothesized to be a potential mechanism of vaccine-induced protection against HIV-1 infection (Gilbert, et al., 2005). Detecting an inverse relationship between HIV-1 infection risk and CD4i_HIV2_IC50 in vaccine recipients would generate hypotheses about the potential role of the antibodies as a “correlate of protection” (Plotkin and Gilbert, 2012), thereby providing guidance for HIV-1 vaccine research and development. In addition, the effect of CD4i_HIV2_IC50 is likely to wane over time, motivating use of a model not assuming proportional hazards.

In this paper, we develop some efficient estimation procedures for analyzing Bernoulli two-phase sampling data under the general semiparametric additive hazards models of Huffer and McKeague (1991). The model includes both time-varying and time-invariant covariate effects to allow for flexible modeling in practice. The approach of Kang, Cai and Chambless (2013) and most of the existing approaches for the additive hazards models under the case-cohort design are based on the inverse probability weighting of complete-case technique of Horvitz and Thompson (1952). With this approach, if a subject has a missing value for one covariate, then the observed values of other covariates together with the observed failure/censoring time of the same subject are not utilized. This leads to loss of efficiency. By adapting the idea of Robins, Rotnitzky and Zhao (1994), we propose an augmented estimating equation on the basis of the inverse probability weighting of complete cases to improve efficiency. It is well known that the augmented inverse probability weighting of complete-case method is doubly robust and is more efficient than the inverse probability weighted complete-case approach when the augmented part is correctly specified (Tsiatis,

2006). The proposed method also utilizes auxiliary variables that have the potential to influence the sampling probabilities and that may improve efficiency through their correlation with the phase-two covariates.

The rest of the paper is organized as follows. Section 2 develops an augmented inverse probability weighted estimation procedure under the semiparametric additive hazards model for failure time data with missing covariates. The asymptotic properties of the proposed estimators are investigated in Section 3. The finite-sample properties of the estimators are examined in Section 4 through a simulation study. The method is applied to analyze data from the VaxGen 004 trial in Section 5. A discussion is given in Section 6. The proofs of the theorems are placed in the online Supplementary Material.

2 Estimation of semiparametric additive hazards models with two-phase sampling data

2.1 Preliminaries

Let $U(\cdot) = \{U(t), 0 \leq t \leq \tau\}$ and $Z(\cdot) = \{Z(t), 0 \leq t \leq \tau\}$ be p and r -dimensional phase-one covariate processes, respectively, where $\tau < \infty$ denotes the time when follow-up ends. Suppose that V is a q -dimensional vector of phase-two covariates. The phase-one covariates $U(\cdot)$ and $Z(\cdot)$ are observed for all the cohort members but the phase-two covariates V are only observed for a subset (subcohort/phase-two sample) of the study subjects. We assume that the conditional hazard function of the failure time T given the covariates $\{U(t), Z(t), V, 0 \leq t \leq \tau\}$ follows the semiparametric additive hazards model

$$h(t|X(t), Z(t)) = \beta_1^T(t)U(t) + \beta_2^T(t)V + \gamma^T Z(t), \quad (1)$$

where $\beta_1(t)$ and $\beta_2(t)$ are p and q -dimensional vectors of time-varying regression coefficients, respectively, and γ is a r -dimensional vector of time-invariant coefficients. Taking the first component of $U(t)$ as 1 yields the time-varying intercept. Under model (1), the effects of the covariates $X(t) = (U^T(t), V^T)^T$ change with time while the effects of $Z(t)$ are time-invariant. We denote $\beta(t) = (\beta_1^T(t), \beta_2^T(t))^T$ as the time-varying coefficient functions for $X(t)$.

Let C denote the censoring time of the subject. The observed right-censored failure time can be denoted by $(\tilde{T}, \delta)$, where $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$. We assume that the censoring C is independent given the covariate history in the sense that the censoring does not alter the risk of failure. This assumption is described by

$$E\{d\tilde{N}(t)|X[0, t], Z[0, t], \tilde{T} \geq t\} = E\{dN^*(t)|X[0, t], Z[0, t], T \geq t\}, \text{ where}$$

$\tilde{N}(t) = I(\tilde{T} \leq t)$, $N^*(t) = I(T \leq t)$ and $X[0, t] = \{X(s), 0 \leq s \leq t\}$ and $Z[0, t] = \{Z(s), 0 \leq s \leq t\}$ are the covariate histories up to time t . Let ξ be the indicator of whether the subject is selected into the phase-two sample (determined via Bernoulli sampling as stated above). A subject with $\xi = 1$ has fully observed covariates $U(\cdot)$, V and $Z(\cdot)$ while a subject with $\xi = 0$ does not have the observed values for V . Let $\Omega = (\tilde{T}, \delta, U(\cdot), Z(\cdot), S)$ be the fully observed

part of the data, where S denotes possible auxiliary variables that have the potential to influence the sampling probabilities and may predict the phase-two covariates. We assume that the missingness pattern of V is non-informative, i.e., not dependent on the unobserved values. This assumption can be expressed as $P(\xi = 1 | V, \Omega) = P(\xi = 1 | \Omega)$, termed as the missing at random (MAR) assumption in Rubin (1976). However, the sampling probability may depend on any of the phase-one information, Ω .

Let (Ω_i, V_i, ξ_i) , $i = 1, \dots, n$, be independent identically distributed (iid) copies of (Ω, V, ξ) , where $\Omega_i = (\tilde{T}_i, \delta_i, U_i(\cdot), Z_i(\cdot), S_i)$. The observed data are $\{\Omega_i, \xi_i V_i, \xi_i, i = 1, \dots, n\}$. That is, $\{\tilde{T}_i, \delta_i, X_i(\cdot), Z_i(\cdot), S_i\}$ are observed for a subject with $\xi_i = 1$, and $\{\tilde{T}_i, \delta_i, U_i(\cdot), Z_i(\cdot), S_i\}$ are observed if $\xi_i = 0$, where $X_i(\cdot) = (U_i^T(\cdot), V_i^T)^T$. The sampling probability, $\theta_i = P(\xi_i = 1 | \Omega_i)$, is the conditional probability that V_i is observed. In particular, this sampling probability depends on the censoring indicator. Under the classical case-cohort Bernoulli sampling design, $\theta_i = 1$ if $\delta_i = 1$ (known as a case) and $\theta_i = P(\xi_i = 1 | \Omega_i, \delta_i = 0) < 1$ if $\delta_i = 0$ (known as a non-case). In this paper, we study the semiparametric additive hazards regression model (1) where the covariates can be missing for the cases as well as for the non-cases. Moreover, we assume Bernoulli two-phase sampling such that within each level of a specified phase-one discrete stratification variable defined by δ , $U(\cdot)$, and/or Z , subjects are selected for measurement of V based on a random draw from a Bernoulli distribution.

2.2 Inverse probability weighted complete-case estimation

Let $N_i(t) = I(\tilde{T}_i \leq t, \delta_i = 1)$, $Y_i(t) = I(\tilde{T}_i \geq t)$ and $\lambda_i(t) = Y_i(t)h(t|X_i(t), Z_i(t))$. Following Horvitz and Thompson (1952), the inverse probability weighting of the complete cases has been commonly used in missing data problems. Suppose that the probability of complete-case $\theta_i = P(\xi_i = 1 | \Omega_i)$ is known. Let $B(t) = \int_0^t \beta(s) ds$. Modifying the estimation equations of McKeague and Sasieni (1994) for the fully observed covariates, model (1) can be estimated based on the following inverse probability weighted estimating equations for $B(t)$ and γ :

$$\sum_{i=1}^n [q_i Y_i(t) X_i(t) W_i(t) (dN_i(t) - \lambda_i(t) dt)] = 0 \quad (2)$$

$$\sum_{i=1}^n \int_0^\tau [q_i Y_i(t) Z_i(t) W_i(t) (dN_i(t) - \lambda_i(t) dt)] = 0, \quad (3)$$

where $q_i = \xi_i / \theta_i$ and $W_i(t)$ is a weight process depending only on phase-one variables. The integrals concerned here and later are the Lebesgue integrals defined at each sample point. The integrals are random variables whose values at each sample point are the values of the Lebesgue integrals. In practice, the sampling probability θ_i is unknown. Let $\hat{\theta}_i$ be an estimate of θ_i , say, based on a parametric model such as logistic regression. The inverse probability weighting of the complete case (IPW) estimators of γ and $B(t)$ can be obtained by solving

(2) and (3) with q_i replaced by $\hat{q}_i = \xi_i / \hat{\theta}_i$. The estimator for $\beta(t)$ can be obtained by kernel smoothing the estimator of $B(t)$.

Under the classical case-cohort design, the sampling probability θ_i is 1 for all the cases and equals $\theta_i = P(\xi_i = 1 | \Omega_i, \delta_i = 0)$ for subcohort members that are not cases. Hence $q_i = \delta_i + (1 - \delta_i)\xi_i/\theta_i$. The estimator of Kulich and Lin (2000) proposed for the additive hazards regression model of Lin and Ying (1994) with case-cohort data is an IPW estimator that can be derived from (2) and (3). The recently proposed estimation method of Kang, Cai and Chambless (2013) for the additive hazards regression model of Lin and Ying (1994) applies allowing $\theta_i < 1$ for cases, and is also an application of the IPW complete-case approach.

Note from (2) and (3) that if a subject has a missing value for V_i , then the observed failure times and the values of $U_i(\cdot)$ and $Z_i(\cdot)$ from the same subject are not fully utilized except through the sampling probability θ_i . Hence the inverse probability weighting of complete cases approach is inefficient. In the following we describe an improved estimation procedure to remedy this potential inefficiency.

2.3 Augmented inverse probability weighted estimation

We adapt the idea of Robins, Rotnitzky and Zhao (1994) and propose an augmented estimation procedure for model (1) with the case-cohort/two-phase sampling data. The procedure augments the inverse probability weighting of complete cases with auxiliary predictors of the first and second moments of the missing values of phase-two covariates. The new procedure utilizes the information on the conditional distribution of the missing covariates and is thus more efficient.

2.3.1 Estimation with known θ_i and known $E(V_i | \Omega_i)$ and $E\{V_i V_i^T | \Omega_i\}$ —First we assume that the sampling probability θ_i and the conditional expectations $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ are known for those with missing values of V_i . Let $de_{i,x}(t)$ and $de_{i,z}(t)$ be the conditional expectations of $Y_i(t)X_i(t)W_i(t)\{dN_i(t) - \lambda_i(t)dt\}$ and $Y_i(t)Z_i(t)W_i(t)\{dN_i(t) - \lambda_i(t)dt\}$ given Ω_i , respectively. Since $\Omega_i = (\tilde{T}_i, \delta_i, U_i(\cdot), Z_i(\cdot), S_i)$ are observed phase-one data, the quantities $de_{i,x}(t)$ and $de_{i,z}(t)$ depend only on Ω_i and $(\beta(\cdot), \gamma)$. Following the augmentation theory of Robins, Rotnitzky and Zhao (1994), we propose the following estimating equations for $(\beta(\cdot), \gamma)$:

$$\tilde{U}_1(\beta(t), \gamma) = \sum_{i=1}^n [q_i Y_i(t) X_i(t) W_i(t) \{dN_i(t) - \lambda_i(t)dt\} + (1 - q_i) de_{i,x}(t)] = 0, \quad (4)$$

$$\tilde{U}_2(\beta(\cdot), \gamma) = \sum_{i=1}^n \int_0^\tau [q_i Y_i(t) Z_i(t) W_i(t) \{dN_i(t) - \lambda_i(t)dt\} + (1 - q_i) de_{i,z}(t)] = 0. \quad (5)$$

The contribution to equation (4) from subject i with $\xi_i = 1$ is the weighted average of the observed residual $Y_i(t)X_i(t)W_i(t)\{dN_i(t) - \lambda_i(t)dt\}$ and its conditional expectation $de_{i,x}(t) =$

$E[Y_i(t)X_i(t)W_i(t)\{dN_i(t)-\lambda_i(t)dt\}|\Omega_i]$ with weights q_i and $1 - q_i$ respectively. The first part of the contribution, $q_i Y_i(t)X_i(t)W_i(t)\{dN_i(t)-\lambda_i(t)dt\}$, represents the inverse probability weighting of complete-case. The second part, $(1 - q_i) de_{i,x}(t)$, is the augmentation to the first part with the knowledge of the conditional expectations $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ for the missing covariates. The contribution from subject i with $\xi_i = 0$ only involves the conditional expectation $de_{i,x}(t)$. A similar interpretation applies to equation (5).

Let

$$\begin{aligned} E_{zx}(t) &= n^{-1} \sum_{i=1}^n W_i(t) Y_i(t) Z_i(t) \{q_i X_i^T(t) + (1 - q_i) E(X_i^T(t) | \Omega_i)\}, \\ E_{zz}(t) &= n^{-1} \sum_{i=1}^n W_i(t) Y_i(t) Z_i(t) Z_i^T(t), \\ E_{xx}(t) &= n^{-1} \sum_{i=1}^n W_i(t) Y_i(t) \{q_i X_i(t) X_i^T(t) + (1 - q_i) E(X_i(t) X_i^T(t) | \Omega_i)\}, \\ E_{zn}(t) &= n^{-1} \sum_{i=1}^n \int_0^t W_i(s) Y_i(s) Z_i(s) dN_i(s), \\ E_{xn}(t) &= n^{-1} \sum_{i=1}^n \int_0^t W_i(s) Y_i(s) \{q_i X_i(s) + (1 - q_i) E(X_i(s) | \Omega_i)\} dN_i(s). \end{aligned} \quad (6)$$

Note that

$$E\{X_i(t) | \Omega_i\} = \begin{pmatrix} U_i(t) \\ E(V_i | \Omega_i) \end{pmatrix}, \quad (7)$$

$$E\{X_i(t) X_i^T(t) | \Omega_i\} = \begin{pmatrix} U_i(t) U_i^T(t) & U_i(t) E\{V_i^T | \Omega_i\} \\ E(V_i | \Omega_i) U_i^T(t) & E\{V_i V_i^T | \Omega_i\} \end{pmatrix}. \quad (8)$$

If the sampling probability θ_i and the conditional expectations $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ for the phase-two covariates are known, then $E_{zx}(t)$, $E_{zz}(t)$, $E_{xx}(t)$, $E_{zn}(t)$ and $E_{xn}(t)$ depend only on the observed two-phase data. The estimators for γ and $B(t)$, denoted by $\tilde{\gamma}$ and $\tilde{B}(t)$, are based on the estimating equations (4) and (5) and are given explicitly in the following theorem. The proof of Theorem 1 is given in the online Supplementary Material.

Theorem 1: Assume that the sampling probability θ_i and the conditional expectations $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ for the phase-two covariates are known. The estimators of γ and $B(t)$ obtained by solving (4) and (5) are respectively given by

$$\tilde{\gamma} = \left\{ \int_0^\tau [E_{zz}(t) - E_{zx}(t) E_{xx}^{-1}(t) E_{zx}^T(t)] dt \right\}^{-1} \left\{ E_{zn}(\tau) - \int_0^\tau E_{zx}(t) E_{xx}^{-1}(t) dE_{xn}(t) \right\}, \quad (9)$$

$$\tilde{B}(t) = \int_0^t E_{xx}^{-1}(s) dE_{xn}(s) - \int_0^t E_{xx}^{-1}(s) E_{xz}(s) ds \tilde{\gamma}. \quad (10)$$

If the sampling probability $\theta_i = 1$ for all subjects, then $q_i = 1$. The estimators $\tilde{\gamma}$ and $\tilde{B}(t)$ become the estimators of McKeague and Sasieni (1994) for the full cohort. The estimators $\tilde{\gamma}$ and $\tilde{B}(t)$ can be viewed as the expectation maximization (EM) estimators based on the estimating functions of McKeague and Sasieni (1994) for the full cohort. The estimators $\tilde{\gamma}$ and $\tilde{B}(t)$ are obtained by replacing the unobserved components V_i and $V_i V_i^T$ in the estimators of McKeague and Sasieni (1994) by their conditional expectations $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$, respectively. We notice from (6), (7) and (8) that all observations on the covariates $Z_i(\cdot)$ and $U_i(\cdot)$ are utilized even for those individuals with missing values of V_i . This estimation procedure utilizes all observed information including $(\tilde{T}_i, \delta_i, Z_i(\cdot), U_i(\cdot))$ for those subjects with unobserved covariates V_i . More efficiency can be achieved with better predictions of V_i and $V_i V_i^T$.

2.3.2 Estimation of θ_i , $E(V_i | \Omega_i)$, and $E(V_i V_i^T | \Omega_i)$ and the AIPW estimator—

Application of the estimators $\tilde{\gamma}$ and $\tilde{B}(t)$ require knowledge of the sampling probabilities θ_i and/or of the conditional expectations $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ of the phase-two covariates, which may be unknown in practice. However, these quantities can be readily estimated under the MAR assumption. Appropriate modelling of the conditional expectations $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ for the phase-two covariates can lead to improved efficiency as we will see later with further discussions and in simulations. It is convenient to estimate the terms in question with some well-established parametric methods. However, they can also be estimated with semi- or nonparametric methods.

Assume that $\pi(\Omega_i, \psi)$ is the working parametric model for the probability of complete-case, $\theta_i = P(\xi_i = 1 | \Omega_i)$, where ψ is a m -dimensional vector of parameters belonging to a compact set Θ_ψ . For example, with only case status the phase-one stratification variable, one can assume the logistic model with $\text{logit}(\pi(\Omega_i, \psi)) = \psi_1^T \Omega_i$ for those with $\delta_i = 1$ and a different logistic model with $\text{logit}(\pi(\Omega_i, \psi)) = \psi_2^T \Omega_i$ for those with $\delta_i = 0$. In this case, $\psi = (\psi_1, \psi_2)$. The parameter ψ can be estimated by the M -estimator (Huber, 1981), $\hat{\psi}$, that maximizes $\log \left[\prod_{i=1}^n \{\pi(\Omega_i, \psi)\}^{\xi_i} \{1 - \pi(\Omega_i, \psi)\}^{1-\xi_i} \right]$. Therefore, we can estimate $\theta_i = \pi(\Omega_i, \psi)$ by $\hat{\theta}_i = \pi(\Omega_i, \hat{\psi})$. One of the advantages of analysis with the M -estimators is that the estimators converge even if the true model is not a member of the assumed parametric family (van der Vaart, 1998).

We estimate $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ with the working models $\mu_1(\Omega_i, \phi_1)$ and $\mu_2(\Omega_i, \phi_2)$, respectively, where ϕ_1 and ϕ_2 are k_1 and k_2 dimensional vectors of parameters belonging to the compact sets Θ_{ϕ_1} and Θ_{ϕ_2} , respectively. For example, one can choose $\mu_1(\cdot, \phi_1)$ and $\mu_2(\cdot, \phi_2)$ as the first order or second order linear functions of the variables in Ω_i or their transformations. In this case, the parameters ϕ_1 and ϕ_2 can be estimated by the M -estimators based on the least squares regressions of V_i on Ω_i and $V_i V_i^T$ on Ω_i , respectively, based on the observations with $\xi_i = 1$ (i.e., those with observed V_i). We denote the estimators of ϕ_1 and ϕ_2 by $\hat{\phi}_1$ and $\hat{\phi}_2$, respectively.

Let $\hat{E}_{zx}(t)$, $\hat{E}_{xx}(t)$ and $\hat{E}_{xn}(t)$ be the counterparts of $E_{zx}(t)$, $E_{xx}(t)$ and $E_{xn}(t)$ defined in (6), obtained by replacing q_i with $\hat{q}_i = \xi_i / \pi(\Omega_i, \hat{\psi})$, and by replacing $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ with $\mu_1(\Omega_i, \hat{\varphi}_1)$ and $\mu_2(\Omega_i, \hat{\varphi}_2)$, respectively. Replacing $E_{zx}(t)$, $E_{xx}(t)$ and $E_{xn}(t)$ by $\hat{E}_{zx}(t)$, $\hat{E}_{xx}(t)$ and $\hat{E}_{xn}(t)$, respectively, in $\hat{\gamma}$ and $\hat{B}(t)$ defined in (9) and (10), we obtain the following augmented inverse probability weighted complete-case (AIPW) estimators for γ and $B(t)$:

$$\hat{\gamma} = \left\{ \int_0^\tau [E_{zz}(t) - \hat{E}_{zx}(t) \hat{E}_{xx}^{-1}(t) \hat{E}_{zx}^T(t)] dt \right\}^{-1} \left\{ E_{zn}(\tau) - \int_0^\tau \hat{E}_{zx}(t) \hat{E}_{xx}^{-1}(t) d\hat{E}_{xn}(t) \right\}, \quad (11)$$

$$\hat{B}(t) = \int_0^t \hat{E}_{xx}^{-1}(s) d\hat{E}_{xn}(s) - \int_0^t \hat{E}_{xx}^{-1}(s) \hat{E}_{xz}(s) ds \hat{\gamma}. \quad (12)$$

The estimators $\hat{B}(t)$ for $\beta(t)$ can be obtained by using the kernel smoothing for the estimator $\hat{B}(t)$.

3 Asymptotic properties

This section investigates the asymptotic properties of the proposed estimators. Since the weights q_i and $\hat{q}_i$ are not generally predictable, the asymptotic properties are investigated using the empirical process theory which does not require predictability. Suppose that $\beta_0(t)$ and γ_0 are the true values of $\beta(t)$ and γ under model (1). Let $B_0(t) = \int_0^t \beta_0(s) ds$. Let $e_{zx}(t) = E\{W_i(t)Y_i(t)Z_i(t)X_i^T(t)\}$, $e_{xz}(t) = e_{zx}^T(t)$, $e_{zz}(t) = E\{W_i(t)Y_i(t)Z_i(t)Z_i^T(t)\}$, $e_{xx}(t) = E\{W_i(t)Y_i(t)X_i(t)X_i^T(t)\}$, $e_{zn}(t) = E\{\int_0^t W_i(s)Y_i(s)Z_i(s)dN_i(s)\}$, and $e_{xn}(t) = E\{\int_0^t W_i(s)Y_i(s)X_i(s)dN_i(s)\}$. Let $A = \int_0^\tau \{e_{zz}(t) - e_{zx}(t)e_{xx}^{-1}(t)e_{zx}^T(t)\} dt$. The regularity conditions for the asymptotic results are stated in Condition A given in the Appendix (which includes the assumption of Bernoulli two-phase sampling).

Suppose that $\pi(\Omega_i, \psi)$ is the working model for $P(\xi_i = 1 | \Omega_i)$, and $\mu_1(\Omega_i, \varphi_1)$ and $\mu_2(\Omega_i, \varphi_2)$ are working models for $E(V_i | \Omega_i)$ and $E\{V_i V_i^T | \Omega_i\}$, respectively. The asymptotic results of the M -estimators are established in Theorem 5.7 and Theorem 5.2 in van der Vaart (1998). The conditions of Theorem 5.7 and Theorem 5.21 in van der Vaart (1998) can be easily checked when $\text{logit}(\pi(\Omega_i, \psi)) = \psi^T \Omega_i$ and $\hat{\psi}$ is the maximizer of the log likelihood function under this working model. The conditions can also be easily checked when $\mu_1(\Omega_i, \varphi_1)$ and $\mu_2(\Omega_i, \varphi_2)$ are linear regression models and $\hat{\varphi}_1$ and $\hat{\varphi}_2$ are the ordinary least squares estimators. Let ψ^* , φ_1^* and φ_2^* be the limits of the M -estimators $\hat{\psi}$, $\hat{\varphi}_1$ and $\hat{\varphi}_2$, respectively. Let $q_1^* = \xi_i / \pi(\Omega_i, \psi^*)$ and let $E^*\{V_i / \Omega_i\} = \mu_1(\Omega_i, \varphi_1^*)$ and $E^*\{V_i V_i^T | \Omega_i\} = \mu_2(\Omega_i, \varphi_2^*)$. Let $E^*(X_i(t) | \Omega_i)$ and $E^*(X_i(t)X_i^T(t) | \Omega_i)$ correspond to $E(X_i(t) | \Omega_i)$ and $E(X_i(t)X_i^T(t) | \Omega_i)$ defined in (7) and (8) with $E(V_i | \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ replaced by $E^*\{V_i | \Omega_i\}$ and $E^*\{V_i V_i^T | \Omega_i\}$, respectively. Replacing q_i , $E(X_i(t) | \Omega_i)$ and $E(X_i(t)X_i^T(t) | \Omega_i)$ by q_i^* , $E^*(X_i(t) | \Omega_i)$ and

$E^*(X_i(t)X_i^T(t)/\Omega_i)$ in (6) to get E_{zx}^* , $E_{xx}^*(t)$ and $E_{xn}^*(t)$ in place of $E_{zx}(t)$, $E_{xx}(t)$ and $E_{xn}(t)$, respectively.

The following theorems show that the AIPW estimators possess the double robustness property, wherein the AIPW estimators $\hat{\gamma}$ and $\hat{B}(t)$ are asymptotically unbiased if the sampling probability $P(\xi_i = 1|\Omega_i)$ and/or both the conditional expectations $E(V_i|\Omega_i)$ and $E(V_iV_i^T|\Omega_i)$ are modeled correctly. The asymptotic weak convergence results presented in Theorem 2 and 3 for $n^{1/2}(\hat{\gamma} - \gamma_0)$ and $G_n(t) = n^{1/2}(\hat{B}(t) - B_0(t))$ over $t \in [0, \tau]$ are useful for construction of a confidence interval for γ , and confidence bands for $B(t)$. The asymptotic weak convergence result in Theorem 3 is also useful for developing hypothesis testing procedures for $\beta(t)$. The proofs of Theorem 2 and 3 are placed in the online Supplementary Material.

Let

$$\eta_{z,i} = \int_0^\tau W_i(t)Y_i(t)[Z_i(t)dN_i(t) - Z_i(t)\{q_i^*X_i^T(t) + (1 - q_i^*)E^*(X_i^T(t)|\Omega_i)\}dB_0(t) - Z_i(t)Z_i^T(t)\gamma_0 dt]. \quad (13)$$

$$\begin{aligned} \eta_{x,i} = & \int_0^\tau e_{zx}(t)e_{xx}^{-1}(t)W_i(t)Y_i(t)[\{q_i^*X_i(t) + (1 - q_i^*)E^*(X_i(t)|\Omega_i)\}dN_i(t) \\ & - \{q_i^*X_i(t)X_i^T(t) + (1 - q_i^*)E^*\{X_i(t)(X_i^T(t)|\Omega_i)\}\}dB_0(t) \\ & - \{q_i^*X_i(t) + (1 - q_i^*)E^*(X_i(t)|\Omega_i)\}Z_i^T(t)\gamma_0 dt]. \end{aligned} \quad (14)$$

Theorem 2

Assuming Condition A, if the sampling probability $P(\xi_i = 1|\Omega_i) = \pi(\Omega_i, \psi)$, and/or both the conditional expectations $E(V_i|\Omega_i) = \mu_1(\Omega_i, \varphi_1)$ and $E(V_iV_i^T|\Omega_i) = \mu_2(\Omega_i, \varphi_2)$ are correctly specified, then the following assertions hold:

- a. $\hat{\gamma} \xrightarrow{P} \gamma_0$;
- b. $n^{1/2}(\hat{\gamma} - \gamma_0) \xrightarrow{\mathcal{D}} N(0, W)$ as $n \rightarrow \infty$, where $W = A^{-1}\Sigma A^{-1}$, $\Sigma = E\{(\eta_{z,i} - \eta_{x,i} + \varepsilon_{\Phi,i} + \varepsilon_{\Psi,i})(\eta_{z,i} - \eta_{x,i} + \varepsilon_{\Phi,i} + \varepsilon_{\Psi,i})^T\}$, and $\varepsilon_{\Phi,i}$ and $\varepsilon_{\Psi,i}$ are given in (S.46) and (S.47) in the Supplementary Material;
- c. In addition, if $P(\xi_i = 1|\Omega_i) = \pi(\Omega_i, \psi)$ is correctly specified then $\varepsilon_{\Phi,i} = 0$; and if $E(V_i|\Omega_i) = \mu_1(\Omega_i, \varphi_1)$ and $E(V_iV_i^T|\Omega_i) = \mu_2(\Omega_i, \varphi_2)$ are modelled correctly, then $\varepsilon_{\Psi,i} = 0$.

The matrix A can be consistently estimated by

$$\hat{A} = n^{-1} \sum_{i=1}^n \int_0^\tau \{\hat{E}_{zz}(t) - \hat{E}_{zx}(t)\hat{E}_{xx}^{-1}(t)\hat{E}_{zx}^T(t)\}dt,$$

and Σ can be consistently estimated by

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n (\hat{\eta}_{z,i} - \hat{\eta}_{x,i} + \hat{\varepsilon}_{\Phi,i} + \hat{\varepsilon}_{\Psi,i})(\hat{\eta}_{z,i} - \hat{\eta}_{x,i} + \hat{\varepsilon}_{\Phi,i} + \hat{\varepsilon}_{\Psi,i})^T,$$

where $\hat{\eta}_{z,i}$ and $\hat{\eta}_{x,i}$ are the empirical counterparts of $\eta_{z,i}$ and $\eta_{x,i}$ obtained by replacing γ_0 , $B_0(t)$, q_i^* , $e_{zx}(t)$ and $e_{xx}(t)$ with $\hat{\gamma}$, $\hat{B}(t)$, $\hat{q}_i = \xi_i / \pi(\Omega_i, \hat{\psi})$, $\hat{E}_{zx}(t)$ and $\hat{E}_{xx}(t)$, respectively, and by replacing the unknown quantities $E^*(V_i | \Omega_i)$ and $E^*(V_i V_i^T | \Omega_i)$ in $E^*(X_i(t) | \Omega_i)$ and $E^*(X_i(t) X_i^T(t) | \Omega_i)$ with $\mu_1(\Omega_i, \hat{\varphi}_1)$ and $\mu_2(\Omega_i, \hat{\varphi}_2)$, respectively. Similarly, $\hat{\varepsilon}_{\Phi,i}$ and $\hat{\varepsilon}_{\Psi,i}$ are the empirical counterparts of $\varepsilon_{\Phi,i}$ and $\varepsilon_{\Psi,i}$ given in (S.46) and (S.47) in the Supplementary Material.

Let

$$\begin{aligned} \zeta_i(t) = & \int_0^t e_{xx}^{-1}(s) W_i(s) Y_i(s) [\{q_i^* X_i(s) + (1 - q_i^*) E^*(X_i(s) | \Omega_i)\} dN_i(s) \\ & - \{q_i^* X_i(s) X_i^T(s) + (1 - q_i^*) E^*\{X_i(s) X_i^T(s) | \Omega_i\}\} dB_0(s) \\ & - \{q_i^* X_i(s) + (1 - q_i^*) E^*(X_i(s) | \Omega_i)\} Z_i^T(s) \gamma_0 ds]. \end{aligned} \quad (15)$$

Theorem 3

Assuming Condition A, if the sampling probability $P(\xi_i = 1 | \Omega_i) = \pi(\Omega_i, \psi)$, and/or both the conditional expectations $E(V_i | \Omega_i) = \mu_1(\Omega_i, \varphi_1)$ and $E(V_i V_i^T | \Omega_i) = \mu_2(\Omega_i, \varphi_2)$ are correctly specified, then the following assertions hold:

- a. $\sup_{t \in [0, \tau]} |\hat{B}(t) - B_0(t)| \xrightarrow{P} 0$ as $n \rightarrow \infty$;
- b. The process $G_n(t) = n^{1/2}(\hat{B}(t) - B_0(t))$ converges weakly to a zero-mean Gaussian process $G(t)$ on $[0, \tau]$ with the covariance matrix

$$\Sigma_G(t) = \{\zeta_i(t) - \int_0^t e_{xx}^{-1}(s) e_{xz}(s) ds A^{-1}(\eta_{z,i} - \eta_{x,i}) + v_{\Phi,i}(t) + v_{\Psi,i}(t)\}^{\otimes 2},$$

where $\eta_{z,i}$ and $\eta_{x,i}$ are defined in (13) and (14), respectively, $\zeta_i(t)$ is defined in (15), and the expressions for $v_{\Phi,i}(t)$ and $v_{\Psi,i}(t)$ are given in (S.59) and (S.60) in the Supplementary Material;

- c. In addition, if $P(\xi_i = 1 | \Omega_i) = \pi(\Omega_i, \psi)$ is correctly specified then $v_{\Psi,i}(t) = 0$ and $\varepsilon_{\Psi,i} = 0$; and if $E(V_i | \Omega_i) = \mu_1(\Omega_i, \varphi_1)$ and $E(V_i V_i^T | \Omega_i) = \mu_2(\Omega_i, \varphi_2)$ are modeled correctly, then $v_{\Phi,i}(t) = 0$ and $\varepsilon_{\Phi,i} = 0$.

Under Theorem 3, if the sampling probability $P(\xi_i = 1 | \Omega_i) = \pi(\Omega_i, \psi)$ and both the conditional expectations $E(V_i | \Omega_i) = \mu_1(\Omega_i, \varphi_1)$ and $E(V_i V_i^T | \Omega_i) = \mu_2(\Omega_i, \varphi_2)$ are correctly specified, then $v_{\Psi,i}(t) = 0$, $v_{\Phi,i}(t) = 0$, $\varepsilon_{\Psi,i} = 0$ and $\varepsilon_{\Phi,i} = 0$. The asymptotic covariance matrix of $G_n(t)$ can be estimated consistently by

$$\sum_G \hat{\Sigma}_G(t) = n^{-1} \sum_{i=1}^n \left\{ \hat{\zeta}_i(t) - \int_0^t \hat{E}_{xx}^{-1}(s) \hat{E}_{xz}(s) ds \hat{A}^{-1}(\hat{\eta}_{z,i} - \hat{\eta}_{x,i}) + \hat{v}_{\Phi,i}(t) + \hat{v}_{\Psi,i}(t) \right\}^{\otimes 2}, \quad (16)$$

where $\hat{\zeta}_i(t)$, $\hat{\eta}_{z,i}$ and $\hat{\eta}_{x,i}$ are the empirical counterparts of $\zeta_i(t)$, $\eta_{z,i}$ and $\eta_{x,i}$ obtained by replacing q_i^* with $\hat{q}_i = \xi_i / \pi(\Omega_i, \hat{\psi})$, and by replacing $E^*(V_i | \Omega_i)$ and $E^*(V_i / V_i^T | \Omega_i)$ with $\mu_1(\Omega_i, \hat{\varphi}_1)$ and $\mu_2(\Omega_i, \hat{\varphi}_2)$, respectively. Here $\hat{v}_{\Phi,i}(t)$ and $\hat{v}_{\Psi,i}(t)$ are the empirical counterparts of $v_{\Phi,i}(t)$ and $v_{\Psi,i}(t)$, respectively.

4 Simulation study

This section presents a simulation study for evaluating the finite-sample properties of the proposed methods. Let $Z = (Z_1, Z_2)^T$ be the phase-one covariates, where Z_1 is a Bernoulli random variable with $P(Z_1 = 1) = 0.5$ and Z_2 is a uniform random variable on $(0, 1)$. Let V be a phase-two covariate following the uniform distribution on $(0, 1)$. The random variables Z_1 , Z_2 and V are independent. We consider the following hazard regression model for the failure time T :

$$h(t|X, Z) = \beta_1(t) + \beta_2(t)V + \gamma_1 Z_1 + \gamma_2 Z_2, \quad 0 \leq t \leq \tau, \quad (17)$$

where $\beta_1(t) = 0.1(t+1)$, $\beta_2(t) = -0.1(t-1)$, $\gamma_1 = 0.02$ and $\gamma_2 = -0.08$, and where $\tau = 1.5$.

Let S be an auxiliary variable for the phase-two covariate V with the relationship $S = (V + \theta\zeta)/(1+\theta)$, where ζ is a uniform random variable on $(0, 1)$ and θ is a parameter dictating the association between S and V . The values $\theta = 1.7321, 0.8819, 0.3287$ yield correlation coefficients $\rho = 0.50, 0.75, 0.95$ between V and S , respectively. The AIPW estimators with $\rho = 0.50, 0.75, 0.95$ are denoted by AIPW-R50, AIPW-R75 and AIPW95, respectively. Let C^* follow an exponential distribution with mean equal to 10. The censoring time is taken as $C = C^* \wedge \tau$, yielding about 80% censoring of the failure time. Let $\tilde{T} = T \wedge C$ and $\delta = I(T \leq C)$.

The phase-one data for subject i are $\Omega_i = (\delta_i, \tilde{T}_i, Z_{1i}, Z_{2i}, S_i)$. We consider two scenarios of the two-phase sampling. The first is the classical case-cohort design where the phase-two covariate V is sampled for all cases and for a selected subset of non-cases. For the non-cases, we assume that the sampling probability $\alpha_i = P(\xi_i = 1 | \Omega_i, \delta_i = 0)$ follows a logistic regression model $\text{logit}(\alpha_i) = \pi_0 + \pi_1(1 + \theta)S_i + \pi_2 Z_{1i} + \pi_3 Z_{2i}$ based on phase-one data. Three different sampling probabilities are considered. The choices of $(\pi_0, \pi_1, \pi_2, \pi_3) = (-1, 0.5, -0.5, -0.5), (0, 0.5, -0.5, -0.5)$ and $(1, 0.5, -0.5, -0.5)$ correspond to the average sampling probabilities $p_0 = 0.1, 0.25$ and 0.5 for the non-cases, respectively. We use the linear model with response variable V_i and predictors S_i, Z_{1i}, Z_{2i} and $\log(\tilde{T}_i)$ to estimate $E(V_i | \Omega_i)$, based on the observations that are non-cases and with observed values of V_i . The linear model with the response variable V_i^2 and the predictors S_i, Z_{1i}, Z_{2i} and $\log(\tilde{T}_i)$ is also used to estimate $E(V_i^2 | \Omega_i)$.

The second two-phase sampling scenario allows V_i to be missing for cases as well as for non-cases. Let $\vartheta_i = P(\xi_i = 1 | \Omega_i, \delta_i = 1)$ and $\alpha_i = P(\xi_i = 1 | \Omega_i, \delta_i = 0)$. Allowing differentiation in the sampling probabilities for cases and non-cases, ϑ_i and α_i are modeled with separate logistic regression models using the predictors S_i, Z_{1i}, Z_{2i} . Our simulation experiment considers the average sampling probabilities $p_1 = 0.5$ for the cases, and $p_0 = 0.1, 0.25$ and 0.5 for the non-cases. As with the first set-up, we use linear models with the predictors S_i, Z_{1i}, Z_{2i} and $\log(\tilde{T}_i)$ to estimate $E(V_i | \Omega_i, \delta_i = 1)$ and $E(V_i^2 | \Omega_i, \delta_i = 1)$ based on the observations that are cases and with observed values of V_i . Similarly, linear models with the predictors S_i, Z_{1i}, Z_{2i} and $\log(\tilde{T}_i)$ are used to estimate $E(V_i | \Omega_i, \delta_i = 0)$ and $E(V_i^2 | \Omega_i, \delta_i = 0)$ based on the observations that are non-cases and with observed values of V_i . The variance estimators are calculated as if the linear models are the correct model specifications.

Tables 1 to 3 present our simulation results for $n = 600, 750, 1000$, and for the average sampling probabilities $p_1 = 1.0$ and 0.5 for the cases, and $p_0 = 0.1, 0.25$ and 0.5 for the non-cases. The weight function $W_i(t) = 1$ is used in the simulations. Each entry of Table 1 to Table 3 is based on 1000 simulation runs. Table 1 summarizes the bias (Bias), the empirical standard error (SSE), the average of the estimated standard error (ESE), and the empirical coverage probability (CP) of 95% confidence intervals of the AIPW-R50 estimator for γ . Table 1 shows that the AIPW-R50 estimator for γ performs well under both scenarios of two-phase sampling with the combinations of the average sampling probabilities $p_1 = 1.0$ and 0.5 and $p_0 = 0.1, 0.25$ and 0.5 . The biases are small for the sample sizes $n = 600, 750$ and 1000 . The averages of the estimated standard errors are very close to the empirical standard errors and the coverage probabilities are very close to the 0.95 nominal level, indicating appropriateness of the proposed estimator for the variance of $\hat{\gamma}$.

We also compare the performance of the proposed AIPW estimator with the IPW estimator described in Section 2.3 and the complete-case (CC) estimator obtained by deleting subjects with missing values of V_i . As a gold standard, we present the estimation results for the full cohort where all the values of V_i are fully observed, which is denoted by Full. Table 2 compares the bias of these estimators for estimating γ . Table 3 compares the relative efficiency (REE) of the AIPW-R50, IPW and CC estimators relative to the Full estimators, where REE for each of the estimators is defined as SSE of the Full estimator divided by SSE of the corresponding estimator. Table 2 shows that the biases of both the IPW and AIPW-R50 estimators for γ are very small at a level comparable to the Full estimator, as if all the values of the covariate V_i were observed. The complete-case estimator (CC) yields larger biases. Table 3 shows that the relative efficiency of the AIPW-R50 estimator for γ is larger than the relative efficiency of the IPW estimator, which is in turn larger than that of the complete-case estimator, in each case. The efficiency of the AIPW estimator gains the most over the IPW estimator when the sampling probability for the non-cases is small, e.g., $p_0 = 0.1$. This is because the AIPW estimator can more efficiently utilize information on the failure time and other fully observable covariates for individuals with missing covariate(s).

Figure 1 presents the comparison of the estimators for the nonparametric component $B_2(t) = \int_0^t \beta_2(s) ds$ for $n = 600$ and with average sampling probabilities $p_1 = 0.5$ for the cases and $p_0 = 0.1$ for the non-cases. Figure 1(a) plots the biases of the estimators for $B_2(t)$ for $0 <$

$t = 1.5$ for each of the estimators, AIPW-R50, AIPW-R75, AIPW-R95, IPW, CC and Full. Figure 1(b) plots the relative efficiencies of these estimators. The coverage probabilities of the 95% pointwise confidence intervals for $B_2(t)$ for each t using the AIPW estimators are given in Figure 1(c). The coverage probability of the IPW estimator is not presented because the estimation of its standard error is much more complicated due to lack of orthogonality, that is, one has to take into consideration the estimation variance for the sampling probability models. Figure 1(a) shows that the estimation biases of both the IPW and AIPW-R50, AIPW-R75, AIPW-R95 estimator for $B_2(t)$ are very small, comparable to the estimator as if all the values of the covariate V_i were observed. The bias of the complete-case estimator of $B_2(t)$ is much larger. Figure 1(b) shows that the relative efficiency of AIPW-R50 estimator for $B_2(t)$ is slightly larger than that of the IPW estimator when the correlation between the auxiliary variable S and the phase-two covariate V is low with $\rho = 0.5$. But the relative efficiency of the AIPW estimator improves significantly for AIPW-R75, AIPW-R95 which corresponds to $\rho = 0.75$ and $\rho = 0.95$, respectively. The pointwise coverage probabilities for $B_2(t)$ shown in Figure 1(c) for the AIPW-R50, AIPW-R75, AIPW-R95 estimators are close to the 0.95 nominal level roughly in the range from 0.92 to 0.95.

The simulation results on the relative efficiency for $n = 600$ with $p_1 = 0.5$ and $p_0 = 0.1$ indicate that the benefit of the AIPW estimator over the IPW estimator is greater for estimating the effects (γ) of those covariates that are fully observed and less for those covariates with missing values. The relative efficiency for estimating γ is 0.94 for AIPW-R50 and 0.73 for IPW, while the relative efficiency for estimating $B_2(t)$ is less than 0.5 for both AIPW-R50 and IPW. This is because the information for the observable covariates for individuals with missing values on other covariate(s) are fully used in the AIPW estimator instead of thrown out as with the IPW estimator (except through the modeling of the sampling probability). However, the relative efficiency of the AIPW estimator for $B_2(t)$ increases greatly as the correlation between the auxiliary variable S and the phase-two covariate V increases.

5 Application

From 1998 to 2003 VaxGen Inc. conducted the first HIV vaccine efficacy trial known as the VaxGen 004 trial (Flynn et al., 2005). This analysis focuses on a dataset for 1168 white vaccinated men of the VaxGen trial who received the first 4 study injections (at the visits at months 0, 1, 6, 12) and were HIV negative (uninfected) at the Month 12 visit. The HIV-1 infection status for those who were HIV-1 negative at the Month 12 visit were determined based on HIV-1 tests at visits at months 18, 24, 30, 36. The infection time is the number of days between the Month 12 visit and the estimated date of HIV-1 infection, as defined in Flynn et al. (2005). For subjects not infected during the trial, the right-censoring time is the number of days between the Month 12 visit and the date of last contact. A behavioral risk score is measured for all subjects at study entry, and is highly predictive of whether a subject acquires HIV infection. The behavioral risk score takes integer values ranging from 0 to 7 with larger values indicating higher risk. The procedure for defining the risk score is described in Flynn et al. (2005).

Three immune responses to vaccination were measured at the Month 12.5 visit in the trial: Neut_SF162_IC50, Neut_MN_IC50 and CD4i_HIV2_IC50. Here we use the short labels Neut_SF162, Neut_MN and CD4i_HIV2 for the three responses, respectively, for convenience. The immune responses Neut_SF162 and Neut_MN were measured for all 1168 subjects, and hence are phase-one covariates. The immune response CD4i_HIV2 is a phase-two covariate measured for 52 of the 95 subjects who became HIV-1 positive during the trial (cases) and for 39 of those (1073) who remained HIV-1 negative (non-cases), where the 52 and 39 subjects were sampled into phase two using Bernoulli sampling with separate success probabilities. Both immune responses Neut_SF162 and Neut_MN were correlated with CD4i_HIV2, with Spearman rank correlation 0.60 between Neut_SF162 and CD4i_HIV2 and 0.34 between Neut_MN and CD4i_HIV2.

The developed method is applied to evaluate the effect of the immune response CD4i_HIV2 at the Month 12.5 visit on the subsequent time to the estimated date of HIV-1 infection, adjusting for the behavioral risk score. Preliminary explorations using the complete-case analysis indicates that the proportional hazards assumption is violated. We re-scale the infection time from days to months and consider the following semiparametric additive hazards regression model

$$h(t|X, Z) = \beta_1(t) + \beta_2(t)V + \gamma Z, \quad (18)$$

for $0 \leq t \leq 2$, where V is the immune response CD4i_HIV2 at the Month 12.5 visit, and Z is the baseline behavioral risk score.

We consider Neut_MN and Neut_SF162 as the auxiliary variables denoted by S_1 and S_2 , respectively. Two logistic regression models are used to model the sampling probabilities for the cases and for the non-cases separately. The estimated sampling probabilities $\hat{\vartheta}_i$ for the cases are given by $\text{logit}(\hat{\vartheta}_i) = -1.4135 + 0.2443S_{1i} + 0.3178S_{2i} + 0.1581Z_i$. The estimated sampling probabilities $\hat{\alpha}_i$ for the non-cases are given by $\text{logit}(\hat{\alpha}_i) = -3.4089 - 0.6409S_{1i} + 1.097S_{2i} - 0.1717Z_i$. The weights q_i are estimated by $\hat{q}_i = \delta_i \xi_i / \hat{\vartheta}_i + (1 - \delta_i) \xi_i / \hat{\alpha}_i$.

Let $Q_i = (1, S_{1i}, S_{2i}, Z_i, \log(\tilde{T}_i))$. The linear models with the predictors Q_i are used to estimate $E(V_i | \Omega_i, \delta_i = 1)$ and $E(V_i^2 | \Omega_i, \delta_i = 1)$ based on the observations that are cases and with observed V_i 's. The term $E(V_i | \Omega_i, \delta_i = 1)$ is estimated by $\hat{\eta}^T Q_i$ where $\hat{\eta} = (3.8649, -0.2140, 1.7379, 0.0623, 0.3264)^T$, and $E(V_i^2 | \Omega_i, \delta_i = 1)$ is estimated by $\hat{\nu}^T Q_i$ where $\hat{\nu} = (7.5491, -3.2593, 23.6284, 1.015, 4.2172)^T$. Similarly, the linear models with the predictors Q_i are used to estimate $E(V_i | \Omega_i, \delta_i = 0)$ and $E(V_i^2 | \Omega_i, \delta_i = 0)$ based on the observations that are non-cases and with observed V_i 's. The term $E(V_i | \Omega_i, \delta_i = 0)$ is estimated by $\hat{\varphi}^T Q_i$ where $\hat{\varphi} = (3.6937, 0.1193, 1.4682, 0.0225, -0.0653)^T$, and $E(V_i^2 | \Omega_i, \delta_i = 0)$ is estimated by $\hat{\kappa}^T Q_i$ where $\hat{\kappa} = (4.1394, 1.9896, 19.5622, 0.2734, -0.2856)^T$.

Our method with $W_{\lambda}(t) = 1$ gives the estimated effect of the behavioral risk score $\hat{\gamma} = 0.03289$ with standard error of 0.0061, yielding p -value < 0.00001 for testing $\gamma = 0$, confirming that the risk score has a strong and significant effect on the hazard of HIV-1 infection. The estimates of the cumulative coefficients $B_1(t) = \int_0^t \beta_1(s) ds$ and $B_2(t) = \int_0^t \beta_2(s) ds$ with 95% pointwise confidence bands are plotted in Figure 2 along with their 95% pointwise confidence bands. Figure 2(b) indicates that an increase in the immune response level of CD4i_HIV2 is associated with a reduced hazard rate of infection, albeit weakly, shortly after receiving the first 4 study injections. This potential predictive utility peaks at about 5 months after the fourth injection and wanes to zero by about 16 months.

6 Discussion

This paper studies the statistical modelling of case-cohort/two-phase sampling data using the semiparametric additive hazards model, allowing the covariates of interest to be missing for cases as well as for non-cases. The semiparametric additive hazards model is an important alternative to the proportional hazards model, where the risk factors contribute additively to the overall risk of a subject. We considered a more flexible form of the additive model that allows the effects of some covariates to be time varying while specifying the effects of others to be constant.

Most existing statistical methods for analyzing case-cohort data are based on modifications of the full data partial likelihood score function for the Cox proportional hazards model, or more recently on the full likelihood, by weighting the contributions from the cases and subcohort members with the inverses of true or estimated sampling probabilities. We explored two estimation approaches of the semiparametric additive hazards model, the first based on the widely adopted inverse probability weighting of complete cases and the other based on augmented inverse probability weighting that incorporates information in phase-one covariates about the missing phase-two covariates. We demonstrated that the IPW estimator is inefficient, by virtue of not utilizing observed information including failure time and other observed covariates from subjects with missing phase-two covariates, and that the AIPW estimator can utilize the available information more efficiently. The simulation results show that the AIPW estimator perform adequately in terms of bias and standard error estimation. The relative efficiency of the AIPW estimator increases as the correlation between the auxiliary variable S_i and the phase-two covariate V_i increases.

As a doubly robust estimator, asymptotically consistent implementation of the AIPW estimator relies on correct modelling of either the sampling probability θ_i and/or the conditional expectations $E(V_i \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ of the missing covariates. Consequently, for case-cohort/two-phase sampling designs with known q_i or consistently estimated q_i , the AIPW estimators $\hat{\gamma}$ and $\hat{B}(t)$ are asymptotically unbiased regardless of whether or not $E(V_i \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ can be modelled correctly. Similarly, if $E(V_i \Omega_i)$ and $E(V_i V_i^T | \Omega_i)$ are both modeled correctly, then the AIPW estimators are asymptotically unbiased if inconsistent estimators of the θ_i are used.

If the covariates are missing by design, then the sampling probability θ_i can be predetermined or consistently estimated with the observed proportions for each subject. In a more general situation a logistic regression model can be used to estimate θ_i . An appropriate modelling of the conditional expectations $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ for the missing covariates can lead to improved efficiency as we showed in the simulation study. Logistic regression models have been used for modeling the sampling probabilities in many existing works under similar situations, cf., Gao and Tsiatis (2005), and Sun and Gilbert (2012).

We propose to use first-order linear regressions to estimate $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ for the missing covariates. Our extensive simulations showed that this approach works well. In fact, in the additional simulations that we conducted for the simulation models not given in Section 4, we found that the second-order linear regression does not improve upon the first-order linear regression for the AIPW estimator in terms of estimation bias or standard error. We chose to estimate these quantities with some well established parametric methods although they can also be estimated with more flexible semi- or nonparametric methods. The derivations of the asymptotic results under semi- or nonparametric models for $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ would be much more complicated unless the conditional expectations $E(V_i|\Omega_i)$ and $E(V_i V_i^T|\Omega_i)$ only depend on some categorical variables.

The developed method requires the weight process $W_\lambda(t)$ depending only on phase-one variables. $W_\lambda(t)$ can be selected to put more weight on early or later failure times such that the variance of the estimator is minimized to improve efficiency. Our simulations and the HIV example use $W_\lambda(t) = 1$ which gives uniform weight across failure times. The choice of $W_\lambda(t)$ is a difficult issue. Theoretically speaking, one should choose $W_\lambda(t)$ such that variance of estimator is minimized. This is obviously an important topic that needs further investigation.

The newly developed method was applied to analyze a data set from the VaxGen 004 trial, to examine the effect of the immune response CD4i_HIV2_IC50 on the risk of HIV infection in vaccinated white men who received the first four study injections. Because the immune responses are expensive and labor-intensive to measure, they are only collected from a subset of the study subjects, and, while it is typically efficient to measure the responses from all infected subjects (cases), they may be unavailable for some infected subjects, as was the case for VaxGen 004. Our analyses showed that the behavioral risk score has a strong significant effect on the hazard of the HIV infection. The analysis also indicates that an increase in the immune response level of CD4i_HIV2_IC50 measured at the Month 12.5 visit is associated with a reduced hazard rate of infection within six months after receiving the first 4 injections. However, its potential association with infection risk disappears afterwards, suggesting that its value as a predictive biomarker may be restricted to infection proximal to the fourth injection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Richard Wyatt, David Montefiori and John Mascola for measuring the immune response data for the VaxGen 004 trial. The authors thank the reviewers for their constructive comments that have improved the paper. This research was partially supported by NSF grants DMS-1208978, DMS-0905777 and DMS-1513072, NIH grant 2 R37 AI054165, and the Reassignment of Duties fund provided by the University of North Carolina at Charlotte.

7 Appendix

Let $f(t)$ be a function $[a, b] \rightarrow \mathcal{R}$. Given any finite partition $\Gamma = \{a = t_0 < \dots < t_K = b\}$ of $[a, b]$, the variation of f over $[a, b]$ is

$V[f; a, b] = \sup \left\{ \sum_{k=1}^K |f(t_k) - f(t_{k-1})| : \Gamma \text{ is a partition of } [a, b] \right\}$. The function f has bounded variation on $[a, b]$ if $V[f; a, b] < \infty$. A vector f of functions has bounded variation if each component of f has bounded variation, and in this case, $V[f; a, b]$ is the vector of the variations of the component functions.

We assume the following regularity conditions throughout the paper:

Condition A

- A1** The processes $X_\lambda(t)$, $Z_\lambda(t)$ and $W_\lambda(t)$, $0 \leq t \leq \tau$, have bounded second moments, their sample paths are left continuous and of bounded variation. The variations of the processes $U_\lambda(\cdot)$, $Z_\lambda(\cdot)$ and $W_\lambda(\cdot)$ satisfy the conditions $(E\{\|\bigvee U_\lambda; s, t\|^2\})^{1/2} \leq C(t-s)^\alpha$, $(E\{\|\bigvee Z_\lambda; s, t\|^2\})^{1/2} \leq C(t-s)^\alpha$, and $(E\{\|\bigvee W_\lambda; s, t\|^2\})^{1/2} \leq C(t-s)^\alpha$, for $s, t \in [0, \tau]$, where $\alpha > 0$ and $C > 0$ are constants, and $\|\cdot\|$ is the Euclidean norm.
- A2** The $\beta(t)$, $e_{zz}(t)$, $e_{xx}(t)$ and $e_{xz}(t)$ are twice differentiable on $[0, \tau]$, and $e_{xx}(t)$ is a nonsingular matrix, and $e_{xx}^{-1}(t)$ is bounded over $0 \leq t \leq \tau$.
- A3** The matrix A is positive definite.
- A4** $W_\lambda(t)$ is a weight process depending only on phase-one variables, $W_i(t) \xrightarrow{P} w_i(t)$ uniformly in $t \in [0, \tau]$ and $1 \leq i \leq n$, and $w_\lambda(t)$ is differentiable with uniformly bounded derivative.
- A5** The censoring is independent in the sense that the censoring does not alter the risk of failure. This assumption is described by $E\{d\tilde{N}_i(t) | X_i[0, t], Z_i[0, t], \tilde{T}_i \geq t\} = E\{dN_i^*(t) | X_i[0, t], Z_i[0, t], T_i \geq t\}$, where $\tilde{N}_i(t) = I(\tilde{T}_i \leq t)$, $N_i^*(t) = I(T_i \leq t)$, and $X_\lambda[0, t] = \{X_\lambda(s), 0 \leq s \leq t\}$ and $Z_\lambda[0, t] = \{Z_\lambda(s), 0 \leq s \leq t\}$ are the covariate histories up to time t .
- A6** The phase-two covariate V_i is missing at random (MAR), i.e., $P(\xi_i = 1 | V_i, \Omega_i) = P(\xi_i = 1 | \Omega_i)$.
- A7** The function $\pi(\Omega_i, \psi)$ is twice differentiable with respect to ψ the compact set Θ_ψ , $\pi'_\psi(\Omega_i, \psi) = \partial \pi(\Omega_i, \psi) / \partial \psi$ is uniformly bounded, and there is a $\varepsilon > 0$ such that $\pi(\Omega_i, \psi) \geq \varepsilon > 0$ for all $i = 1, \dots, n$.

- A8** The functions $\mu_1(\Omega_i, \varphi_1)$ and $\mu_2(\Omega_i, \varphi_2)$ are twice differentiable with respect to φ_1 and φ_2 on the compact sets Θ_{φ_1} and Θ_{φ_2} , respectively.

References

- Aalen, OO. Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory. Springer Verlag; New York: 1980. A model for nonparametric regression analysis of counting processes; p. 1-25.
- Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics*. 1994; 50:1064–1072. [PubMed: 7786988]
- Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis*. 2000; 6:39–58. [PubMed: 10763560]
- Breslow, NE., Lumley, T. Semiparametric models and two-phase samples: Applications to Cox regression. In: Banerjee, M., Bunea, F., Huang, J., Koltchinskii, V., Maathuis, M.H., editors. *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A Wellner*. Vol. 9. Beachwood, Ohio, USA: Institute of Mathematical Statistics; 2013. p. 65-77.
- Breslow N, Lumley T, Ballantyne C, Chambless L, Kulich M. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*. 2009a; 1:32–49. [PubMed: 20174455]
- Breslow N, Lumley T, Ballantyne C, Chambless L, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*. 2009b; 169:1398–1405. [PubMed: 19357328]
- Breslow N, Wellner J. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*. 2007; 34:86–102.
- Chen K. Generalized case-cohort sampling. *Journal of the Royal Statistical Society*. 2001; 63:791–809. Ser. B
- Chen K, Lo SH. Case-cohort and case-control analysis with Cox's model. *Biometrika*. 1999; 86:755–764.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82:835–845.
- Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF, the rgp120 HIV Vaccine Study Group. Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases*. 2005; 191:654–665. [PubMed: 15688278]
- Gao G, Tsiatis AA. Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. *Biometrika*. 2005; 92:875–891.
- Gilbert PB, Peterson ML, Follmann D, Hudgens MG, Francis DP, Gurwith M, Heyward WL, Jobes DV, Popovic V, Self SG, Sinangil F, Burke D, Berman PW. 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]
- Gottschalk P, Dunn J. The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Analytical Biochemistry*. 2005; 343:54–65. [PubMed: 15953581]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Huber, PJ. *Robust Statistics*. Wiley; New York: 1981.
- Huffer FW, McKeague IW. Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association*. 1991; 86:114–129.
- Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003; 90:341–353.
- Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*. 1988; 7:149–160. [PubMed: 3353602]

- Kang S, Cai J, Chambless L. Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis Risk in Communities (ARIC) study. *Biostatistics*. 2013; 14:28–41. [PubMed: 22826550]
- Kong L, Cai J. Case-cohort analysis with accelerated failure time model. *Biometrics*. 2009; 65:135–142. [PubMed: 18537948]
- Kulich M, Lin DY. Additive hazard regressions for case-cohort studies. *Biometrika*. 2000; 87:73–87.
- Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*. 2004; 99:832–844.
- Li Z, Gilbert PB, 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]
- Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*. 1993; 88:1341–1349.
- Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994; 81:61–71.
- Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*. 2001; 96:103–113.
- McKeague IW, Sasieni PD. A partly parametric additive risk model. *Biometrika*. 1994; 81:501–514.
- Murphy SA, Rossini AJ, van der Vaart AW. Maximum Likelihood Estimation in the Proportional Odds Model. *Journal of the American Statistical Association*. 1997; 92:968–976.
- Nan B, Wellner JA. A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica*. 2013; 23:1155–1180. [PubMed: 24489449]
- Plotkin SA, Gilbert PB. Nomenclature for immune correlates of protection after vaccination. *Clin Infect Dis*. 2012; 54:1615–1617. [PubMed: 22437237]
- Prentice RL. A Case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Saegusa T, Wellner JA. Weighted likelihood estimation under two-phase sampling. *Annals of Statistics*. 2013; 41:269–295. [PubMed: 24563559]
- Samuelsen SO, Ånestad H, Skrandal A. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*. 2007; 34:103–119.
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*. 1988; 16:64–81.
- Sun Y, Gilbert PB. Estimation of stratified mark-specific proportional hazards models with missing marks. *Scandinavian Journal of Statistics*. 2012; 39:34–52.
- Tsiatis, AA. *Semiparametric Theory and Missing Data*. Springer; New York: 2006.
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge University Press; New York: 1998.

Figure 1.

Comparison of the AIPW-R50, AIPW-R75, AIPW-R95, IPW, CC and Full estimators for the cumulative coefficient $B_2(t)$ under model (17) based on 1000 simulations with $n = 600$, and with sampling probabilities 0.5 for the cases and 0.1 for the non-cases: (a) The plots of the biases of the estimates; (b) The plots of the relative efficiencies of the estimators; (c) The coverage probabilities of the 95% pointwise confidence intervals for $B_2(t)$ for each t using the AIPW estimators.

Figure 2.

The AIPW estimates of the cumulative coefficients under model (18) using Neut_MN_IC50 and Neut_SF162_IC50 as the auxiliary variables: (a) The plot of $\hat{B}_1(t)$ for the cumulative baseline function with 95% pointwise confidence bands; (b) The plot of $\hat{B}_2(t)$ for the cumulative effect of CD4i_HIV2_IC50 with 95% pointwise confidence bands.

Table 1

Bias, empirical standard error (SSE), average of the estimated standard error (ESE), and empirical coverage probability (CP) of 95% confidence intervals for the AIPW-R50 estimator of γ under model (17) with $\rho = 0.5$ and about 80% censoring percentage based on 1000 simulations, where p_l is the sampling probability for the cases and p_0 is the sampling probability for the non-cases.

Size	Select. P.			γ_1			γ_2				
	n	p_1	p_0	Bias	SSE	ESE	CP	Bias	SSE	ESE	CP
600	1.0	0.10	0.0010	0.0306	0.0306	0.0305	0.948	0.0019	0.0506	0.0527	0.967
		0.25	-0.0001	0.0292	0.0290	0.0290	0.943	0.0015	0.0486	0.0504	0.961
		0.50	-0.0005	0.0292	0.0288	0.0288	0.940	0.0014	0.0480	0.0499	0.963
	0.5	0.10	0.0008	0.0309	0.0307	0.0307	0.948	0.0019	0.0506	0.0531	0.964
		0.25	-0.0003	0.0294	0.0291	0.0291	0.944	0.0017	0.0484	0.0505	0.961
		0.50	-0.0006	0.0292	0.0288	0.0288	0.944	0.0015	0.0480	0.0499	0.965
750	1.0	0.10	0.0004	0.0274	0.0274	0.0270	0.950	0.0013	0.0453	0.0465	0.963
		0.25	-0.0004	0.0266	0.0266	0.0260	0.950	0.0013	0.0438	0.0450	0.965
		0.50	-0.0006	0.0262	0.0262	0.0258	0.949	0.0012	0.0437	0.0258	0.963
	0.5	0.10	0.0002	0.0275	0.0275	0.0272	0.950	0.0013	0.0455	0.0469	0.965
		0.25	-0.0005	0.0266	0.0266	0.0261	0.951	0.0013	0.0440	0.0451	0.964
		0.50	-0.0007	0.0263	0.0263	0.0258	0.952	0.0012	0.0438	0.0447	0.960
1000	1.0	0.10	0.0017	0.0228	0.0228	0.0230	0.952	0.0006	0.0388	0.0398	0.950
		0.25	0.0011	0.0224	0.0224	0.0224	0.951	0.0004	0.0386	0.0389	0.948
		0.50	0.0008	0.0223	0.0223	0.0223	0.949	0.0004	0.0384	0.0387	0.950
	0.5	0.10	0.0016	0.0228	0.0228	0.0231	0.952	0.0005	0.0390	0.0400	0.945
		0.25	0.0010	0.0225	0.0225	0.0225	0.950	0.0004	0.0387	0.0389	0.948
		0.50	0.0007	0.0223	0.0223	0.0223	0.949	0.0004	0.0384	0.0387	0.948

Table 2

Comparison of Bias for the AIPW-R50 (AIPW shown in table), IPW, CC and Full estimators of γ under model (17) with $\rho = 0.5$ and about 80% censoring percentage based on 1000 simulations, where p_1 is the sampling probability for the cases and p_0 is the sampling probability for the non-cases.

Size	Select. P.		Bias(γ_1)				Bias(γ_2)			
	p_1	p_0	Full	AIPW	IPW	CC	Full	AIPW	IPW	CC
600	1.0	0.10	-0.0007	0.0010	0.0038	0.1924	0.0015	0.0019	0.0009	0.0075
	1.0	0.25		-0.0001	0.0011	0.1444		0.0015	0.0010	0.0069
	1.0	0.50		-0.0005	-0.0003	0.0754		0.0014	0.0015	0.0065
	0.5	0.10		0.0008	0.0041	0.0849		0.0019	0.0010	-0.0518
	0.5	0.25		-0.0003	0.0014	0.0361		0.0017	0.0013	-0.0212
	0.5	0.50		-0.0006	-0.0002	-0.0017		0.0015	0.0015	0.0023
750	1.0	0.10	-0.0007	0.0004	0.0025	0.1933	-0.0013	0.0013	0.0012	0.0161
	1.0	0.25		-0.0004	0.0001	0.1452		0.0013	0.0012	0.0117
	1.0	0.50		-0.0006	-0.0005	0.0768		0.0012	0.0012	0.0067
	0.5	0.10		0.0002	0.0022	0.0825		0.0013	0.0012	-0.0484
	0.5	0.25		-0.0005	-0.0001	0.0353		0.0013	0.0012	-0.0203
	0.5	0.50		-0.0007	-0.0006	-0.0002		0.0012	0.0013	0.0011
1000	1.0	0.10	0.0007	0.0017	0.0038	0.2022	0.0004	0.0006	-0.0003	-0.0003
	1.0	0.25		0.0011	0.0020	0.1493		0.0004	0.0003	0.0062
	1.0	0.50		0.0008	0.0010	0.0791		0.0004	0.0005	0.0060
	0.5	0.10		0.0016	0.0036	0.0852		0.0005	-0.0009	-0.0620
	0.5	0.25		0.0010	0.0018	0.0351		0.0004	-0.0002	-0.0235
	0.5	0.50		0.0007	0.0007	-0.0008		0.0004	-0.0001	0.0011

Table 3

Relative efficiencies (REE) of the AIPW-R50 (AIPW shown in table), IPW, CC estimators compared to the Full estimator for γ under model (17) with $\rho = 0.5$ and about 80% censoring percentage based on 1000 simulations, where p_1 is the sampling probability for the cases and p_0 is the sampling probability for the non-cases.

Size	Select. P.		REE(γ_1)			REE(γ_2)		
	p_1	p_0	AIPW	IPW	CC	AIPW	IPW	CC
600	1.0	0.10	0.9477	0.7572	0.2003	0.9486	0.7154	0.2017
	1.0	0.25	0.9932	0.9119	0.3173	0.9877	0.9040	0.3160
	1.0	0.50	0.9932	0.9764	0.5061	1.0000	0.9639	0.5128
	0.5	0.10	0.9385	0.7342	0.1975	0.9486	0.6847	0.2032
	0.5	0.25	0.9864	0.8788	0.3766	0.9917	0.8618	0.3774
	0.5	0.50	0.9932	0.9385	0.6532	1.0000	0.9125	0.6540
750	1.0	0.10	0.9562	0.8062	0.2017	0.9625	0.7279	0.2001
	1.0	0.25	0.9850	0.9258	0.3112	0.9954	0.9083	0.3283
	1.0	0.50	1.0000	0.9850	0.5078	0.9977	0.9732	0.5135
	0.5	0.10	0.9527	0.7939	0.2000	0.9582	0.7101	0.2111
	0.5	0.25	0.9850	0.9066	0.3743	0.9909	0.8826	0.3967
	0.5	0.50	0.9962	0.9562	0.6701	0.9954	0.9397	0.6636
1000	1.0	0.10	0.9737	0.8043	0.2015	0.9897	0.7918	0.2111
	1.0	0.25	0.9911	0.9250	0.3122	0.9948	0.9253	0.3380
	1.0	0.50	0.9955	0.9780	0.4912	1.0000	0.9846	0.5348
	0.5	0.10	0.9737	0.7929	0.2056	0.9846	0.7680	0.2136
	0.5	0.25	0.9867	0.8988	0.3731	0.9922	0.8930	0.3926
	0.5	0.50	0.9955	0.9487	0.6529	1.0000	0.9505	0.6621