Published in final edited form as:

Stat Biosci. 2017 June ; 9(1): 259–283. doi:10.1007/s12561-016-9177-5.

Estimation of Stratified Mark-Specific Proportional Hazards Models under Two-Phase Sampling with Application to HIV Vaccine Efficacy Trials

Guangren Yang,

School of Economics, Jinan University, Guangzhou, 510632, China

Yanqing Sun,

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

Li Qi, and

Biostatistics and Programming, Sanofi, Bridgewater, NJ 08807, USA

Peter B. Gilbert

Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Center, Seattle, WA 98109, USA

Abstract

An objective of preventive HIV vaccine efficacy trials is to understand how vaccine-induced immune responses to specific protein sequences of HIV-1 associate with subsequent infection with specific sequences of HIV, where the immune response biomarkers are measured in vaccine recipients via a two-phase sampling design. Motivated by this objective, we investigate the stratified mark-specific proportional hazards model under two-phase biomarker sampling, where the mark is the genetic distance of an infecting HIV-1 sequence to an HIV-1 sequence represented inside the vaccine. Estimation and inference procedures based on inverse probability weighting of complete-cases and on augmented inverse probability weighting of complete-cases are developed. Asymptotic properties of the estimators are derived and their finite-sample performances are examined in simulation studies. The methods are shown to have satisfactory performance, and are applied to the RV144 vaccine trial to assess whether immune response correlates of HIV-1 infection are stronger for HIV-1 infecting sequences similar to the vaccine than for sequences distant from the vaccine. Augmented inverse probability weighting; auxiliary variables; case-cohort design; censored failure time; competing risks; HIV vaccine efficacy trial.

1 Introduction

The statistical methods developed in this article are motivated by an objective in preventive vaccine efficacy trials. As an illustrative example throughout, RV144 was a preventive HIV vaccine efficacy trial that randomized 16,395 HIV-1 negative volunteers to receive vaccine

or placebo and monitored them for 42 months for occurrence of the primary study endpoint of HIV-1 infection. The rate of HIV-1 infection was significantly lower in the vaccine than placebo group, with estimated vaccine efficacy 31% on a multiplicative reduction scale estimated with a Cox model (95% confidence interval 1% to 51%; $p = 0.04$) (Rerks-Ngarm et al., 2009). A secondary objective assessed immune response biomarkers in vaccine recipients measured at the Week 26 visit as correlates of subsequent HIV-1 infection through Month 42 (Haynes et al., 2012). Because the HIV-1 infection rate was low (fewer than 1% of vaccine recipients acquired HIV), this objective was assessed using two-phase sampling of the Week 26 biomarkers, being measured in all participants with Week 26 samples who subsequently acquired HIV-1 infection (cases) and in a stratified random sample of participants with Week 26 samples who completed follow-up HIV-1 uninfected. Based on Borgan et al.'s (2000) estimator II with a Cox proportional hazards model, certain immune response biomarkers were significantly associated with HIV-1 infection; in particular vaccine recipients with higher levels of antibodies binding to the V1V2 portion of the HIV-1 envelope protein had a significantly lower rate of HIV-1 infection (Haynes et al., 2012; Zolla-Pazner et al., 2014; Yates et al., 2014).

A different secondary objective of RV144 assessed whether and how vaccine efficacy depends on genetic distances of infecting HIV-1 sequences to the HIV-1 sequences represented in the vaccine. This objective has been assessed using a mark-specific proportional hazards model (Sun, Gilbert, and McKeague, 2009; Sun and Gilbert, 2012; Juraska and Gilbert, 2013; Gilbert and Sun, 2015), where the mark is a measure of HIV-1 genetic distance to the vaccine construct only observed in HIV-1 infected cases. Application of these methods showed a non-significant trend of decreasing vaccine efficacy with HIV-1 genetic distance in V1V2 to one of the vaccine strains (92TH023) (Juraska and Gilbert, 2013), where the genetic distances were weighted Hamming distances based on a multiple sequence alignment that restricted to amino acid positions in V1V2.

The scientific goal of this work is to assess the associations of incompletely observed immune response biomarkers with the rate of subsequent HIV-1 infection, in a way that accounts for the genetic distances of the infecting HIVs. In RV144, the immune response biomarkers were measured at the Week 26 visit from 34 of 41 vaccine recipients who subsequently acquired HIV-1 infection (cases) and from 205 of 7010 vaccine recipients who completed follow-up HIV-1 uninfected (controls). To our knowledge no research has been published on statistical methods that address this problem by incorporating data both on incompletely observed covariates and on competing risks outcomes where the competing risks are described by a continuous mark variable (HIV-1 genetic distances in the RV144 illustration). This research can improve understanding of how vaccine-induced immune responses to specific protein sequences of HIV-1 associate with the hazard of infection with HIVs of specific protein sequences. This problem is important for any candidate vaccine designed to protect against a genetically heterogeneous pathogen (e.g., malaria, TB, dengue, HPV, HIV), because in general the adaptive immune system generates sequence-specific immune responses such that it is a fundamental issue that these immune responses may associate differently with infection depending on the sequence of the infection and how those sequences compare to those in the pathogen reagents used for detecting the immune responses.

Our investigation is carried out for the stratified mark-specific proportional hazards model in the competing risks failure time framework (Prentice, *et al.*, 1978) with a continuous mark variable playing the role of the cause-of-failure (Huang and Louis, 1998). The model was originally proposed and investigated by Sun, Gilbert and McKeague (2009) for one stratum to assess mark-specific HIV-1 vaccine efficacy, where the mark is a measure of the genetic distance of an HIV-1 sequence measured from a trial participant after acquiring HIV infection to an HIV-1 sequence contained inside the vaccine. Sun and Gilbert (2012) and Gilbert and Sun (2015) developed estimation and inference procedures for the stratified mark-specific proportional hazards model with missing marks. The methods have applications in evaluating mark-specific HIV-1 vaccine efficacy when the mark variable of interest is subject to missingness, for example due to rapid evolution of acquired HIV-1 sequences. This work investigates a different problem motivated by a different scientific question as mentioned above. We develop estimation procedures for the stratified mark-specific proportional hazards model when covariates, e.g. the immune responses, may have missing values. Since the mark is only observable in failures, it cannot be treated as a covariate. Therefore the statistical methods of Sun and Gilbert (2012) and Gilbert and Sun (2015) for missing marks are not applicable to the situation of missing covariates. Yet, vaccine efficacy trials in general have a large amount of missing data on immune responses, as they are only measured in a sub-sample such as through a two-phase sampling design (e.g., Breslow and Lumley, 2013).

Two-phase sampling is a more general form of case-cohort sampling. The original classical case-cohort sampling design (Prentice, 1986; Self and Prentice, 1988) measures the covariates of interest in a subcohort randomly sampled from all enrollees and also in all failure cases. For two-phase sampling data, phase-one data are all variables measured from all participants (including the minimum of the failure time and the censoring time, the indicator of failure or censoring, and some covariates), and the phase-two data are measured from a stratified random sample where the categorical stratification variable may depend on any phase-one information, and hence can be outcome-dependent (with nested case-control sampling the special case where the stratification variable is case-control status). Vaccine efficacy trials in general measure immune response biomarkers with a two-phase sampling scheme. Under Breslow and Lumley's (2013) terminology, two-phase sampling may be implemented via Bernoulli sampling or without replacement sampling within each phase-one stratum, and here our methods assume Bernoulli sampling. The motivating RV144 study used without replacement two-phase sampling, such that our methods do not exactly apply; however, our methods that assume Bernoulli sampling do provide conservative p-values and confidence intervals, and the degree of conservatism tends to be small, as shown by Breslow et al. (2009).

In this article, we develop and compare two estimation procedures for analyzing Bernoulli two-phase sampling with the stratified continuous mark-specific proportional hazards (PH) model. First we investigate an estimation method based on the inverse probability weighting (IPW) of complete-case technique of Horvitz and Thompson (1952). With this approach, if a participant has a missing value for one covariate, then the observed values of other covariates together with the observed failure/censoring time of the same participant are not fully utilized, losing efficiency. In the second approach, we adapt the approach of Robins,

Rotnitzky and Zhao (1994) to develop an augmented IPW estimating equation to improve efficiency, both by allowing the sampling probabilities to depend on the phase-one variables and by leveraging correlations of the phase-one variables with the phase-two covariates. The augmented IPW estimators are doubly robust and more efficient than the IPW estimators when the augmented part is correctly specified; see Tsiatis (2006) for the general theory and Sun and Gilbert (2012) for mark-specific vaccine efficacy applications.

The rest of the article is organized as follows. In Section 2, we introduce the stratified mark-specific proportional hazards model and the missing data mechanism. The estimation procedures are proposed in Section 3. The asymptotic properties of the proposed estimators are derived in Section 4. The finite-sample performance of the estimators is studied in simulations in Section 5. The methods are applied to analyze the RV144 data in Section 6. A discussion is given in Section 7. The proofs of the theorems are given in Appendix B placed in the online Supplementary Material at the journal's website.

2 Model, competing risks and the two-phase sampling

The stratified mark-specific proportional hazards (PH) model (Sun and Gilbert, 2012) is formulated under the competing risks framework. Suppose that T is the failure time of interest, V is a continuous mark variable, and $Z(t)$ is a possibly time-dependent p -dimensional covariate. Under the competing risks model, the mark V is only observed when T is observed. Let C be the censoring time, $X = \min\{T, C\}$ and $\delta = I(T \leq C)$ the censoring indicator. We assume that C is independent of (T, V) . The failure time T is observed if $\delta = 1$.

Suppose that the mark-specific hazard of failure at time t conditional on the covariate history $\mathcal{Z}(t) = \{Z(s), s \leq t\}$ only depends on the current value $Z(t)$. The model under investigation is based on the conditional mark-specific hazard function (Sun *et al.*, 2009) defined as $\lambda(t, v|z) = \lim_{h_1, h_2 \rightarrow 0} P\{T \in [t, t+h_1), V \in [v, v+h_2) | T \geq t, Z(t) = z\} / h_1 h_2$, which represents the instantaneous failure rate at time t due to $V = v$ conditional on the current covariate value $Z(t) = z$. Let τ be the end of follow-up time. Assume that V has a known bounded support on $[0, 1]$ and rescale it if necessary. The stratified mark-specific proportional hazards (PH) model (Sun and Gilbert, 2012) assumes

$$\lambda_k(t, v|z(t)) = \lambda_{0k}(t, v) \exp\{\beta(v)^T z(t)\}, \quad k=1, \dots, K, \quad (1)$$

for $t \in [0, \tau]$ and $v \in [0, 1]$, where K is the number of baseline covariate strata, $\beta(v)$ is a p -dimensional unknown function of v , and $\lambda_{0k}(t, v)$ is the unspecified baseline hazard function of (t, v) for the k th stratum. The mark-specific proportional hazards model studied by Sun *et al.* (2009) corresponds to $K = 1$.

We use the RV144 trial as an example for model interpretation. Suppose that T is the time to HIV-1 infection. Let V be the mark defined as the genetic distance of the V1V2 HIV-1 sequence measured after HIV-1 infection to one of the vaccine V1V2 sequences (92TH023 or A244), based on a multiply aligned data set of V1V2 amino acid sequences. Let Z be a

given biomarker measured after the vaccinations (at Week 26) that measures a vaccine-induced antibody response. Under model (1), the relative hazard of infection with HIVs of genetic distance v from a vaccine sequence for a vaccine recipient with antibody immune response $Z = a_2$ compared to a_1 is $RR(v) = \exp\{\beta(v)(a_2 - a_1)\}$. Thus, $RR(v)$ measures the association of the antibody response biomarker with the instantaneous rate of infection with viruses that are distance v from the vaccine. A negative value of $\beta(v)$ implies that vaccine recipients with higher biomarker values have a lower rate of v -specific HIV-1 infection. Furthermore, If $\beta(v)$ increases with v and has larger negative value for smaller v , then vaccine recipients exposed to HIVs with V1V2 sequences closer to the vaccine sequences 92TH023 or A244 may be more likely to be protected by antibodies than vaccine recipients exposed to HIVs with V1V2 sequences with larger genetic distances.

Sun and Gilbert (2012) and Gilbert and Sun (2015) studied estimation and hypothesis testing procedures for model (1) when the marks V are subject to missingness. This work investigates a different problem for model (1) motivated by a different scientific question. We develop estimation procedures for model (1) when covariates may have missing values, for example emanating from a two-phase sampling design as for the RV144 trial. The statistical methods developed here are used to investigate the roles of biomarkers measuring immune responses elicited by the vaccine on HIV-1 infection taking into account the genetic diversity of the infecting HIVs and the fact that the biomarkers are incompletely observed. Although general principles for dealing with missing covariates are analogous to those for missing marks, covariates cannot be treated as mark variables, such that the statistical procedures and their theoretical justifications must be established differently.

Suppose the covariate $Z(\cdot) = (Z_1^T(\cdot), Z_2^T(\cdot))^T$ consists of two parts; the phase-one covariates $Z_1(\cdot)$ measured in all participants and the phase-two covariates $Z_2(\cdot)$ measured in a random sample. Let ξ be the indicator random variable for whether a participant has complete covariate information. The probability a participant is selected into phase two may depend on the phase-one information available, $\Omega = (X, \delta, \delta V, Z_1(\cdot), A(\cdot))$, at the time of sampling, where $A(\cdot)$ are phase-one auxiliary variables predictive of phase-two covariates but need not be part of model (1). We assume $Z_2(\cdot)$ is missing at random (MAR), which can be expressed as $P(\xi = 1/\Omega, Z_2(\cdot)) = P(\xi = 1/\Omega)$. That is, whether $Z_2(\cdot)$ is missing may depend on the phase-one information but not on unobserved variables; see Rubin (1976).

3 Estimation procedures

We develop IPW and augmented IPW estimation procedures for analyzing the two-phase sampling data under the stratified mark-specific PH model. Let n_k be the number of

participants in the k th stratum; the total sample size is $n = \sum_{k=1}^K n_k$. We label the i th participant in the k th stratum with a pair of subscripts $\{ki\}$. Let $Z_{1,ki}(\cdot)$ and $Z_{2,ki}(\cdot)$ be the corresponding parts of covariates $Z_1(\cdot)$ and $Z_2(\cdot)$ for participant i in stratum k . The observed data are $(\Omega_{ki}, Z_{2,ki}(\cdot))$ for participants with $\xi_{ki} = 1$ and Ω_{ki} for participants with $\xi_{ki} = 0$, where $\Omega_{ki} = (X_{ki}, \delta_{ki}, \delta_{ki} V_{ki}, Z_{1,ki}(\cdot), A_{ki}(\cdot))$ is the phase-one data. Let $N_{ki}(t, v) = I(X_{ki} = t, \delta_{ki} = 1, V_{ki} = v)$ and $Y_{ki}(t) = I(X_{ki} = t)$. We consider Bernoulli two-phase sampling, with sampling probability for participant i in stratum k given by $\pi_{ki}(t) = P_k\{\xi_{ki} = 1/\Omega_{ki}, Y_{ki}(t) =$

1}. We assume that $\{T_{ki}, C_{ki}, V_{ki}, Z_{ki}(\cdot), \xi_{ki}, A_{ki}(\cdot); i = 1, \dots, n_k\}$ are i.i.d. replicates of $(T, C, V, Z(\cdot), \xi, A(\cdot))$ from stratum k .

3.1 Inverse probability weighted estimator

We propose a pseudo-score function using the inverse probability weighting (IPW) of the complete cases (Horvitz and Thompson, 1952). This approach modifies the full data estimating equations by weighting the contributions from the complete cases through the inverses of estimated sampling probabilities. Let $\hat{\pi}_k(t)$ be an estimator of $\pi_k(t)$ based on some parametric or nonparametric methods to be discussed in Section 3.3. Let $W_k(t) = \xi_k(\pi_k(t))^{-1}$ and $\hat{W}_k(t) = \xi_k(\hat{\pi}_k(t))^{-1}$. For $\beta \in \mathbb{R}_p$, $t \geq 0$, let

$$\tilde{S}_k^{(j)}(t, \beta) = n_k^{-1} \sum_{i=1}^{n_k} W_{ki}(t) Y_{ki}(t) \exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j} \text{ for } j = 0, 1, 2, \text{ where for any } z \in$$

$\mathbb{R}_p$, $z^{\otimes 0} = 1$, $z^{\otimes 1} = z$ and $z^{\otimes 2} = zz^T$. Let $\hat{\tilde{S}}_k^{(j)}(t, \beta)$ be the corresponding term of $\tilde{S}_k^{(j)}(t, \beta)$

obtained by replacing $W_{ki}(t)$ with $\hat{W}_k(t)$. Define $\tilde{Z}_k(t, \beta) = \tilde{S}_k^{(1)}(t, \beta) / \tilde{S}_k^{(0)}(t, \beta)$ and

$$\hat{\tilde{Z}}_k(t, \beta) = \hat{\tilde{S}}_k^{(1)}(t, \beta) / \hat{\tilde{S}}_k^{(0)}(t, \beta).$$

For each $v \in (0, 1)$, the local IPW pseudo-score function for $\beta(v)$ is defined as

$$U_w(v, \beta) = \sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^1 \int_0^v K_h(u-v) \hat{W}_{ki}(t) \left[Z_{ki}(t) - \hat{\tilde{Z}}_k(t, \beta) \right] N_{ki}(dt, du), \quad (2)$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a kernel function with compact support on $[-1, 1]$ and h is the bandwidth that depends on n . The IPW estimator $\hat{\beta}_w(v)$ for $\beta(v)$ is the root of (2).

The estimator of the doubly cumulative baseline function $\Lambda_{0k}(t, v) = \int_0^t \int_0^v \lambda_{0k}(s, u) ds du$

$$\text{given by } \hat{\Lambda}_{w,0k}(t, v) = \sum_{i=1}^{n_k} \int_0^t \int_0^v \{n_k \hat{\tilde{S}}_k^{(0)}(s, \hat{\beta}_w(u))\}^{-1} N_{ki}(ds, du).$$

3.2 Augmented inverse probability weighted estimator

Robins, Rotnitzky and Zhao (1994) proposed subtracting the projection term of the simple weighted estimating equation onto the nuisance tangent space to increase the efficiency of the estimators. Let

$$S_k^{(j)}(t, \beta) = n_k^{-1} \sum_{i=1}^{n_k} \{W_{ki}(t) Y_{ki}(t) \exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j} + (1 - W_{ki}(t)) Y_{ki}(t) E_k[\exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j} | \Omega_{ki}]\}$$

, for $j = 0, 1$, and 2 . Let $\bar{Z}_k(t, \beta) = S_k^{(1)}(t, \beta) / S_k^{(0)}(t, \beta)$. Here and henceforth, $E_k\{\cdot\}$ denotes the conditional expectation for an individual in stratum k . If the conditional expectations

$E_k[\exp\{\beta^T Z_{ki}(t)\} | \Omega_{ki}]$ and $E_k[\exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t) | \Omega_{ki}]$ in $S_k^{(0)}(t, \beta)$ and $S_k^{(1)}(t, \beta)$ are known, then the augmented IPW estimating equation for $\beta = \beta(v)$ can be written as

$$\begin{aligned}
 U_{aw}(v, \beta) &= \sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u-v) \left(W_{ki}(t) \{Z_{ki}(t) - \bar{Z}_k(t, \beta)\} N_{ki}(dt, du) + (1 - W_{ki}(t)) [E_k\{Z_{ki}(t) | \Omega_{ki}\} - \bar{Z}_k(t, \beta)] N_{ki}(dt, du) \right). \\
 &\quad (3)
 \end{aligned}$$

Let $\beta_1(v)$ and $\beta_2(v)$ be the coefficients for $Z_{1,kt}(t)$ and $Z_{2,kt}(t)$, respectively. For given β , $E_k\{Z_{kt}(t) | \Omega_{ki}\}$, $E_k[\exp\{\beta^T Z_{kt}(t)\} | \Omega_{ki}]$ and $E_k[Z_{kt}(t) \exp\{\beta^T Z_{kt}(t)\} | \Omega_{ki}]$ only depend on the phase-one data, and the conditional expectations $E_k\{Z_{2,kt}(t) | \Omega_{ki}\}$, $E_k\{\exp(\beta_2^T Z_{2,ki}(t)) | \Omega_{ki}\}$ and $E_k\{Z_{2,ki}(t) \exp(\beta_2^T Z_{2,ki}(t)) | \Omega_{ki}\}$ are the unknown components in (3). Let $\hat{E}_k\{g(Z_{2,kt}(t)) | \Omega_{ki}\}$ be the estimate of $E_k\{g(Z_{2,kt}(t)) | \Omega_{ki}\}$ based on some parametric or semiparametric models, where $g(Z_{2,kt}(t))$ is a function of $Z_{2,kt}(t)$ such as $Z_{2,kt}(t)$, $\exp(\beta_2^T Z_{2,ki}(t))$ or $Z_{2,ki}(t) \exp(\beta_2^T Z_{2,ki}(t))$. Although the exact models for the conditional expectations may be unknown, our simulations show that approximating these conditional expectations of the functions of the phase-two covariates using linear models with the phase-one predictors δ_{ki} , $Z_{1,kt}(t)$, $A_{kt}(t)$ and $\log(X_{ki})$ works well.

The estimating equation for the augmented IPW estimator for $\beta = \beta(v)$ is then

$$\begin{aligned}
 U_{aw}(v, \beta) &= \sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u-v) \left(\hat{W}_{ki}(t) \{Z_{ki}(t) - \hat{Z}_k(t, \beta)\} N_{ki}(dt, du) + (1 - \hat{W}_{ki}(t)) [\hat{E}_k\{Z_{ki}(t) | \Omega_{ki}\} - \hat{Z}_k(t, \beta)] N_{ki}(dt, du) \right) \\
 &\quad (4)
 \end{aligned}$$

where $\hat{Z}_k(t, \beta) = \hat{S}_k^{(1)}(t, \beta) / \hat{S}_k^{(0)}(t, \beta)$, and

$$\begin{aligned}
 \hat{S}_k^{(j)}(t, \beta) &= n_k^{-1} \sum_{i=1}^{n_k} \left\{ \hat{W}_{ki}(t) Y_{ki}(t) \exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j} + (1 - \hat{W}_{ki}(t)) Y_{ki}(t) \hat{E}_k[\exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j} | \Omega_{ki}] \right\}, \\
 &\quad (5)
 \end{aligned}$$

for $j = 0, 1$, and 2 . Here the first part of $\hat{E}_k\{Z_{kt}(t) | \Omega_{ki}\}$ is $Z_{1,kt}(t)$ and the second part is $\hat{E}_k\{Z_{2,kt}(t) | \Omega_{ki}\}$. The estimate $\hat{E}_k[\exp\{\beta^T Z_{kt}(t)\} Z_{kt}(t)^{\otimes j} | \Omega_{ki}]$ can be similarly worked out as a function of the phase-one data and $\hat{E}_k[\exp\{\beta_2^T Z_{2,ki}(t)\} Z_{2,ki}(t)^{\otimes r} | \Omega_{ki}]$, for $r = 0, 1$, and 2 .

The augmented IPW estimator (AIPW) of $\beta_0(v)$ is obtained by solving $U_{aw}(v, \beta) = 0$ and is denoted by $\hat{\beta}_{aw}(v)$.

Let $\hat{E}_k\{Z_{2,k}(t)/\Omega_{ki}\}$ be the estimate of $E_k\{Z_{2,k}(t)/\Omega_{ki}\}$. The estimator $\hat{\beta}_{aw}(v)$ can be implemented using the Newton-Raphson iterative algorithm. Starting with an initial value $\beta^{(0)}(v)$, let $\beta^{(m)}(v)$ be the estimate of $\beta(v)$ at step m . The estimator $\hat{\beta}_{aw}(v)$ is obtained by iterating the following steps (i) and (ii) until convergence: (i) Estimate the conditional expectations $E_k[\exp\{(\beta_2^{(m)}(v))^T Z_{2,ki}(t)\} Z_{2,ki}(t)^{\otimes j} | \Omega_{ki}]$ for $j = 0, 1, 2$, and calculate $\hat{Z}_k(t, \beta^{(m)}(v))$; (ii) The updated estimate $\beta^{(m+1)}(v)$ at step $m + 1$ is obtained by $\beta^{(m+1)}(v) = \beta^{(m)}(v) - (U_{aw}(v, \beta^{(m)}(v)) / \beta)^{-1} U_{aw}(v, \beta^{(m)}(v))$.

The baseline function $\lambda_{0k}(t, v)$ can be estimated by $\hat{\lambda}_{aw,0k}(t, v)$, obtained by smoothing the increments of the estimator $\hat{\Lambda}_{aw,0k}(t, v) = \sum_{i=1}^{n_k} \int_0^t \int_0^v \{n_k \hat{S}_k^{(0)}(s, \hat{\beta}_{aw}(u))\}^{-1} N_{ki}(ds, du)$ of the doubly cumulative baseline function $\Lambda_{0k}(t, v) = \int_0^t \int_0^v \lambda_{0k}(s, u) ds du$. For example, one can use kernel smoothing $\hat{\lambda}_{aw,0k}(t, v) = \int_0^t \int_0^1 K_{h_1}^{(1)}(t-s) K_{h_2}^{(2)}(v-u) \hat{\Lambda}_{aw,0k}(ds, du)$, where $K_{h_1}^{(1)}(x) = K^{(1)}(x/h_1)/h_1$ and $K_{h_2}^{(2)}(x) = K^{(2)}(x/h_2)/h_2$, with $K^{(1)}(\cdot)$ and $K^{(2)}(\cdot)$ kernel functions and h_1 and h_2 bandwidths.

3.3 Estimation of sampling probabilities

In a two-phase sampling design, the probability that a participant is selected into phase two may depend on the phase-one information Ω_{ki} . When the sampling probability $\pi_{ki}(t)$ does not depend on the at-risk set at time t , $\pi_{ki}(t) \equiv \pi_{ki}$, and parametric models such as logistic regression models are commonly used to estimate π_{ki} . Assume that $\pi_{ki} = \pi_k(\Omega_{ki}, \psi_k)$, where $\pi_k(\cdot, \psi_k)$ is a known function of Ω_{ki} up to a q -dimensional vector ψ_k of unknown parameters. The parameter $\psi = (\psi_1, \dots, \psi_K)^T$ can be estimated by the maximum likelihood estimator $\hat{\psi} = (\hat{\psi}_1, \dots, \hat{\psi}_K)^T$, which maximizes the likelihood function

$L_\pi(\psi) = \prod_{k=1}^K \prod_{i=1}^{k_i} \pi_k(\Omega_{ki}, \psi_k)^{\xi_{ki}} \{1 - \pi_k(\Omega_{ki}, \psi_k)\}^{1 - \xi_{ki}}$. Then π_{ki} is estimated by $\hat{\pi}_{ki} = \pi_k(\Omega_{ki}, \hat{\psi}_k)$. Let ψ_{k0} be the true value of ψ_k . By standard likelihood based analysis, we have

the asymptotic approximation $\hat{\psi}_k - \psi_{k0} = (I_k^\psi)^{-1} n_k^{-1} \sum_{i=1}^{n_k} G_{ki}^\psi + o_p(n^{-1/2})$, where

$$G_{ki}^\psi = \{\pi_k(\Omega_{ki}, \psi_{k0}) (1 - \pi_k(\Omega_{ki}, \psi_{k0}))\}^{-1} (\xi_{ki} - \pi_k(\Omega_{ki}, \psi_{k0})) \frac{\partial \pi_k(\Omega_{ki}, \psi_{k0})}{\partial \psi_k} \text{ and } I_k^\psi = E\{(G_{ki}^\psi)^{\otimes 2}\}. \text{ It follows that}$$

$$\pi_k(\Omega_{ki}, \hat{\psi}_k) - \pi_k(\Omega_{ki}, \psi_{k0}) = \pi_k'(\Omega_{ki}, \psi_{k0}) (I_k^\psi)^{-1} n_k^{-1} \sum_{j=1}^{n_k} G_{kj}^\psi + o_p(n^{-1/2}). \quad (6)$$

To allow for more flexibility, different parametric models can be used for cases ($\delta_{ki} = 1$) and controls ($\delta_{ki} = 0$), as well as for different phase-one covariate strata.

When $\pi_{ki}(t)$ depends on the at-risk set at time t , nonparametric methods can be used to estimate $\pi_{ki}(t)$, cf. Qi, Wang and Prentice (2005). Often in practice, a simple and reliable method for estimating the sampling probabilities can be based on the cell probabilities, cf. Kulich and Lin (2004). Suppose that the sampling depends on discrete phase-one random variables D_{ki} and the failure status $Y_{ki}(t)$ at time t , where D_{ki} may be formulated based on Ω_{ki} . In this case, for an individual in the k th stratum, the sampling probability at time t is $\pi_{ki}(t) = P_k(\xi_{ki} = 1/\Omega_{ki}, Y_{ki}(t) = 1) = P_k(\xi_{ki} = 1/D_{ki}, Y_{ki}(t) = 1)$. Let B_{ki} be a function of Ω_{ki} . If the sampling probability depends on the time of sampling, then we can take $A_{ki}(t) = Y_{ki}(t)B_{ki}$ and $A_{ki}(t) = B_{ki}$ otherwise. Then $\pi_{ki}(t)$ can be estimated by

$$\hat{\pi}_{ki}(t) = \left(\sum_{j=1}^{n_k} A_{kj}(t) I\{D_{kj} = D_{ki}\} \right)^{-1} \left(\sum_{j=1}^{n_k} \xi_{kj} A_{kj}(t) I\{D_{kj} = D_{ki}\} \right). \quad (7)$$

Different choices of $A_{ki}(t)$ may lead to different estimators. Self and Prentice (1988) used constant weights $A_{ki}(t) = I_p$, while Barlow (1994), Borgan et al. (2000) and Chen (2001) considered time-varying weights $A_{ki}(t) = Y_{ki}(t)I_p$. Under some regularity conditions, $\hat{\pi}_{ki}(t)$ has the following asymptotic approximation:

$$n_k^{1/2}(\hat{\pi}_{ki}(t) - \pi_{ki}(t)) = I_{k,\pi}^{-1}(\Omega_{ki}, t) n_k^{-1/2} \sum_{j=1}^{n_k} G_{kj}(t) + o_p(1), \quad (8)$$

uniformly in i and t , where $G_{kj}(t)$ are independent identically distributed random vectors with $E_k(G_{kj}(t)/\Omega_{kj}) = 0$ for $j = 1, \dots, n_k$ and $I_{k,\pi}(\omega, t)$ is a nonsingular matrix.

In view of (6) and (8), we assume the estimator $\hat{\pi}_{ki}(t)$ of the sampling probability satisfies (8), which is stated as Condition (A.6) in Appendix B.

4 Asymptotic properties

This section investigates the asymptotic properties of the proposed estimators. Let $\beta_0(v)$ be the true regression function of $\beta(v)$. Let $s_k^{(j)}(t, \beta) = E\{Y_{ki}(t) \exp\{\beta^T Z_{ki}(t)\} Z_{ki}(t)^{\otimes j}\}$ for $j = 0, 1, 2$. Let $\bar{z}_k(t, \beta) = s_k^{(1)}(t, \beta)/s_k^{(0)}(t, \beta)$ and $J_k(t, \beta) = s_k^{(2)}(t, \beta)/s_k^{(0)}(t, \beta) - (\bar{z}_k(t, \beta))^{\otimes 2}$. Let $\sum_k(v) = \int_0^\tau J_k(t, \beta_0(v)) \lambda_{0k}(t, v) s_k^{(0)}(t, \beta_0(v)) dt$ and $\sum(v) = \sum_{k=1}^K p_k \sum_k(v)$, where $p_k = \lim_{n_K \rightarrow \infty} n_k/n$ and $n = \sum_{k=1}^K n_k$. Let $\nu_0 = \int K^2(x) dx$ and $\mu_2 = \int x^2 K(x) dx$. Let $\mathcal{F}_t = \sigma\{I(X_{ki} \leq s, \delta_{ki} = 1), I(X_{ki} \leq s, \delta_{ki} = 0), V_{ki} I(X_{ki} \leq s, \delta_{ki} = 1), Z_{ki}(s); 0 \leq s \leq t, i = 1, \dots, n_k, k = 1, \dots, K\}$ be the right-continuous filtration generated by the data processes $\{N_{ki}(s, v), Y_{ki}(s), Z_{ki}(s); 0 \leq s \leq t, 0 \leq v \leq 1, i = 1, \dots, n_k, k = 1, \dots, K\}$. Under model (1), the mark-specific intensity of $N_{ki}(t, v)$ with respect to $\mathcal{F}_t$ is $Y_{ki}(t) \lambda_{ki}(t, v)$, where $\lambda_{ki}(t, v) = \lambda_{0ki}(t, v) \exp\{\beta_0(v)^T Z_{ki}(t)\}$. Let $\mathcal{F}_t^* = \mathcal{F}_t \cup \{\xi_{ki}, A_{ki}(s), 0 \leq s \leq t; i = 1, \dots, n_k, k = 1, \dots, K\}$ be the enlarged filtration over $\lambda_{ki}^*(t, v)$ is defined through

$E(N_{ki}(dt, dv)|\mathcal{F}_{t-}^*) = Y_{ki}(t)\lambda_{ki}^*(t, v) dt dv$, where $\lambda_{ki}^*(t, v)Y_{ki}(t)$ is the mark-specific intensity of $N_{ki}(t, v)$ with respect to the enlarged filtration $\mathcal{F}_t^*$.

Under some smoothness conditions, the Glivenko-Cantelli Theorem (cf. van der Vaart, 1998) can be used to show that $\sup_k |S_k^{(j)}(t, \beta) - s_k^{(j)}(t, \beta)| \xrightarrow{P} 0$ uniformly in $t \in [0, \tau]$ for $j = 0, 1$,

2. If $\sup_{k,i} |\hat{W}_{ki}(t) - W_{ki}(t)| \xrightarrow{P} 0$ uniformly in $t \in [0, \tau]$, then

$\sup_k |\hat{S}_k^{(j)}(t, \beta) - s_k^{(j)}(t, \beta)| \xrightarrow{P} 0$ uniformly in $t \in [0, \tau]$. This leads to the consistency of the estimator $\hat{\beta}_{aw}(v)$ regardless of whether $\sup_{k,i} |\hat{E}_k\{Z_{ki}(t)|\Omega_{ki}\} - E_k\{Z_{ki}(t)|\Omega_{ki}\}| \xrightarrow{P} 0$ and $\hat{Z}_k(t, \beta) \xrightarrow{P} \bar{z}_k(t, \beta)$ hold. That is, both the estimators $\hat{\beta}_w(v)$ and $\hat{\beta}_{aw}(v)$ are consistent under $\sup_{k,i} |\hat{W}_{ki}(t) - W_{ki}(t)| \xrightarrow{P} 0$. On the other hand, if

$\sup_{k,i} |\hat{E}_k\{Z_{ki}(t)|\Omega_{ki}\} - E_k\{Z_{ki}(t)|\Omega_{ki}\}| \xrightarrow{P} 0$ and $\hat{Z}_k(t, \beta) \xrightarrow{P} \bar{z}_k(t, \beta)$ hold, then $\hat{\beta}_{aw}(v)$ is also consistent even if $\sup_{k,i} |\hat{W}_{ki}(t) - W_{ki}(t)| \xrightarrow{P} 0$ fails. This is the so-called double robustness property. Under the two-phase sampling designs and in other situations with missing covariates, $\pi_{ki}(t)$ can often be consistently estimated. However, the consistent estimators for $E_k\{Z_{ki}(t)/\Omega_{ki}\}$ and $\bar{z}_k(t, \beta)$ are often hard to come by. Nevertheless, the regression techniques can still be used to provide reasonable estimations for $E_k\{Z_{ki}(t)/\Omega_{ki}\}$ and $\bar{z}_k(t, \beta)$ to recover some of the missing information, and thus improve efficiency for estimating $\beta(v)$. The next two subsections investigate the asymptotic properties of $\hat{\beta}_w(v)$ and $\hat{\beta}_{aw}(v)$ under the ideal situation that both the sampling probabilities and the conditional moments concerning $Z_{ki}(t)$ can be consistently estimated. The conditions for the asymptotic results are specified in Condition A of Appendix B.

4.1 Asymptotic results of the IPW estimator

Let $\vartheta_w(v) = \sum_{k=1}^K p_k E_k\{\int_0^\tau [Z_{ki}(t) - \bar{z}_k(t, \beta)] \frac{\partial^2 \lambda_{ki}(t, v)}{\partial v^2} Y_{ki}(t) dt\}$, and $\sum_{w,k}^*(v)$ be the derivative of $E_k\{\int_0^\tau \int_0^\tau [Z_{ki}(t) - \bar{z}_k(t, \beta_0(u))]^{\otimes 2} \xi_{ki}(\pi_{ki}(t))^{-2} N_{ki}(dt, du)\}$ with respect to v . The following theorem presents the asymptotic consistency and asymptotic normality of the IPW estimator $\hat{\beta}_w(v)$.

Theorem 1—Under Condition A,

- a. $\hat{\beta}_w(v) \xrightarrow{P} \beta_0(v)$, uniformly in $v \in [a, b] \subset (0, 1)$ as $n \rightarrow \infty$;
- b. $(nh)^{1/2} [\hat{\beta}_w(v) - \beta_0(v) - \frac{1}{2}\mu_2 h^2 \sum^{-1}(v) \vartheta_w(v)] \xrightarrow{D} N(0, \nu_0 \sum^{-1}(v) \sum_w^*(v) \sum^{-1}(v))$, for $v \in [a, b]$ as $n \rightarrow \infty$, where $\sum_w^*(v) = \sum_{k=1}^K p_k \sum_{w,k}^*(v)$.

By the first order Taylor expansion of the score function,

$$\hat{\beta}_w(v) - \beta_0(v) = -(U'_w(v, \beta^*(v)))^{-1} U_w(v, \beta_0(v)), \text{ where}$$

$$U'_w(v, \beta) = - \sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u, v) \hat{J}_k(t, \beta) N_{ki}(dt, du),$$

$\hat{J}_k(t, \beta) = \hat{S}_k^{(2)}(t, \beta) / \hat{S}_k^{(0)}(t, \beta) - (\hat{Z}_k(t, \beta))^{\otimes 2}$, and $\beta^*(v)$ is on the line segment between $\hat{\beta}_w(v)$ and $\beta_0(v)$. It can be shown that $\sum_w(v) \equiv -n^{-1} U'_w(v, \hat{\beta}_w(v)) \xrightarrow{P} \sum(v)$ as $n \rightarrow \infty$. By the asymptotic approximation for $n^{-1/2} h^{1/2} U_w(v, \beta_0(v))$ given in (W.7) in Appendix B, $\nu_0 \sum_w^*(v)$ can be consistently estimated by

$$\sum_w^*(v) = n^{-1} h \sum_{k=1}^K \sum_{i=1}^{n_k} \left[\int_0^1 \int_0^\tau K_h(u-v) \hat{W}_{ki}(t) \left[Z_{ki}(t) - \hat{Z}_k(t, \hat{\beta}_w(u)) \right] \hat{M}_{ki}(dt, du) \right]^{\otimes 2}, \quad (9)$$

where $\hat{M}_{ki}(dt, du) = N_{ki}(dt, du) - Y_{ki}(t) \exp\{\hat{\beta}_w(u)^T Z_{ki}\} \hat{\Lambda}_{w,0k}(dt, du)$. Thus, the asymptotic variance $\nu_0 \sum^{-1}(v) \sum_w^*(v) \sum^{-1}(v)$ can be consistently estimated by $(\sum_w(v))^{-1} \sum_w^*(v) (\sum_w(v))^{-1}$.

4.2 Asymptotic results of the AIPW estimator

Let $O_{ki} = (\delta_{ki} V_{ki}, Z_{1,ki}(\cdot), A_{ki}(\cdot))$. We can write $\Omega_{ki} = (X_{ki}, \delta_{ki}, O_{ki})$. For convenience of the stochastic analysis, we explicitly spell out the dependence of $\pi_{ki}(t, \Omega_{ki})$ on (X_{ki}, δ_{ki}) and define $W_{ki}^*(t, X_{ki}, \delta_{ki}) = \xi_{ki}(\pi_{ki}(t, X_{ki}, \delta_{ki}, O_{ki}))^{-1}$. We define $Q_{ki}^*(t) = W_{ki}^*(t, t, 1)$, $Z_{ki}^o(t) = Q_{ki}^*(t) Z_{ki}(t) + (1 - Q_{ki}^*(t)) E\{Z_{ki}(t) | (X_{ki}=t, \delta_{ki}=1, O_{ki})\}$,

$$\begin{aligned} \vartheta_{aw}(v) = & \sum_{k=1}^K \int_0^\tau p_k E_k \left(W_{ki}(t) Y_{ki}(t) \{Z_{ki}(t) - \bar{z}_k(t, \beta_0(v))\} \frac{\partial^2 \lambda_{ki}(t, v)}{\partial v^2} + (1 - W_{ki}(t)) Y_{ki}(t) \right. \\ & \left. \left[E \left\{ Z_{ki}(t) \frac{\partial^2 \lambda_{ki}(t, v)}{\partial v^2} | \Omega_{ki} \right\} - \bar{z}_k(t, \beta_0(v)) E \left\{ \frac{\partial^2 \lambda_{ki}(t, v)}{\partial v^2} | \Omega_{ki} \right\} \right] \right) dt, \\ \sum_{aw,k}^*(v) = & E \left\{ \int_0^\tau [Z_{ki}^o(t) - \bar{z}_k(t, \beta_0(v))]^{\otimes 2} \lambda_{ki}^*(t, v) Y_{ki}(t) dt \right\}. \end{aligned}$$

By the property of double expectation, it is easy to see that $\sum_{aw,k}^*(v)$ is the derivative of $E_k \left\{ \int_0^\tau \int_0^\tau [Z_{ki}^o(t) - \bar{z}_k(t, \beta_0(u))]^{\otimes 2} N_{ki}(dt, du) \right\}$ with respect to v .

The following theorem establishes asymptotic consistency and asymptotic normality of the AIPW estimator $\hat{\beta}_{aw}(v)$.

Theorem 2—Under Condition A,

- a.** $\hat{\beta}_{aw}(v) \xrightarrow{P} \beta_0(v)$, uniformly in $v \in [a, b] \subset (0, 1)$ as $n \rightarrow \infty$;

b.

$$(nh)^{1/2} \left[\hat{\beta}_{aw}(v) - \beta_0(v) - \frac{1}{2} \mu_2 h^2 \sum^{-1}(v) \vartheta_{aw}(v) \right] \xrightarrow{D} N(0, \nu_0 \sum^{-1}(v) \sum_{aw}^*(v) \sum^{-1}(v))$$

, for $v \in [a, b]$ as $n \rightarrow \infty$, where $\sum_{aw}^*(v) = \sum_{k=1}^K p_k \sum_{aw,k}^*(v)$.

Note that $Z_{ki}^o(t) = Z_{ki}(t) - (1 - Q_{ki}^*(t)) \{Z_{ki}(t) - E_k[Z_{ki}(t) | (X_{ki}=t, \delta_{ki}=1, O_{ki})]\}$. Under MAR, $\{Z_{ki}(t) - \bar{z}_k(t, \beta(u))\} \lambda_{ki}^*(t, v) Y_{ki}(t)$ and $(1 - Q_{ki}^*(t)) \{Z_{ki}(t) - E_k[Z_{ki}(t) | (X_{ki}=t, \delta_{ki}=1, O_{ki})]\} \lambda_{ki}^*(t, v) Y_{ki}(t)$ are uncorrelated. It follows that

$$\sum_{aw,k}^*(v) = \sum_k (v) + E_k \left\{ \int_0^\tau ((\pi_{ki}^*(t))^{-1} - 1) \{Z_{ki}(t) - E_k[Z_{ki}(t) | (X_{ki}=t, \delta_{ki}=1, O_{ki})]\}^{\otimes 2} \lambda_{ki}^*(t, v) Y_{ki}(t) dt \right\},$$

(10)

where $\pi_{ki}^*(t) = P_k\{\xi_{ki}=1 | (X_{ki}=t, \delta_{ki}=1, O_{ki})\}$. The second term of (10) indicates the efficiency loss due to the missing data.

The $\Sigma(v)$ for $\hat{\beta}_{aw}(v)$ can be estimated consistently by $\hat{\Sigma}_{aw}(v) = -n^{-1} U'_{aw}(v, \hat{\beta}_{aw}(v))$, where $U'_{aw}(v, \beta) = -\sum_{k=1}^K \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u-v) \hat{J}_k(t, \beta) N_{ki}(dt, du)$ is the derivative of $U_{aw}(v, \beta)$ with respect to β , and $\hat{J}_k(t, \beta) = \hat{S}_k^{(2)}(t, \beta) / \hat{S}_k^{(0)}(t, \beta) - (\hat{Z}_k(t, \beta))^{\otimes 2}$.

By the asymptotic approximation (W.13) given in Appendix B, the middle part $\nu_0 \sum_{aw}^*(v)$ of the asymptotic variance for $\hat{\beta}_{aw}(v)$ can be consistently estimated by

$$\begin{aligned} & \hat{\Sigma}_{aw}^*(v) \\ &= n^{-1} h \sum_{k=1}^K \sum_{i=1}^{n_k} \left\{ \int_0^1 \int_0^\tau K_h(u-v) \hat{W}_{ki}(t) \left[Z_{ki}(t) - \hat{Z}_k(t, \hat{\beta}_{aw}(u)) \right] \hat{M}_{ki}(dt, du) - \int_0^1 \int_0^\tau K_h(u-v) (1 - \hat{W}_{ki}(t)) \hat{M}_{ki, \bar{z}}^o(dt, du) \right\}^{\otimes 2} \end{aligned}$$

(11)

where $\hat{M}_{ki}(dt, du) = N_{ki}(dt, du) - Y_{ki}(t) \exp\{\hat{\beta}_{aw}(u)^T Z_{ki}(t)\} \hat{\Lambda}_{aw,0k}(dt, du)$ and $\hat{M}_{ki, \bar{z}}^o(dt, du)$ is the estimator of $M_{ki, \bar{z}}^o(dt, du) = E[\{Z_{ki}(t) - \bar{z}_k(t, \beta_0(u))\} M_{ki}(dt, du) | \Omega_{ki}]$. Since

$$\begin{aligned}
M_{ki,\bar{z}}^o(dt, du) &= [E\{Z_{ki}(t)|\Omega_{ki}\} \\
&\quad - \bar{z}_k(t, \beta_0(u))] N_{ki}(dt, du) \\
&\quad - (E[Z_{ki}(t)\exp\{\beta_0^T(v)Z_{ki}(t)\}|\Omega_{ki}] \\
&\quad - \bar{z}_k(t, \beta_0(u))E[\exp\{\beta_0^T(v)Z_{ki}(t)\}|\Omega_{ki}]) Y_{ki}(t)\lambda_{0k}(t, u) dt du,
\end{aligned}$$

$\hat{M}_{ki,\bar{z}}^o(dt, du)$ is obtained by replacing $\bar{z}_k(t, \beta_0(u))$ with $\hat{Z}_k(t, \hat{\beta}_{aw}(u))$, $\lambda_{0k}(t, u) dt du$ with $\hat{\Lambda}_{aw,0k}(dt, du)$, and by replacing $E\{Z_{ki}(t)|\Omega_{ki}\}$, $E[\exp\{\beta_0^T(v)Z_{ki}(t)\}|\Omega_{ki}]$ and $E[Z_{ki}(t)\exp\{\beta_0^T(v)Z_{ki}(t)\}|\Omega_{ki}]$ with their estimators. Thus, the asymptotic variance $\nu_0 \sum^{-1}(v) \sum_{aw}^*(v) \sum^{-1}(v)$ can be estimated by $(\sum_{aw}(\hat{v}))^{-1} \sum_{aw}^*(\hat{v}) (\sum_{aw}(\hat{v}))^{-1}$.

5 Simulation study

Simulation studies are conducted to examine and compare the finite-sample properties of the IPW estimator (IPW) and the augmented IPW estimator (AIPW), and also to compare their performance with that of the full-cohort (Full) and complete-case (CC) estimators. We consider the case with $K=1$ stratum, in which case the CC and full estimators are based on the methods of Sun et al. (2009).

Let Z_1 be a binary covariate taking value 0 or 1 with probability 0.5 for each value and let Z_2 be uniformly distributed on $[0, 1]$. The variables (T, V) are generated from the following mark-specific proportional hazards model:

$$\lambda(t, v|z) = \exp\{\gamma v + \beta_1(v)Z_1 + \beta_2(v)Z_2\}, \quad t \geq 0, 0 \leq v \leq 1, \quad (12)$$

where $\gamma = 0.3$, $\beta_1(v) = -0.6 + 0.6v$ and $\beta_2(v) = 0.3v$. Under model (12), the mark-specific baseline function is $\lambda_0(t, v) = \exp(0.3v)$ and the mark-specific relative risk effect of Z_1 is $RR(v) = \exp(\beta_1(v))$.

Failure times greater than $\tau = 2.0$ are administratively right-censored. Censoring times are generated from an exponential distribution, independent of (T, V) , with parameter adjusted so that the overall censoring rates range from 25% to 35%. A cohort size of $n = 500$ is used for all simulation settings. A total of 1000 simulation runs are generated for each setting. We assume that Z_1 is a phase-one covariate observed for all participants, while Z_2 is a phase-two covariate. Suppose that the auxiliary variable $A = Z_2 + \varepsilon$, where ε is normal with mean 0, independent of Z_2 . The standard deviation of ε is chosen as 0.3 and 0.11, which corresponds to a correlation coefficient of $\rho = 0.7$ and 0.93 between Z_2 and A , respectively. We use the Epanechnikov kernel $K(x) = 0.75(1-x^2)I\{|x| \leq 1\}$. The bandwidth is selected using the formula $h = 4\sigma_v n^{-1/3}$, where σ_v is the estimated standard error of the observed marks for uncensored failure times. Similar bandwidth selection has been used in Zhou and Wang (2000) and Sun, Li and Gilbert (2015). Our simulations use the bandwidth $h = 4\sigma_v n^{-1/3} = 4(0.285)(500^{-1/3}) = 0.15$.

We consider four simulation settings in terms of how the phase-two covariate is sampled and whether the sampling probabilities are modeled correctly, all of which measure Z_2 in about 50% of participants. The first three settings compare efficiency of the estimators under correctly specified models for the sampling probabilities and the fourth setting examines robustness of these estimators to model misspecification of the sampling probabilities. The first three simulation settings use the following sampling probability models: (S1) the phase-two covariate is a simple random sample from the phase-one sample with sampling probability $P(\xi_{ki} = 1) = 0.6$; (S2) the phase-two covariate is a stratified random sample stratified by the failure status δ_{ki} with $P(\xi_{ki} = 1/\delta_{ki} = 1) = 0.7$ and $P(\xi_{ki} = 1/\delta_{ki} = 0) = 0.5$; and (S3) the sampling probability depends on both the failure status δ_{ki} and the auxiliary information A_{ki} with $\text{logit}(P(\xi_{ki} = 1/\Omega_{ki})) = 1.2Z_{1,ki} - 0.4\delta_{ki} - 0.3$.

The sampling probabilities are estimated using the logit model $\text{logit}(\pi_k(\Omega_{ki})) = \psi_0 + \psi_1\delta_{ki} + \psi_2Z_{1,ki}$ and the conditional expectations for the AIPW estimators are estimated using linear models with predictors δ_{ki} , $Z_{1,ki}$, A_{ki} and $\log(X_{ki})$. When the correlation coefficient between $Z_{2,ki}$ and A_{ki} is $\rho \approx 0.93$, the AIPW estimator is denoted by AIPW-A2, and by AIPW-A1 for $\rho \approx 0.7$. AIPW-A0 is the AIPW estimator in the case where the auxiliary information A_{ki} is not available and is not included in the linear model for estimating the conditional expectations. The performances of the estimators, IPW, AIPW, CC and Full, under the settings (S1)–(S3) are shown in Figures 1–3, respectively, in terms of bias (Bias), the sample standard error of the estimators (SEE), the sample mean of the estimated standard errors (ESE), and the 95% empirical coverage probability (CP).

For (S1), the results show that the biases of all estimators are very small at a level comparable to that of the full data likelihood estimator. This is because the data are missing completely at random (MCAR) under this setting. For (S2) and (S3), the complete-case estimator yields large bias, as expected due to the non-MCAR missingness. Across the simulation settings the AIPW estimators have smaller standard errors than the CC and IPW estimators, validating that the AIPW estimator is more efficient than the IPW estimator even without using the auxiliary variables as in (S3). The standard error of the AIPW estimator decreases as the correlation between the auxiliary variable and the missing variable increases, with the standard errors getting closer to that of the Full estimator. In particular, we notice that the standard error of the AIPW estimator for the coefficient of the phase-one covariate Z_1 is much smaller than that of the IPW estimator and is nearly the same as the standard error of the full estimator. This phenomenon highlights the advantage of the AIPW estimator in that it retains much of the observed phase-one information. The simulations also indicate that the linear models with the phase-one predictors δ_{ki} , $Z_{1,ki}$, A_{ki} and $\log(X_{ki})$ track well the conditional expectations of the phase-two covariates $E_k\{Z_{2,k}(t)/\Omega_{ki}\}$, $E_k\{\exp(\beta_2 Z_{2,k}(t))/\Omega_{ki}\}$ and $E_k\{Z_{2,k}(t) \exp(\beta_2 Z_{2,k}(t))/\Omega_{ki}\}$. The pointwise coverage probabilities of the IPW and AIPW estimators for $\beta(v)$ are almost all in the range of 92.5% and 97.5% for $v \in (0, 1)$, indicating adequate performance of the proposed variance estimators for the IPW and AIPW estimators.

Robustness of the estimators to misspecification of $\pi_k(\Omega_{ki})$ is examined by assuming the model $\text{logit}(\pi_k(\Omega_{ki})) = \psi_{k0} + \psi_{k1}\delta_{ki}$ while the actual complete-case indicator ξ_{ki} was generated under model (S3). Figure 4(a) shows that the bias of the IPW estimator for $\beta_1(v)$

is large when $\pi_k(\Omega_{ki})$ is misspecified. Figures 4(a) and (b) show that the AIPW estimators have very little bias, tracking closely the full data likelihood estimator when $\pi_k(\Omega_{ki})$ is misspecified, demonstrating the double robustness property of the AIPW estimator and that the suggested linear models for the conditional expectations of the phase-two covariates work well.

6 Analysis of the RV144 HIV vaccine efficacy trial

The RV144 preventive vaccine efficacy trial randomized 16,395 HIV-1 negative volunteers to the vaccine ($n = 8198$) and placebo ($n = 8197$) groups. We apply the newly proposed methods to the vaccine group, which included 5035 men and 3163 women, with 41 vaccine recipients acquiring the primary endpoint of HIV-1 infection after the Week 26 biomarker sampling time point through to the end of follow-up at 42 months. Vaccine recipients were distributed in the Low, Medium, and High baseline behavioral risk scores (defined as in Rerks-Ngarm et al., 2009) with 3863 Low, 2370 Medium, and 1965 High.

Three HIV-1 gp120 sequences were included in the vaccine construct: 92TH023 in the ALVAC canary-pox vector prime component; and A244 and MN in the AIDSVAX recombinant glycoprotein 120 (gp120) boost component. 92TH023 and A244 are subtype E HIVs whereas MN is subtype B. The subtype E vaccine-insert sequences are much closer genetically to the infecting (and regional circulating) HIV-1 sequences than MN, and thus are more likely to stimulate protective immune responses. Accordingly, the analysis focuses on the 92TH023 and A244 insert sequences, not considering MN. The failure time T is the time from the Week 26 visit to HIV-1 infection diagnosis.

The published results cited in the Introduction suggest that the V1V2 region of gp120 may have been involved in the partial vaccine efficacy conferred by the vaccine regimen, by containing epitopes recognized by antibodies induced by the vaccine. Therefore, we investigate mark variables defined based on the genetic distance of an infecting HIV-1 V1V2 sequence to the corresponding V1V2 sequence in the vaccine construct (using a multiple sequence alignment). Two mark variables V are defined, based on the 92TH023 and A244 vaccine construct sequences. The genetic distances were computed as HIV-1-specific point accepted mutation weighted Hamming amino acid distances (Nickle et al., 2007). Between 2 and 13 HIV-1 sequences (total 1030 sequences) were measured per infected participant, and V was defined based on an infected participant's sequence that was closest to his or her consensus sequence (the consensus sequence is comprised of the majority amino acids at each amino acid site). Finally, the distances V were re-scaled to take values between 0 and 1. We refer to these two genetic distance marks as 92TH023V1V2 and A244V1V2.

The Haynes et al. (2012) study evaluated immune response biomarkers measured in vaccine recipients HIV-1 negative at Week 26 as correlates of HIV-1 infection by the Month 42 visit. The immune response biomarkers that measured levels of IgG binding antibodies to the 92TH023 and A244 V1V2 sequences are only observed for 34 of 41 HIV-1 infected vaccine recipients with HIV-1 V1V2 sequence data and for 205 of 7010 vaccine recipients who completed follow-up through Month 42 HIV-1 negative. They were found to be significantly inversely correlated with the rate of HIV-1 infection based on a Cox model and Borgan et

al.'s (2000) estimator II (Haynes et al., 2012; Zolla-Pazner et al., 2014). In addition to finding that IgG antibodies were significant correlates of HIV-1 infection, the analyses also showed that IgG3 sub-class antibodies to 92TH023 and A244 V1V2 were significant correlates of HIV-1 infection, and the hypothesis was generated that IgG3 may be markers of unmeasured functional immune responses that caused the vaccine protection (Yates et al., 2014).

However, these previous results did not take into account the V1V2 sequence of the infecting HIV-1s. Here we extend the previous case-control analysis of the IgG and IgG3 biomarkers, to study these biomarkers as correlates of 92TH023V1V2 and A244V1V2 mark-specific HIV-1 infection. In particular, paired to the 92TH023V1V2 mark variable we study the two biomarkers Week 26 IgG and IgG3 binding antibodies to 92TH023 V1V2 (which we name IgG-92TH023V1V2 and IgG3-92TH023V1V2); and, paired to the A244V1V2 mark variable, we study Week 26 IgG and IgG3 binding antibodies to A244 V1V2 (IgG-A244V1V2, IgG3-A244V1V2).

These analyses are informative because antibody responses generated by the vaccine are highest against the HIV-1 V1V2 sequences included in the vaccine (92TH023 and A244), and are lower and sometimes absent against HIV-1 sequences mismatched to the vaccine sequences in the V1V2 subprotein. As such, vaccine recipients exposed to HIVs with V1V2 sequences close to 92TH023 or A244 (small marks/genetic distances) may be more likely to be protected by antibodies than vaccine recipients exposed to HIVs with V1V2 sequences with large marks/genetic distances. Therefore, if a given antibody biomarker among the four is important for vaccine protection, its inverse correlation with risk would be expected to be strongest against HIV-1 infecting sequences with small V1V2 distances and be weakest or absent against HIV-1 infecting sequences with large V1V2 distances, and the newly proposed methods are designed to detect or refute this type of result. If this pattern is seen for some of the antibody biomarkers but not the others, then it would provide evidence for which type of antibody (IgG or IgG3) and/or which vaccine insert sequence (92TH023 or A244) was more important for the protection conferred by the vaccine. This evidence would increase insight into how the vaccine partially worked and would suggest next steps of iterative vaccine research.

We consider the following mark-specific PH model for each of the four immune response/mark pairs with $K = 1$ baseline stratum:

$$\lambda_k(t, v|z(t)) = \lambda_{k0}(t, v) \exp \{ \beta_1(v)Sex + \beta_2(v)RS_L + \beta_3(v)RS_M + B_4(v)IR \}, \quad (13)$$

where we control for the HIV-1 infection prognostic factors Sex (1=male or 0=female) and baseline behavioral risk score group RS_L (1=Low, 0=High) and RS_M (1=Medium, 0=High). Each antibody response biomarker IR was standardized to have mean 0 and variance 1. Because the unstandardized biomarkers had similar ranges, this standardization has little influence on the results. For the mark 92TH023V1V2, each of IgG-92TH023V1V2 and IgG3-92TH023V1V2 is analyzed, and for the mark A244V1V2, each of IgG-A244V1V2 and IgG3-A244V1V2 is analyzed. To illustrate the interpretation of model (13), if IgG

binding antibodies to 92TH023 are important for protection, then we would expect to see the infection risk $\lambda_k(t, v/z)$ decreasing with IR at the smallest values of v but not (or less so) at the highest values of v .

To predict the probability of observing the IR , we use a logistic regression model with $\text{logit}(P(\xi = 1|\Omega))$ a linear function of $(1, \delta, Sex, RS_L, RS_M, \delta Sex, \delta RS_L, \delta RS_M)$. To implement the AIPW method, we use linear models for $E(IR|\Omega)$ and $E(\exp(\beta_4(v)IR^{\otimes j})|\Omega)$ for $j = 0, 1, 2$, with the predictors $(1, \delta, Sex, BS_L, BS_M, \log(X), \delta Sex, \delta BS_L, \delta BS_M, \delta \log(X))$. The bandwidth formula $h = 4\sigma_v n_o^{-1/3} = 4(0.192)(41^{-1/3})$ leads to $h = 0.22$ for the 92TH023V1V2 mark and $h = 0.21$ for the A244V1V2 mark. Here σ_v is the estimated standard error of the observed marks for cases. We modified the bandwidth formula with n_o replaced by the number of cases (41) due to the very low HIV-1 infection rate.

Figures 5–8 show the estimated mark-specific log-relative risk, $\hat{\beta}(v)$, along with the 95% pointwise confidence bands for the four different immune response/mark pairs. Figures 5–8 show that for all four immune response/mark pairs, the High risk group tends to have higher risk of HIV-1 infection than the Medium and Low risk groups. In addition, men tend to have higher mark-specific risk than women for small values of the mark around <0.3 .

For the IgG-92TH023V1V2, IgG-A244V1V2, and IgG3-A244V1V2 immune responses, Figures 5, 6 and 8 show that IgG antibody responses to 92TH023V1V2 and to A244V1V2 appear to be inversely associated with corresponding 92TH023V1V2 and A244V1V2 mark-specific infection risk for marks in the lower range, but not in the upper range, and similarly for IgG3 antibody responses to A244V1V2. In contrast, for IgG3-92TH023V1V2, Figure 7 shows no association of IgG3 binding antibody responses to 92TH023V1V2 with 92TH023V1V2 mark-specific infection risk, for any values of the mark. These results support the hypotheses stated above for three of the four immune response/mark pairs, suggesting that both 92TH023 and A244 vaccine insert sequences were important for stimulating protective antibodies, and that these antibodies protected less against exposures to HIV-1 V1V2 sequences highly distant from the 92TH023 and A244 V1V2 vaccine sequences. Moreover, because IgG3 antibodies to 92TH023V1V2 did not associate with risk for any mark values, it suggests that other IgG subclasses besides type 3 induced by 92TH023 may be relevant for protection, and that A244 was more important than 92TH023 for induction of protective IgG3 antibodies, at least as a marker for an underlying causal functional immune response.

7 Discussion

The mark-specific proportional hazards model with a continuous mark has been studied to evaluate mark-specific relative risks (Sun, Gilbert and McKeague, 2009). This research studied IPW and augmented IPW estimation methods for the stratified mark-specific proportional hazards model under two-phase sampling of covariates, which apply for the general problem of missing covariates. Asymptotic properties of the estimators were established and satisfactory operating characteristics were demonstrated in simulation studies, indicating that the augmented IPW approach effectively gains back efficiency lost with the IPW approach.

This research was motivated by preventive HIV vaccine efficacy trials for assessing the association of vaccine-induced immune response biomarkers with the mark-specific incidence of acquiring HIV-1 infection, where the mark is the genetic distance of an infecting HIV-1 sequence to an HIV-1 sequence represented inside the vaccine. This research breaks new ground as the first statistical method to our knowledge that combines the analysis of pathogen-specific immune responses with pathogen sequence distance-specific infection outcomes, which is a critical area for clinical vaccine research in general for developing more efficacious vaccines.

Ideally, statistical methods would assess whether and how mark-specific vaccine efficacy varies over vaccinated subgroups defined by immune response biomarkers (measured at Week 26 in RV144). However, this problem is quite challenging because the immune response biomarker subgroups are defined by the potential biomarker outcome if assigned to receive the vaccine, and these potential outcomes are not observable in placebo recipients. Here we tackle the more tractable problem of assessing within the vaccine group alone how the immune response biomarkers associate with the mark-specific hazard rate, where the biomarkers are measured via a two-phase sampling design. This approach advantageously conducts inference on a target parameter that can be directly observed without counterfactuals and identifiability challenges. In the application section we elaborate on the scientific value of this analysis, and in Appendix A placed in the online Supplementary Material at the journal's website, we sketch a way to extend the method to assess mark-specific vaccine efficacy conditional on potential biomarker response if assigned vaccination under a simplifying assumption.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was partially supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R37AI054165 and by the Henry Jackson Foundation under Contract Number 694251. Dr. Sun's research was also partially supported by the NSF grant DMS1513072, and the Reassignment of Duties fund provided by the University of North Carolina at Charlotte. Yang's research was supported by the National Nature Science Foundation of China grant 11471086, the National Social Science Foundation of China grant 16BTJ032, the Fundamental Research Funds for the Central Universities 15JNQM019 and 21615452, the National Statistical Scientific Research Center Projects 2015LD02, the China Scholarship Council 201506785010 and Science and Technology Program of Guangzhou 2016201604030074. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

- 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, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*. 2009; 1:32–49. [PubMed: 20174455]

- Breslow NE, Lumley T. Semiparametric models and two-phase samples: Applications to Cox regression. *IMS Collections: From Probability to Statistics and Back: High-Dimensional Models and Processes*. 2013; 9:65–77.
- Chen K. Generalized case-cohort sampling. *Journal of the Royal Statistical Society, SerB*. 2001; 63:791–809.
- Gilbert PB, Sun Y. Inferences on relative failure rates in stratified mark-specific proportional hazards models with missing marks, with application to HIV vaccine efficacy trials. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*. 2015; 64:49–73.
- Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, Evans DT, Montefiori DC, Karnasuta C, Sutthent R, Liao HX, DeVico AL, Lewis GK, Williams C, Pinter A, Fong Y, Janes H, deCamp A, Huang Y, Rao M, Billings E, Karasavvas N, Robb ML, Ngauy V, de Souza MS, Paris R, Ferrari G, Bailer RT, Soderberg KA, Andrews C, Berman PW, Frahm N, De Rosa SC, Alpert MD, Yates NL, Shen X, Koup RA, Pitisuttithum P, Kaewkungwal J, Nitayaphan S, Rerks-Ngarm S, Michael NL, Kim JH. Immune-correlates analysis of an HIV vaccine efficacy trial. *New England Journal of Medicine*. 2012; 366:1275–1286. [PubMed: 22475592]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85:785–798.
- Juraska M, Gilbert PB. Mark-specific hazard ratio model with multivariate continuous marks: an application to vaccine efficacy. *Biometrics*. 2013; 69:328–337. [PubMed: 23421613]
- 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.
- Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. HIV-specific probabilistic models of protein evolution. *PLoS ONE*. 2007; 2(6):e503. [PubMed: 17551583]
- Prentice RL. A Case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34:541–554. [PubMed: 373811]
- Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 472:1250–1263.
- Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil JG, Francis DP, Stablein D, Birx DL, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, Robb ML, Michael NL, Kunasol P, Kim JH. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine*. 2009; 361:2209–2220. [PubMed: 19843557]
- 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.
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *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.
- Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. *The Annals of Statistics*. 2009; 37:394–426. [PubMed: 19838313]
- Sun Y, Li M, Gilbert PB. Goodness-of-fit of stratified proportional hazards models with continuous marks. *Computational Statistics and Data Analysis*. 2016; 93:348–358. [PubMed: 26461462]
- Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer; New York: 2006.
- van der Vaart, AW. *Asymptotic Statistics*. Cambridge University Press; Cambridge: 1998.
- Yates NL, Liao HX, Fong Y, deCamp A, Vandergift NA, Williams WT, Alam SM, Ferrari G, Yang ZY, Seaton KE, Berman PW, Alpert MD, Evans DT, O'Connell RJ, Francis D, Sinangil F, Lee C, Nitayaphan S, Rerks-Ngarm S, Kaewkungwal J, Pitisuttithum P, Tartaglia J, Pinter A, Zolla-Pazner S, Gilbert PB, Nabel GJ, Michael NL, Kim JH, Montefiori DC, Haynes BF, Tomaras GD. Vaccine-

induced env V1–V2 IgG3 correlates with lower HIV-1 infection risk and declines soon after vaccination. *Science Translational Medicine*. 2014; 6:228–239.

Zolla-Pazner S, deCamp A, Gilbert PB, Williams C, Yates NL, Williams WT, Howington R, Fong Y, Morris DE, Soderberg KA, Irene C, Reichman C, Pinter A, Parks R, Pitisuttithum P, Kaewkungwal J, Rerks-Ngarm S, Nitayaphan S, Andrews C, O’Connell RJ, Yang ZY, Nabel GJ, Kim JH, Michael NL, Montefiori DC, Liao HX, Haynes BF, Tomaras GD. Vaccine-induced IgG antibodies to V1V2 regions of multiple HIV-1 subtypes correlate with decreased risk of HIV-1 infection. *PLoS One*. 2014; 9(2):e87572. [PubMed: 24504509]

Zhou H, Wang CY. Failure time regression with continuous covariates measured with error. *Journal Royal Statistical Society B*. 2000; 62:657–665.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n=500$ and $h=0.15$ under the sampling probability model (S1) based on 1000 simulations.

Fig. 2. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n=500$ and $h=0.15$ under the sampling probability model (S2) based on 1000 simulations.

Fig. 3. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n=500$ and $h=0.15$ under the sampling probability model (S3) based on 1000 simulations.

Fig. 4. Bias, SEE, ESE and CP for $\hat{\beta}_1(v)$ and $\hat{\beta}_2(v)$ for $n=500$ and $h=0.15$ under the misspecified sampling probability model based on 1000 simulations, where $\pi_{ki}(\Omega_{ki})$ is generated under (S3) but misspecified using $\text{logit}(\pi_k(\Omega_{ki})) = \psi_{k0} + \psi_{k1}\delta_{ki}$.

Fig. 5. AIPW estimates of $\beta(v)$ and 95% pointwise confidence bands with $h = 0.22$ for IgG-92TH023V1V2 in the RV144 trial.

Fig. 6. AIPW estimates of $\beta(v)$ and 95% pointwise confidence bands with $h = 0.21$ for IgG-A244V1V2 in the RV144 trial.

Fig. 7.
AIPW estimates of $\beta(v)$ and 95% pointwise confidence bands with $h = 0.22$ for IgG3-92TH023V1V2 in the RV144 trial.

Fig. 8. AIPW estimates of $\beta(v)$ and 95% pointwise confidence bands with $h = 0.21$ for IgG3-A244V1V2 in the RV144 trial.