Published in final edited form as:

Comput Stat Data Anal. 2018 June ; 122: 59–79. doi:10.1016/j.csda.2018.01.003.

Analysis of Generalized Semiparametric Regression Models for Cumulative Incidence Functions with Missing Covariates

Unkyung Lee^a, Yanqing Sun^{b,*}, Thomas H. Scheike^c, and Peter B. Gilbert^{d,e}

^aDepartment of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

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

^cDepartment of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, DK-1014, Denmark

^dDepartment of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.

^eVaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

Abstract

The cumulative incidence function quantifies the probability of failure over time due to a specific cause for competing risks data. The generalized semiparametric regression models for the cumulative incidence functions with missing covariates are investigated. The effects of some covariates are modeled as non-parametric functions of time while others are modeled as parametric functions of time. Different link functions can be selected to add flexibility in modeling the cumulative incidence functions. The estimation procedures based on the direct binomial regression and the inverse probability weighting of complete cases are developed. This approach modifies the full data weighted least squares equations by weighting the contributions of observed members through the inverses of estimated sampling probabilities which depend on the censoring status and the event types among other subject characteristics. The asymptotic properties of the proposed estimators are established. The finite-sample performances of the proposed estimators and their relative efficiencies under different two-phase sampling designs are examined in simulations. The methods are applied to analyze data from the RV144 vaccine efficacy trial to investigate the associations of immune response biomarkers with the cumulative incidence of HIV-1 infection.

Keywords

Competing risks; Inverse probability weighted complete-case; Semiparametric regression model; RV144 vaccine efficacy trial; Time-varying effects; Two-phase sampling

*Corresponding author yasun@uncc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Competing risks data are often encountered in medical researches where study participants are exposed to two or more mutually exclusive causes of failure. Because of the inherited nonidentifiability problem with the competing risks data, the cause-specific hazard function and cumulative incidence function have been used as the primary quantities for analyzing competing risks data (Tsiatis (1978); Kalbfleisch and Prentice (1980)). The cause-specific hazard function measures the instantaneous failure risk due to a specific cause while the cumulative incidence function describes the probability distribution for a specific cause of failure, both are the measurements of the crude failure risk in the presence of all other risks. The two measures provide different perspectives for cause-specific failure times, cf. Fine and Gray (1999), Katsahian et al. (2006) and Latouche et al. (2007).

The statistical methods developed for failure time hazard regression models such as Cox (1975), Lin and Ying (1994) and Scheike and Zhang (2002) can often be utilized directly for modeling the cause-specific hazard functions. The cumulative incidence function $F_j(t)$ for cause j relates the cause-specific hazard function $\lambda_j(t)$ through the simple formula $F_j(t) = \int_0^t \lambda_j(s) S_T(s) ds$, where $S_T(t)$ is the survival function of T of all causes. Cumulative incidence functions can be estimated through modeling the cause-specific hazard functions for all causes (Cheng, Fine and Wei (1988) and Shen and Cheng (1999)). This approach provides a straightforward way to estimate cumulative incidence functions, but the effects of covariates on the cumulative incidence functions are not clear. In addition, indirect modelling $F_j(t)$ through the cause-specific hazard functions requires setting up models for the cause-specific hazard functions of all causes. Alternative approaches that directly model the cumulative incidence functions have been studied by Klein and Andersen (2005) and Scheike, Zhang and Gerds (2008). Direct modeling of the cumulative incidence functions allows direct evaluation of covariate effects on the probability of failure over time for a specific cause without the need to model for other causes.

This paper investigates the generalized semiparametric additive model for cumulative incidence functions with missing covariate values. This research is motivated by RV144, a preventive HIV-1 vaccine efficacy trial. RV144 randomized 16,395 HIV-1 negative volunteers in 1:1 allocation to receive vaccine or placebo and followed them for 42 months for occurrence of the primary study endpoint of HIV-1 infection (Rerks-Ngarm, Pitisuttithum, Nitayaphan, Kaewkungwal, Chiu, Paris, Prensri, Namwat, de Souza, Adams, Benenson, Gurunathan, Tartaglia, McNeil, Francis, Stablein, Birx, Chunsuttiwat, Khamboonruang, Thongcharoen, Robb, Michael, Kunasol and Kim (2009)), showing partial beneficial efficacy of the vaccine to lower the incidence of HIV-1 infection. An objective of RV144 is to examine the effects of certain biomarkers measuring immune responses to vaccination on the cumulative probability of becoming infected with specific genetic types of HIV-1 (the competing risks of failure). The immune response biomarkers were measured from vaccine recipients at the Week 26 visit, two weeks after the vaccination series. Since the rate of HIV-1 infection was low (fewer than 1% of participants acquired HIV), it would be very expensive and unnecessary to measure the biomarkers from all vaccine recipients. A classical case-cohort design would measure the biomarkers from all participants who

experience the event of interest after Week 26 (HIV-1 infection) and from a relatively small random sample from the original cohort of vaccine recipients (Prentice, 1986). However, since the immune response biomarkers were not measured from all failure events after Week 26, we follow the generalized case-cohort design, which is in the form of two-phase sampling data (Breslow, Lumley, Ballantyne, Chambless and Kulich (2009a), Breslow, Lumley, Ballantyne, Chambless and Kulich (2009b), Haneuse, Saegusa and Lumley (2011)), where the phase-one data are variables measured from all participants and the phase-two data are measured in random samples (without replacement or Bernoulli) of participants within each level of a stratification variable defined by the phase-one data. The case-cohort or two-phase sampling of covariate data are common forms of missing covariates. In the application section we describe the specific genetic types of HIV-1 infection that constitute the failure types of interest.

Our estimation procedure for generalized semiparametric additive models is based on the direct binomial regression model of Scheike et al. (2008) using an inverse probability weighting of complete cases (IPW) estimating equation. This approach modified the full data weighted least squares equations by weighting the contributions from participants with fully observed data by the inverses of the estimated sampling probabilities. Under the competing risks situations, the sampling probabilities may be different for different types of events. Previous related methodological research includes the following papers. Kang and Cai (2009) developed methods for fitting failure time data from case-cohort studies with multiple disease outcomes under a marginal proportional hazards model. Kang, Cai and Chambless (2013) developed a weighted estimating equations approach for the marginal additive hazards regression model for case-cohort studies with multiple disease outcomes. Sun, Qian, Shou and Gilbert (2017) studied a semiparametric additive hazards model for case-cohort and two-phase sampling data. We transport this approach to develop an IPW estimating equations procedure for the generalized semiparametric additive model in Section 2. We develop asymptotic properties of the proposed IPW estimator in Section 3. The finite-sample performances of the IPW estimators and their relative efficiencies are examined in Section 4 for different two-phase sampling designs. The proposed method is applied to analyze data from the RV144 vaccine efficacy trial in Section 5.1 to investigate the association of antibody responses with the vaccine on the cumulative incidence of infection with HIVs of specific genetic types of interest. An additional simulation study based on the RV144 data is conducted in Section 5.2 to examine the performance of the IPW method when the sample size is very large, and both the event rate and sampling percentage of phase-two variables are very small. Proofs of the asymptotic results are given in Appendix A.

2. Estimation of the Semiparametric Model for the Cumulative Incidence Function with Missing Covariates

2.1. Generalized semiparametric model and missing data

Let T_i be the failure time and $J_i \in \{1, 2, \dots, k\}$ denote the k different failure types for the i th subject. Assume that cause $J_i = 1$ is the primary cause of interest and $J_i > 1$ for other competing causes. Let $X_i = (1, X_{i1}, \dots, X_{ip})^T$ and $Z_i = (Z_{i1}, \dots, Z_{iq})^T$ be the $(p+1)$ - and q -

dimensional possibly time-dependent covariate vectors, respectively. Let C_i be the right censoring time. Let $I_i = I(T_i \leq C_i)$ be an indicator for uncensored failure time. The observed independent identically distributed (iid) competing risks data can be represented by $Y_i = (X_i, Z_i, \tilde{T}_i, \tilde{J}_i)$ for $i = 1, 2, \dots, n$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\tilde{J}_i = J_i I_i$. The value $\tilde{J}_i = J_i$ indicates that the failure time is observed at $\tilde{T}_i$ and the cause of failure is of type J_i . Let $[0, \tau]$ be the follow-up period during which data are collected. We assume that the (T_i, J_i) are independent of the C_i given the covariates (X_i, Z_i) . The covariates of subject i are only meaningful in the time interval when the subject is at-risk and still in the study, i.e., $t \leq \tilde{T}_i$.

Let $F_1(t; X_i, Z_i) = P(T_i \leq t, J_i = 1 | X_i, Z_i)$ be the conditional cumulative incidence function given covariates (X_i, Z_i) . We consider the following generalized semiparametric model:

$$F_1(t; X_i, Z_i) = \mathbf{h}\{X_i^T \boldsymbol{\eta}(t), g(\boldsymbol{\gamma}, Z_i, t)\} \quad (1)$$

where $\mathbf{h}(\cdot, \cdot)$ is a known bivariate link function, $\boldsymbol{\eta}(t)$ is a $(p+1)$ -dimensional vector of time-dependent regression coefficients, $\boldsymbol{\gamma}$ is a q -dimensional vector of time-invariant coefficients, and $g(\cdot)$ is a known function of $(\boldsymbol{\gamma}, Z_i, t)$ that can be constant or piecewise-constant parametric functions. Under model (1), the effects of the covariates X_i change over time while the effects of Z_i are time-invariant. Note that model (1) can be used to model failure times with only one type of events or without competing risks, in this case the cumulative incidence function is cumulative distribution function. Model (1) was studied by Scheike, Zhang and Gerds (2008) for fully observed covariates. The purpose of this paper is to develop an estimation procedure when some covariates of X_i may be missing.

Model (1) includes the following important submodels

$$F_1(t; X_i, Z_i) = \mathbf{h}\{X_i^T \boldsymbol{\eta}(t)g(\boldsymbol{\gamma}, Z_i, t)\}, \quad (2)$$

$$F_1(t; X_i, Z_i) = \mathbf{h}\{X_i^T \boldsymbol{\eta}(t) + g(\boldsymbol{\gamma}, Z_i, t)\}, \quad (3)$$

where $\mathbf{h}(\cdot)$ is a known monotone link function. With $\mathbf{h}(x) = 1 - \exp(-x)$ and $g(\boldsymbol{\gamma}, Z_i, t) = \exp(\boldsymbol{\gamma}^T Z_i)$, model (2) reduces to a Cox-Aalen model that contains both the Cox model and Aalen's additive model (Scheike and Zhang, 2003). Further if $X_i = 1$, model (2) is equivalent to the proportional subdistribution hazard model of Fine and Gray (1999). When $\mathbf{h}(x) = 1 - \exp(-x)$ and $g(\boldsymbol{\gamma}, Z_i, t) = \boldsymbol{\gamma}^T Z_i t$, model (3) generates a partially semiparametric additive model (McKeague and Sasieni, 1994). The additive model of Lin and Ying (1994) is a special case by letting $X_i = 1$ in model (3).

Suppose that X_i has two parts $(X_i^{(1)}, X_i^{(2)})$. The covariates $X_i^{(1)}$ and Z_i are observed for all cohort members, but $X_i^{(2)}$ is only observed for a subset (subcohort, phase-two sample) of the study cohort. Let ξ_i be the indicator of whether subject i is selected into the phase-two

sample. Let $\mathcal{V}_i = \{\tilde{T}_i, \Delta_i, \tilde{J}_i, X_i^{(1)}, Z_i\}$. We assume that the probability a subject is missing the phase-two covariates $X_i^{(2)}$ does not depend on the values of the $X_i^{(1)}$. This missing at random (MAR) assumption (Rubin, 1976) can be expressed as $P(\xi_i = 1 | X_i^{(2)}, \mathcal{V}_i) = P(\xi_i = 1 | \mathcal{V}_i)$.

However, the missingness probability may depend on any of the phase-one information $\mathcal{V}_i$.

Let $(\mathcal{V}_i, X_i^{(2)}, \xi_i)$, $i = 1, \dots, n$ be the full data iid sample under the two-phase sampling design.

The observed two-phase sampling data are $(\mathcal{V}_i, \xi_i X_i^{(2)}, \xi_i)$. That is, $\{\tilde{T}_i, \tilde{J}_i, X_i, Z_i\}$ are observed for a subject with $\xi_i = 1$, and $\{\tilde{T}_i, \Delta_i, \tilde{J}_i, X_i^{(1)}, Z_i\}$ are observed if $\xi_i = 0$. Let $\vartheta_i =$

$P(\xi_i = 1 | \mathcal{V}_i)$ be the conditional probability that the phase-two covariate $X_i^{(2)}$ is observed. This selection probability ϑ_i may depend on outcomes $\tilde{J}_i$ based on the competing causes of failure and the censoring indicator. Under the competing risks data with type-1 case of interest, the classical case-cohort design implies that $\vartheta_i = 1$ if $\tilde{J}_i = 1$ (a type-1 case) and $\vartheta_i = P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 0) = 1$ if $\tilde{J}_i = 0$ (a non-case free of the endpoint). In this paper, we consider the flexible semiparametric model for the cumulative incidence function (1) where the covariates can be missing for the cases as well as for the non-cases. In addition, we assume Bernoulli sampling of ξ within each phase-one sampling stratum.

2.2. Estimation of the selection probability and direct binomial estimation

Our estimation procedure for the generalized semiparametric model (1) is based on the direct binomial regression model of Scheike et al. (2008). We apply the approach of Horvitz and Thompson (1952) for missing data through inverse probability weighting of complete cases (IPW). This approach modifies the full data estimating equations by weighting the contributions of subjects with fully observed data by the inverses of the selection probabilities. The selection probability ϑ_i is often unknown in practice. However, by using the MAR assumption, it can be estimated based on parametric or semiparametric models. Parametric models have been widely used to model the selection probabilities.

Assume that $\varphi(\mathcal{V}_i, \theta)$ is a parametric model for $\vartheta_i = P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i)$, where θ is a r -dimensional vector of parameters belonging to a compact set Θ . The choice of $\varphi(\mathcal{V}_i, \theta)$ depends on the two-phase sampling designs. Under two-phase sampling with competing risks data, for example, the sampling probability ϑ_i may depend on $\tilde{J}_i$ which can be estimated under different models depending on whether $\tilde{J}_i$ is type-1 cases, non-type-1 cases or for the censored subjects. In this case, one may consider separate parametric models for different $\tilde{J}_i$ values such as $\varphi_1(\mathcal{V}_i, \theta_1)$ for $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 1)$, $\varphi_2(\mathcal{V}_i, \theta_2)$ for $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i > 1)$ and $\varphi_3(\mathcal{V}_i, \theta_3)$ for $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 0)$, where θ_1 , θ_2 and θ_3 are column vectors of dimensions r_1 , r_2 and r_3 , respectively. The logistic models have been commonly used to model such probabilities. Let $\theta = (\theta_1^T, \theta_2^T, \theta_3^T)^T$. We can write $\varphi(\mathcal{V}_i, \theta) = \mathbb{I}(\tilde{J}_i = 1)\varphi_1(\mathcal{V}_i, \theta_1) + \mathbb{I}(\tilde{J}_i > 1)\varphi_2(\mathcal{V}_i, \theta_2) + \mathbb{I}(\tilde{J}_i = 0)\varphi_3(\mathcal{V}_i, \theta_3)$. The maximum likelihood estimator (mle) $\hat{\theta}$ of θ is the maximizer of the following observed data likelihood $L(\theta) = L_1(\theta_1)L_2(\theta_2)L_3(\theta_3)$, where

$$L_1(\theta_1) = \prod_{i=1}^n \{\varphi_1(\mathcal{V}_i, \theta_1)\}^{\xi_i I(\tilde{J}_i = 1)} \{1 - \varphi_1(\mathcal{V}_i, \theta_1)\}^{(1 - \xi_i) I(\tilde{J}_i = 1)},$$

$$L_2(\theta_2) = \prod_{i=1}^n \{\varphi_2(\mathcal{V}_i, \theta_2)\}^{\xi_i I(\tilde{J}_i > 1)} \{1 - \varphi_2(\mathcal{V}_i, \theta_2)\}^{(1 - \xi_i) I(\tilde{J}_i > 1)}, \quad (4)$$

$$L_3(\theta_3) = \prod_{i=1}^n \{\varphi_3(\mathcal{V}_i, \theta_3)\}^{\xi_i I(\tilde{J}_i = 0)} \{1 - \varphi_3(\mathcal{V}_i, \theta_3)\}^{(1 - \xi_i) I(\tilde{J}_i = 0)}$$

are the observed data likelihoods for the type-1 cases, the non-type-1 cases and the censored, respectively. Here $I(D)$ is the indicator function of the set D .

In general, $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ can be the M-estimators each satisfying the asymptotic linearity property Condition (I.4) given in Appendix A. Let $\theta_0 = (\theta_{10}^T, \theta_{20}^T, \theta_{30}^T)^T$ be the true value of θ . Then $n_k^{1/2}(\hat{\theta}_k - \theta_{k0}) = n_k^{-1/2}[\mathcal{J}_{k, \theta}(\theta_{k0})]^{-1} \sum_{i \in D_k} U_k(\mathcal{V}_i, \theta_{k0}) + o_p(1)$, for $k = 1, 2, 3$, where $D_1 = \{\tilde{J}_i = 1\}$, $D_2 = \{\tilde{J}_i > 1\}$, and $D_3 = \{\tilde{J}_i = 0\}$, n_k is the number of observations in D_k , $\mathfrak{U}_{k, \theta}(\theta_k)$ is the positive definite Fisher information matrix and $U_k(\mathcal{V}_i, \theta_k)$ for $i \in D_k$ are independent identically distributed mean zero random variables. Thus,

$$n^{1/2}(\hat{\theta} - \theta_0) = n^{-1/2}[\mathcal{J}_{\theta}(\theta_0)]^{-1} \sum_{i=1}^n U(\mathcal{V}_i, \theta_0) + o_p(1), \quad (5)$$

where $\mathfrak{U}_{\theta}(\theta_0) = \text{diag}(\mathfrak{U}_1, \mathfrak{U}_2, \mathfrak{U}_3)$, and

$U(\mathcal{V}_i, \theta_0) = \left((n/n_1)I(i \in D_1)U_1^T(\mathcal{V}_i, \theta_{10}), (n/n_2)I(i \in D_2)U_2^T(\mathcal{V}_i, \theta_{20}), (n/n_3)I(i \in D_3)U_3^T(\mathcal{V}_i, \theta_{30}) \right)^T$. Under the standard case-cohort sampling where all phase-two covariates are sampled for $\tilde{J}_i = 1$, we take $U_1(\mathcal{V}_i, \theta_{10}) = 0$.

2.3. Inverse probability weighted complete-case estimation

Let $N_i(t) = I(T_i \leq t, J_i = 1)$ be the counting process associated with cause-1 failure and let $G_i(t|X_i, Z_i) = P(C_i > t|X_i, Z_i)$ be the conditional survival function of the censoring time for 0 $t \leq \tau$. Assume there exists a positive number $0 < \delta \leq 1$ such that $G_i(\tau|x, z) \geq \delta > 0$ for (x, z) in the range of (X_i, Z_i) . Under conditional independence between C_i and (T_i, J_i) given the covariates, we have $E\{I_i|X_i, Z_i, T_i, J_i\} = G_i(T_i|X_i, Z_i)$. Using the double expectation property, it follows that

$$E\left\{\frac{\Delta_i N_i(t)}{G(T_i|X_i, Z_i)} | X_i, Z_i\right\} = P(T_i \leq t, J_i = 1 | X_i, Z_i). \quad (6)$$

Thus, we can consider the estimating equation based on the weighted response $\delta_i N_i(t)/G(T_i|X_i, Z_i)$. The censoring distribution $G(t|X_i, Z_i)$ is unknown in practice, but it can be estimated by using a regression model for censoring times to improve efficiency. For simplicity, the Kaplan-Meier estimator $\hat{G}(t)$ is used to estimate $G(t)$.

We modify the estimating equations of Scheike et al. (2008) for the fully observed covariates with the inverse probability weighting of the complete cases. Let $D_{\eta,i}(t, \eta(t), \gamma) = F_1(t; X_i, Z_i)/\eta(t)$ and $D_{\gamma,i}(t, \eta(t), \gamma) = F_1(t; X_i, Z_i)/\gamma$. Let $\psi_{1,i}(\theta_1) = \xi_i/\phi_1(\gamma_i, \theta_1)$, $\psi_{2,i}(\theta_2) = \xi_i/\phi_2(\gamma_i, \theta_2)$ and $\psi_{3,i}(\theta_3) = \xi_i/\phi_3(\gamma_i, \theta_3)$. Let $\psi_i(\theta) = \xi_i/\phi(\gamma_i, \theta)$. Then $\psi_i(\theta) = I(\tilde{J}_i = 1)\psi_{1,i}(\theta_1) + I(\tilde{J}_i > 1)\psi_{2,i}(\theta_2) + I(\tilde{J}_i = 0)\psi_{3,i}(\theta_3)$. Following Horvitz and Thompson (1952), the cumulative regression functions $\eta(t)$ and parameters γ in model (1) can be estimated by the following IPW estimating functions:

$$U_{\eta}(t, \eta(t), \gamma, \hat{\theta}) = \sum_{i=1}^n \psi_i(\hat{\theta}) D_{\eta,i}(t, \eta(t), \gamma) w_i(t) \left(\frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - F_1(t; X_i, Z_i) \right), \quad (7)$$

$$U_{\gamma}(\tau, \eta(\cdot), \gamma, \hat{\theta}) = \sum_{i=1}^n \int_0^{\tau} \psi_i(\hat{\theta}) D_{\gamma,i}(t, \eta(t), \gamma) w_i(t) \left(\frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - F_1(t; X_i, Z_i) \right) dt, \quad (8)$$

where $w_i(t)$ is a possibly random weight function. Under the MAR assumption, it is clear that $E\{\psi_i(\theta)|\gamma_i\} = 1$.

Let $F_1(t, \eta(t), \gamma)$ be the $n \times 1$ vector of the model $F_1(t; X_i, Z_i)$ in (1) for $i = 1, \dots, n$, $R(t)$ be the $n \times 1$ vector of weighted responses $\delta_i N_i(t)/\hat{G}(T_i)$, $D_{\eta}(t, \eta(t), \gamma)$ and $D_{\gamma}(t, \eta(t), \gamma)$ be the $n \times (p+1)$ and $n \times q$ matrices with the i th rows equal to $D_{\eta,i}(t, \eta(t), \gamma)$ and $D_{\gamma,i}(t, \eta(t), \gamma)$, respectively. Let $W(t) = \text{diag}(w_i(t))$ and $\Psi(\theta) = \text{diag}(\psi_i(\theta))$. The estimating functions given in (7) and (8) are equivalent to

$$U_{\eta}(t, \eta(t), \gamma, \hat{\theta}) = (D_{\eta}(t, \eta(t), \gamma))^T W(t) \Psi(\hat{\theta}) \{R(t) - F_1(t, \eta(t), \gamma)\}, \quad (9)$$

$$U_{\gamma}(\tau, \eta(\cdot), \gamma, \hat{\theta}) = \int_0^{\tau} (D_{\gamma}(t, \eta(t), \gamma))^T W(t) \Psi(\hat{\theta}) \{R(t) - F_1(t, \eta(t), \gamma)\} dt. \quad (10)$$

Let $\mathcal{F}[0, \tau]$ be the set of uniformly bounded functions on $[0, \tau]$ and $\tilde{B} = (\mathcal{F}[0, \tau])^{p+1} \times \mathbb{R}^q$. The estimating functions $U_{\eta, \gamma}(\eta(t), \gamma, \hat{\theta}) = \{U_{\eta}(t, \eta(t), \gamma, \hat{\theta}), U_{\gamma}(\tau, \eta(\cdot), \gamma, \hat{\theta})\}$ are the mappings from $\tilde{B}$ to $\tilde{B}$. The inverse probability weighting of the complete-case estimators $\hat{\eta}(t)$ and $\hat{\gamma}$ of $\eta(t)$ and γ solve the joint estimating equations $U_{\eta, \gamma}(\hat{\eta}(t), \hat{\gamma}, \hat{\theta}) = 0$.

2.4. Numerical algorithm

The estimating equations (9) and (10) for $\hat{\eta}(t)$ and $\hat{\gamma}$ can be solved by using an iterative algorithm. Consider the following Taylor expansion of $F_1(t, \hat{\eta}(t), \hat{\gamma})$ around the values $(\eta_0(t), \gamma_0)$:

$$F_1(t, \hat{\eta}(t), \hat{\gamma}) = F_1(t, \eta_0(t), \gamma_0) + D_{\eta}(t, \eta_0(t), \gamma_0)\{\hat{\eta}(t) - \eta_0(t)\} + D_{\gamma}(t, \eta_0(t), \gamma_0)\{\hat{\gamma} - \gamma_0\} + o_p(n^{-\frac{1}{2}}). \quad (11)$$

Substituting (11) into the score equations (9) and (10), and denoting $D_{\eta}(t) = D_{\eta}(t, \eta_0(t), \gamma_0)$, $D_{\gamma}(t) = D_{\gamma}(t, \eta_0(t), \gamma_0)$ and $F_1(t) = F_1(t, \eta_0(t), \gamma_0)$, we have

$$U_{\eta}(t, \hat{\eta}(t), \hat{\gamma}, \hat{\theta}) = D_{\eta}^T(t)W(t)\Psi(\hat{\theta}) \left[R(t) - F_1(t) - D_{\eta}(t)\{\hat{\eta}(t) - \eta_0(t)\} - D_{\gamma}(t)\{\hat{\gamma} - \gamma_0\} \right] + o_p(n^{-\frac{1}{2}}), \quad (12)$$

$$U_{\gamma}(\tau, \hat{\eta}(\cdot), \hat{\gamma}, \hat{\theta}) = \int_0^{\tau} D_{\gamma}^T(t)W(t)\Psi(\hat{\theta}) \left[R(t) - F_1(t) - D_{\eta}(t)\{\hat{\eta}(t) - \eta_0(t)\} - D_{\gamma}(t)\{\hat{\gamma} - \gamma_0\} \right] dt + o_p(n^{-\frac{1}{2}}). \quad (13)$$

Solving equations (12) and (13) for $\hat{\eta}(t)$ and $\hat{\gamma}$, the estimators for η and $\gamma(t)$ are given by

$$\hat{\gamma} = \gamma_0 + \{\mathcal{J}_{\gamma}(\hat{\theta})\}^{-1} B_{\gamma}(\hat{\theta}) + o_p(n^{-\frac{1}{2}}), \quad (14)$$

$$\hat{\eta}(t) = \eta_0(t) + \{\mathcal{J}_{\eta}(t, \hat{\theta})\}^{-1} \{D_{\eta}(t)\}^T W(t) \Psi(\hat{\theta}) \left[R(t) - F_1(t) - D_{\gamma}(t) \{\mathcal{J}_{\gamma}(\hat{\theta})\}^{-1} B_{\gamma}(\hat{\theta}) \right] + o_p(n^{-\frac{1}{2}}),$$

(15)

where

$$\mathcal{J}_{\gamma}(\theta) = \int_0^{\tau} \mathbf{D}_{\gamma}^T(t) \mathbf{W}(t) \Psi(\theta) \mathbf{H}(t, \theta) \mathbf{D}_{\gamma}(t) dt,$$

$$\mathbf{B}_{\gamma}(\theta) = \int_0^{\tau} \mathbf{D}_{\gamma}^T(t) \mathbf{W}(t) \Psi(\theta) \mathbf{H}(t, \theta) \{\mathbf{R}(t) - \mathbf{F}_1(t)\} dt,$$

$$\mathbf{H}(t, \theta) = \mathbf{I} - \mathbf{D}_{\eta}(t) [\mathcal{J}_{\eta}(t, \theta)]^{-1} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\theta),$$

$$\mathcal{J}_{\eta}(t, \theta) = \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\theta) \mathbf{D}_{\eta}(t). \quad (16)$$

The estimators $\hat{\boldsymbol{\eta}}(t)$ and $\hat{\boldsymbol{\gamma}}$ can be solved iteratively similar to Scheike, Zhang and Gerds (2008), based on equations (14) and (15). Specifically, the $(m+1)$ -th iterative estimators are obtained by replacing $\hat{\boldsymbol{\gamma}}^{(m)}$ and $\hat{\boldsymbol{\eta}}^{(m)}(t)$ for $\boldsymbol{\gamma}_0$ and $\boldsymbol{\eta}_0(t)$ on the right sides of (14) and (15) as the m -th step estimators, and replacing $\hat{\boldsymbol{\gamma}}^{(m+1)}$ and $\hat{\boldsymbol{\eta}}^{(m+1)}(t)$ for $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\eta}}(t)$ on the left sides (14) and (15) as the $(m+1)$ -th step estimators.

3. Asymptotic Properties

This section investigates the asymptotic properties of the estimators $\hat{\boldsymbol{\eta}}(t)$ and $\hat{\boldsymbol{\gamma}}$. The estimating equations for $\boldsymbol{\eta}(t)$ and $\boldsymbol{\gamma}$ involve the estimators of the selection probabilities ϑ_i and the censoring survival distribution $G(t)$, whose asymptotic linearity properties are often used to show asymptotic results. Some regularity conditions for the results are stated in Condition I in Appendix A.

Let $Y_i(t) = \mathcal{Q}(\tilde{T}_i - t)$, $y(t) = P(\tilde{T}_i - t)$ and $A_i(\theta) = \psi_i(\theta)/\theta$. Let $M_i^c(t) = \mathcal{J}(\tilde{T}_i \leq t, \Delta_i = 0) - \int_0^t Y_i(s) d(-\log G(s))$. For simplicity, we denote $F_1(t; \mathbf{X}_i, \mathbf{Z}_i)$, $\mathbf{D}_{\eta,i}(t, \boldsymbol{\eta}(t), \boldsymbol{\gamma})$ and $\mathbf{D}_{\gamma,i}(t, \boldsymbol{\eta}(t), \boldsymbol{\gamma})$ by $F_1(t)$, $\mathbf{D}_{\eta,i}(t)$ and $\mathbf{D}_{\gamma,i}(t)$, respectively. Let $\mathbf{K}(t, \theta) = \mathbf{D}_{\gamma}^T(t) \mathbf{W}(t) \Psi(\theta) \mathbf{D}_{\eta}(t) [\mathcal{J}_{\eta}(t, \theta)]^{-1}$. We also introduce the following notations

$$\mathbf{k}(t, \theta) = E \left[\mathbf{D}_{\gamma,i}^T(t) w_i(t) \psi_i(\theta) \mathbf{D}_{\eta,i}(t) \right] \left[E \left[\mathbf{D}_{\eta,i}^T(t) w_i(t) \psi_i(\theta) \mathbf{D}_{\eta,i}(t) \right] \right]^{-1},$$

$$\mathbf{q}_{\gamma}(s, t, \theta) = E \left[\left[\mathbf{D}_{\gamma,j}^T(t) - \mathbf{k}(t, \theta) \mathbf{D}_{\eta,j}^T(t) \right] w_j(t) \psi_j(\theta) \frac{\Delta_j N_j(t)}{G(T_j)} \mathcal{J}(s \leq \tilde{T}_j \leq t) \right],$$

$$\xi_{\gamma,i}(t, \theta) = \left[D_{\gamma,i}^T(t) - k(t, \theta) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\},$$

$$\kappa_{\gamma,i}(t, \theta) = \left\{ \int_0^\tau \frac{q_\gamma(s, t, \theta)}{y(s)} dM_i^c(s) \right\}, \quad (17)$$

$$\mathbf{Q}_\gamma(\theta) = E \left\{ \int_a^\tau \left[D_{\gamma,i}^T(t) - k(t, \theta) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta) D_{\gamma,i}(t) dt \right\},$$

$$g(\tau, \theta) = E \left\{ \int_0^\tau \left[D_{\gamma,i}^T(t) - k(t, \theta) D_{\eta,i}^T(t) \right] w_i(t) A_i(\theta) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt \right\},$$

$$\mathbf{W}_{\gamma,i}(\tau, \theta) = \int_0^\tau \xi_{\gamma,i}(t, \theta) dt + \int_0^\tau \kappa_{\gamma,i}(t, \theta) dt + g(\tau, \theta) [\mathcal{J}_\theta(\mathcal{V}_i, \theta)]^{-1} U(\mathcal{V}_i, \theta).$$

Note that $\mathbf{A}_i(\theta) = (A_{1,i}(\theta_1), A_{2,i}(\theta_2), A_{3,i}(\theta_3))$, where $A_{1,i}(\theta_1) = I(\tilde{J}_i = 1) \psi_{1,i}(\theta_1) / \theta_1$, $A_{2,i}(\theta_2) = I(\tilde{J}_i > 1) \psi_{2,i}(\theta_2) / \theta_2$, and $A_{3,i}(\theta_3) = I(\tilde{J}_i = 0) \psi_{3,i}(\theta_3) / \theta_3$ are row vectors. Let

$$g_k(\tau, \theta) = E \left\{ \int_0^\tau \left[D_{\gamma,i}^T(t) - k(t, \theta) D_{\eta,i}^T(t) \right] w_i(t) A_{k,i}(\theta_k) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt \right\}.$$

Then $g(\tau, \theta) = (g_1(\tau, \theta), g_2(\tau, \theta), g_3(\tau, \theta))$ and the third term in $\mathbf{W}_{\gamma,i}(\tau, \theta)$ equals $g(\tau, \theta) [\mathcal{J}_\theta(\mathcal{V}_i, \theta)]^{-1} U(\mathcal{V}_i, \theta) = \sum_{k=1}^3 (n/n_k) g_k(\tau, \theta_k) [\mathcal{J}_{k,\theta}(\mathcal{V}_i, \theta_k)]^{-1} U_k(\mathcal{V}_i, \theta_k)$.

Theorem 1

Let $\varphi_1(\mathcal{V}_i, \theta_1)$, $\varphi_2(\mathcal{V}_i, \theta_2)$ and $\varphi_3(\mathcal{V}_i, \theta_3)$ be the parametric models for $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 1)$, $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i > 1)$, and $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 0)$, respectively. Under Condition I in Appendix A, $\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_\gamma)$, where $\Sigma_\gamma = \mathbf{Q}_\gamma(\theta_0)^{-1} E \{ \mathbf{W}_{\gamma,i}(\tau, \theta_0) \}^{\otimes 2} \mathbf{Q}_\gamma(\theta_0)^{-1}$.

The asymptotic covariance matrix of $\sqrt{n}(\hat{\gamma} - \gamma_0)$ can be consistently estimated by

$$\widehat{\Sigma}_\gamma = \widehat{\mathbf{Q}}_\gamma^{-1}(\hat{\theta}) n^{-1} \sum_{i=1}^n \left\{ \widehat{\mathbf{W}}_{\gamma,i}(\tau, \hat{\theta}) \right\}^{\otimes 2} \widehat{\mathbf{Q}}_\gamma^{-1}(\hat{\theta}), \quad (18)$$

where

$$\widehat{W}_{\gamma,i}(\tau, \theta) = \int_0^\tau \widehat{\xi}_{\gamma,i}(t, \theta) dt + \int_0^\tau \widehat{\kappa}_{\gamma,i}(t, \theta) dt + \widehat{g}(\tau, \theta) [\widehat{\mathcal{F}}_\theta(\mathcal{V}_i, \theta)]^{-1} \widehat{U}(\mathcal{V}_i, \theta),$$

$$\widehat{Q}_\gamma(\theta) = n^{-1} \sum_{i=1}^n \int_0^\tau \left[\widehat{D}_{\gamma,i}^T(t) - \widehat{K}(t, \theta) \widehat{D}_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta) \widehat{D}_{\gamma,i}(t) dt.$$

Here $\widehat{\xi}_{\gamma,i}(t, \theta)$, $\widehat{\kappa}_{\gamma,i}(t, \theta)$ and $\widehat{g}(\tau, \theta)$ are the empirical counterparts of $\xi_{\gamma,i}(t, \theta)$, $\kappa_{\gamma,i}(t, \theta)$ and $g(\tau, \theta)$ defined in (17). $\widehat{\mathcal{Q}}_\theta(\mathcal{V}_i, \theta)$ and $\widehat{U}(\mathcal{V}_i, \theta)$ are the empirical counterparts of $\mathcal{Q}_\theta(\mathcal{V}_i, \theta)$ and $U(\mathcal{V}_i, \theta)$, which are components of the influence function of the asymptotically linear M-estimator $\widehat{\theta}$ given in (5). $\widehat{K}(t, \widehat{\theta})$ is the empirical counterpart of $K(t, \theta)$. These estimators can be obtained by replacing θ with $\widehat{\theta}$ and replacing $F_{1i}(t)$, $D_{\eta,i}(t)$, $D_{\gamma,i}(t)$, $k(t, \theta)$, $M_i^c(t)$, $q_\gamma(s, t, \theta)$, $y(t)$ with the empirical counterparts $\widehat{F}_{1i}(t)$, $\widehat{D}_{\eta,i}(t)$, $\widehat{D}_{\gamma,i}(t)$, $\widehat{K}(t, \theta)$, $\widehat{q}_\gamma(s, t, \theta)$, $\widehat{y}(t)$, $\widehat{M}_i^c(t)$, which can be estimated by inserting the estimators $\widehat{\theta}$, $\widehat{\eta}(t)$, $\widehat{\gamma}$, $\widehat{G}(t)$.

Next, we derive the asymptotic results for $\widehat{\eta}(t)$. Let

$$\xi_{\eta,i}(t, \theta) = D_{\eta,i}^T(t) w_i(t) \psi_i(\theta) \left[\frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right],$$

$$q_\eta(s, t, \theta) = E \left\{ D_{\eta,j}^T(t) w_j(t) \psi_j(\theta) \frac{\Delta_j N_j(t)}{G(T_j)} \mathcal{I}(s \leq \widetilde{T}_j \leq t) \right\},$$

$$\kappa_{\eta,i}(t, \theta) = \left\{ \int_0^\tau \frac{q_\eta(s, t, \theta)}{y(s)} dM_i^c(s) \right\},$$

$$Q_{\eta,\gamma}(t, \theta) = E \{ D_{\eta,i}^T(t) w_i(t) \psi_i(\theta) D_{\gamma,i}(t) \}. \quad (19)$$

Define

$$W_{\eta,i}(t, \theta) = \xi_{\eta,i}(t, \theta) + \kappa_{\eta,i}(t, \theta) - Q_{\eta,\gamma}(t, \theta) \{ Q_\gamma(\theta) \}^{-1} W_{\gamma,i}(\tau, \theta), \quad (20)$$

$$Q_\eta(t, \theta) = E \{ D_{\eta,i}^T(t) w_i(t) \psi_i(\theta) D_{\eta,i}(t) \}. \quad (21)$$

Theorem 2

Let $\varphi_1(\mathcal{V}_i; \theta_1)$, $\varphi_2(\mathcal{V}_i; \theta_2)$ and $\varphi_3(\mathcal{V}_i; \theta_3)$ be the parametric models for $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 1)$, $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i > 0)$, and $P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 0)$, respectively. Under Condition I in Appendix A,

$$\sqrt{n}(\hat{\eta}(t) - \eta_0(t)) = \{Q_{\eta}(t, \theta_0)\}^{-1} n^{-\frac{1}{2}} \sum_{i=1}^n W_{\eta,i}(t, \theta_0) + o_p(1) \quad (22)$$

uniformly in $t \in [0, \tau]$.

Thus, $\sqrt{n}(\hat{\eta}(t) - \eta_0(t))$ converges weakly to a mean zero Gaussian process on $t \in [0, \tau]$. The asymptotic covariance matrix of $\sqrt{n}(\hat{\eta}(t) - \eta_0(t))$ is consistently estimated by

$$\widehat{\Sigma}_{\eta}(t) = \widehat{Q}_{\eta}^{-1}(t, \hat{\theta}) n^{-1} \sum_{i=1}^n \left\{ \widehat{W}_{\eta,i}(t, \hat{\theta}) \right\}^{\otimes 2} \widehat{Q}_{\eta}^{-1}(t, \hat{\theta}),$$

where

$$\widehat{W}_{\eta,i}(t, \theta) = \left\{ \hat{\xi}_{\eta,i}(t, \theta) + \hat{\kappa}_{\eta,i}(t, \theta) - \widehat{Q}_{\eta,\gamma}(t, \theta) \left\{ \widehat{Q}_{\gamma}(\theta) \right\}^{-1} \widehat{W}_{\gamma,i}(\tau, \theta) \right\},$$

$$\widehat{Q}_{\eta}(t, \theta) = n^{-1} \sum_{i=1}^n \widehat{D}_{\eta,i}^T(t) w_i(t) \psi_i(\theta) \widehat{D}_{\eta,i}(t).$$

Here $\hat{\xi}_{\eta,i}(t, \theta)$, $\hat{\kappa}_{\eta,i}(t, \theta)$, and $\widehat{Q}_{\eta,\gamma}(t, \theta)$ are the empirical counterparts of $\xi_{\eta,i}(t, \theta)$, $\kappa_{\eta,i}(t, \theta)$, and $Q_{\eta,\gamma}(t, \theta)$ defined in (19).

4. Simulation Studies

We conduct a simulation study to evaluate the finite-sample properties of the IPW estimators of $(\eta(t), \gamma)$ for two different models under different sampling schemes. Let X_i be Bernoulli random variables with $P(X_i = 1) = 0.5$ and let Z_i be uniform random variables on $(0, 1)$ for $i = 1, 2, \dots, n$. The covariate X_i is always observed and the phase-two covariate Z_i can be missing. Let $J_i = k$ be the types of failure with $k \in \{1, 2\}$, where the event of interest is $k = 1$. We consider the following semiparametric models for the cumulative incidence function. The first model is given by

$$F_{1i}(t | X_i, Z_i) = 1 - \exp(-\eta_0(t) \exp(\gamma_1 X_i + \gamma_2 Z_i)), \quad 0 \leq t \leq \tau, \quad (23)$$

where $\gamma_1 = -0.1$, $\gamma_2 = 0.1$, $\eta_0(t) = 0.1t$, and the second model is given by

$$F_{1i}(t|X_i, Z_i) = \frac{\exp(\eta_0(t) + \gamma_1 X_i + \gamma_2 Z_i)}{1 + \exp(\eta_0(t) + \gamma_1 X_i + \gamma_2 Z_i)}, \quad 0 \leq t \leq \tau, \quad (24)$$

where $\gamma_1 = 0.1$, $\gamma_2 = -0.1$, and $\eta_0(t) = \log\{p(t)/(1 - p(t))\}$ with $p(t) = 0.01 + 0.1\sqrt{t}$. We take $\tau = 3$. Model (23) has the link function $h(x) = 1 - \exp(-x)$ and is equivalent to the proportional subdistribution hazard model studied by Fine and Gray (1999). Model (24) is a logistic model with $h(x) = \exp(x)/(1 + \exp(x))$.

Given X_i, Z_i , the conditional probabilities of type-1 failure during the time period $[0, \tau]$ equals $F_{1i}(\tau) = F_{1i}(\tau|X_i, Z_i)$. We use a simulation framework to generate the competing risks failure times such that the conditional type-1 cumulative incidence functions follow models (23) and (24). The types of failure J_i are determined by a Bernoulli random variable with probability $P(J_i = 1|X_i, Z_i) = F_{1i}(\tau)$ and $P(J_i = 2|X_i, Z_i) = 1 - F_{1i}(\tau)$. Conditional on a failure being the type-1 failure, T_i is generated by the conditional probability $\tilde{F}_{1i}(t) = F_{1i}(t)/F_{1i}(\tau)$. The failure time of type-2 is not of the main interest and is set as $\min(U_i, \tau)$, where U_i is a uniformly distributed on $(0, 6)$ for (23) and on $(0, 5)$ for (24). Let the censoring time $C_i = \min(C_i^*, \tau)$, where C_i^* follows an exponential distribution with rate $\lambda = 0.15$ for (23) and $\lambda = 0.2$ for (24). This setting generates approximately 50% censoring of the failure time for both (23) and (24).

Our simulation framework provides a way to examine the performance of the proposed estimation procedure under the models for the cumulative incidence functions. As one reviewer pointed out, this simulation framework has limitations. The method can not be used to generate competing risks failure times from the cause-specific hazard functions. A more realistic approach requires generating bivariate type-1 and type-2 failure times (T_1, T_2) from a bivariate distribution model. The cumulative incidence function for the type-1 failure would be $P(\min(T_1, T_2) \leq t, T_1 \leq T_2)$. Although this can be done for simple situations (cf. Sun (2001); Sun, Tiwari and Zalkikar (2001)), it is quite challenging to generate the failure times of type-1 from a specific regression model such as (23) and (24).

We consider three simulation scenarios in terms of whether the phase-two sampling probabilities depend on the outcome variables $\tilde{J}_i$ and how the phase-two covariate Z_i is missing. The first sampling scenario (I) is the classical case cohort sampling design, where Z_i is sampled for all type-1 cases $\tilde{J}_i = 1$ and is subject to missingness for the non-type-1 cases, $\tilde{J}_i > 1$ or the censored $\tilde{J}_i = 0$. The second sampling scenario (II) is the generalized case-cohort sampling design, which allows the phase-two covariate Z_i to be missing for all, such as type-1 cases, non-type-1 cases, and the censored. These two scenarios are examples of two-phase sampling designs. The phase-two sampling probability in the third scenario (III) does not depend on $\tilde{J}_i$, such that the observed Z_i are a random sample from the cohort with the sampling probability depending on X_i . Let s_1, s_2 and s_3 be the probabilities of selected Z_i for the type-1 cases, for the non-type-1 cases and for censored, respectively, and let s_0 be the overall probability of selected Z_i .

First, we consider model (23). For the sampling scenario I, the selection probability $\vartheta_{1i} = P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 1) = 1$ for the cases, such that $s_1 = 1$. For non-type-1 cases $\tilde{J}_i > 1$, and for censored $\tilde{J}_i = 0$, we assume that the selection probabilities $\vartheta_{2i} = P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i > 1)$ and $\vartheta_{3i} = P(\xi_i = 1 | \mathcal{V}_i, \tilde{J}_i = 0)$ follows the logistic regression models $\text{logit}(\vartheta_{2i}) = \theta_{20} + \theta_{21}X_i + \theta_{22}\tilde{T}_i$ and $\text{logit}(\vartheta_{3i}) = \theta_{30} + \theta_{31}X_i$ yielding approximately $s_2 = 0.50$ and $s_3 = 0.30$ by choosing $\theta_2 = (0.4, -0.2, -0.3)$ and $\theta_3 = (-0.6, -0.2)$.

For the sampling scenario II, the selection probabilities are given by the following logistic regression models $\text{logit}(\vartheta_{1i}) = \theta_{10} + \theta_{11}X_i + \theta_{12}\tilde{T}_i$, $\text{logit}(\vartheta_{2i}) = \theta_{20} + \theta_{21}X_i + \theta_{22}\tilde{T}_i$ and $\text{logit}(\vartheta_{3i}) = \theta_{30} + \theta_{31}X_i$ which gives approximately $s_1 = 0.75$, $s_2 = 0.55$ and $s_3 = 0.40$ with $\theta_1 = (1.2, -0.1, -0.1)$, $\theta_2 = (0.5, -0.2, -0.1)$ and $\theta_3 = (-0.4, -0.1)$.

For the sampling scenario III, $\vartheta_i = P(\xi_i = 1 | \mathcal{V}_i)$ follows the logistic model $\text{logit}(\vartheta_i) = \theta_0 + \theta_1X_i$ yielding $\vartheta_i = 0.5$ where $\theta = (-0.1, 0.3)$. The overall selection probability for each of three sampling scenarios is about $s_0 = 0.5$.

Similar settings are applied to model (24) to yield overall selection probability $s_0 = 0.5$. For the sampling scenario I, we have approximately $s_1 = 1$, $s_2 = 0.60$ and $s_3 = 0.35$ by choosing $\theta_2 = (0.6, -0.2, -0.1)$ and $\theta_3 = (-0.6, -0.1)$. For scenario II, we have approximately $s_1 = 0.85$, $s_2 = 0.55$ and $s_3 = 0.35$ when $\theta_1 = (1.5, 0.2, 0.3)$, $\theta_2 = (-0.3, -0.1, 0.5)$ and $\theta_3 = (-0.5, -0.2)$. For the sampling scenario III, we obtain $s_1 = s_2 = s_3 = 0.5$ by choosing $\theta = (-0.1, 0.2)$. The sampling probabilities for each scenario are estimated under the correctly specified models.

The performance of the IPW estimation procedure is evaluated for sample sizes $n = 600, 700$ and 800 , and is compared with the full estimator (Full) which is the case when the values of Z_i are fully observed and with the complete-case estimator (CC) obtained by removing subjects having missing covariates Z_i . Each entry of the table is estimated based on 1000 simulation runs. For the IPW estimator, Table 1 summarizes the bias (Bias), the empirical standard deviation (ESD), the average of the estimated standard error (ASE), and the empirical coverage probability (CP) of 95% confidence intervals for γ_1 and γ_2 for models (23) and (24). The Full estimator is presented as an unattainable gold standard. The relative efficiency (RE) of the IPW estimator compared to the Full estimator is ESD of the Full estimator divided by ESD of the IPW estimator. Table 1 also includes the biases and the relative efficiencies of the IPW and CC estimators.

The performance of the Full, IPW and CC estimators of the baseline cumulative coefficient $\eta_0(t)$, $t \in [0, \tau]$ under models (23) and (24) are shown in Figures 1 and 2, respectively, for $n = 700$. In each figure, the Bias, ESD and CP for estimating $\eta_0(t)$ are compared among the three estimators. Results under the sampling scenarios I, II and III are presented moving from the left to the right.

Table 1 shows that the IPW estimators for γ_1 and γ_2 perform well for both models (23) and (24) under three different sampling scenarios. The biases of the IPW estimators of γ_1 and γ_2 are small comparable to the biases of the Full estimators. The empirical standard errors tend to decrease as the sample size increases. The averages of the estimated standard errors are close to the empirical standard errors. The coverage probabilities are very close to the 0.95

nominal level. Figure 1 shows that the biases of the IPW estimator for $\eta_0(t)$ are very small comparable to those of the Full estimator and the coverage probabilities of the IPW estimator are around the 0.95 nominal level for all three different sampling scenarios. The relative efficiencies of the IPW estimators under sampling scenario I are higher than under sampling scenario II, which in turn are higher than under sampling scenario III.

Table 1, Figure 1 and Figure 2 show that the CC estimators have the largest biases when the selection probabilities are maximally different between the type-1 cases, the non-type-1 cases and the censored subjects as in the sampling scenario I, and have smaller biases when the differences in the sampling probabilities are smaller as in the sampling scenario III.

When the cumulative incidence functions follow model (24), the pattern of performances of the Full, IPW and CC estimators are similar to that observed for model (23). The numerical algorithm described in Section 2.4 is not overly sensitive to the initial values. The proposed IPW estimators are theoretically justified for the two-phase sampling designs or missing covariates that are missing at random in general. Our simulations demonstrate validity of the IPW estimators and their relative efficiencies under different two-phase sampling designs. On the other hand, the CC estimators are obtained by applying standard methods to the two-phase sampling data, which are biased samples from the study cohort. Therefore, the CC estimators can be biased, the magnitude of the bias depends on the missing mechanism.

5. Application

5.1. Analysis of the RV144 data

The RV144 vaccine efficacy trial randomized 16,395 HIV-1 negative volunteers into the vaccine ($n = 8198$) and placebo ($n = 8197$) groups. We apply the proposed IPW estimation procedures to the vaccine group, who were assigned to receive four vaccinations at weeks 0, 4, 12 and 24. Blood samples were collected from all participants at the week 26 visit for potentially measuring biomarkers of immune response to the vaccine. Participants were monitored for 36 months after the week 26 visit for occurrence of the primary endpoint of HIV-1 infection, with 41 observed HIV-1 infections among the 8198 vaccine recipients.

The tested vaccine contained three specific HIV-1 gp120 sequences— 92TH023 in the ALVAC canarypox vector prime component administered at week 0, 4, 12, 24 and A244 and MN in the AIDSVAX protein boost component administered at week 12 and 24. The 92TH023 and A244 sequences are CRF01 AE HIV-1s whereas the MN sequence is a subtype B HIV-1. Since the CRF01 AE vaccine-insert sequences were genetically much closer to the population of HIV-1 sequences that trial participants were exposed to than MN, we expect they were more likely to have induced protective immune responses. Moreover, the A244 vaccine-insert sequence is of special interest because various analyses have supported that it may have the best induced protective V2 antibodies (e.g., Alam, Liao, Tomaras, Bonsignori, Tsao, Hwang et al. (2013)).

Accordingly, our analysis focuses on the A244 vaccine-insert sequence and on post Week 26 HIV-1 infection endpoints of vaccine recipients. The observed failure time $\tilde{T}_i$ is the

minimum of the time to HIV-1 infection diagnosis (the failure time) and the time to right censoring (study dropout, administrative censoring at 42 months, etc.).

Many papers (e.g., Haynes et al. (2012); Yates et al. (2014); Zolla-Pazner et al. (2014)) reported analyses supporting that vaccine recipients with higher week 26 levels of antibodies binding to the V1V2 portion of the HIV-1 envelope protein had a significantly lower rate of HIV-1 infection over the subsequent 36 months. This observation was made for antibodies measured to each of several specific V1V2 sequences, including to the A244 vaccine-insert sequence. In general, vaccines induce immune responses to the vaccine-insert sequences but not necessarily to sequences not inside the vaccine construct, such that vaccines tend to confer greater efficacy to prevent infection or disease with sequences matching the vaccine-insert sequences than with sequences mismatching the vaccine-insert sequences. Therefore, it is of interest to refine the analysis of the association of V1V2 antibodies with HIV-1 infection to account for the genetic type of the HIV-1 infection defined in terms of the V1V2 sequence. In particular, the theory is that if A244-induced V1V2 antibodies are a cause of vaccine efficacy, then we would expect to see that the association of V1V2 antibodies with HIV-1 infection would be stronger against infection with V1V2 sequences close to the A244 sequence than against infection with V1V2 sequences far from the A244 sequence.

To test this hypothesis, we defined the HIV-1 genetic type of interest $k = 1$ as infecting HIV-1 that is below the median V1V2 genetic distance among the set of 41 distances from HIV-1 infected vaccine recipient sequences to the A244 sequence. These genetic distances were weighted Hamming distances with weights defined in Nickle et al. (2007), computed based on a multiple alignment; the minimum, median, and maximum genetic distances were 0.0289, 0.1478, and 0.3861, respectively. We defined $k = 2$ for subjects infected with HIV-1 with genetic distance > 0.1478 .

Our analysis studies the same week 26 biomarker measuring level of IgG antibodies to the V1V2 portion of A244 that was previously studied by Yang et al. (2016); we label this week 26 biomarker IgG-A244V1V2. This biomarker was measured for 34 of 41 HIV-1 infected vaccine recipients with HIV-1 V1V2 sequence data and for a random sample of 205 vaccine recipient controls. The observed biomarker was standardized to have mean 0 and variance 1 for the analysis.

Let R_i denote the biomarker IgG-A244V1V2 for subject i . Let I_i be infection status, with value 1 if subject i is diagnosed with HIV-1 after week 26 and by 42 months, and 0 if subject i is censored over the 42-month follow-up period. The HIV-1 genetic type J_i is observed for HIV-1 infected subjects. It is unobserved for the censored subjects and in this case $\tilde{J}_i = 0$. The analysis adjusts for the amount of exposure to HIV-1 by including the phase-one covariate baseline behavioral risk score in the model, where this risk score was defined to have three levels (Low, Medium, High) in Rerks-Ngarm, Pitisuttithum, Nitayaphan, Kaewkungwal, Chiu, Paris, Prensri, Namwat, de Souza, Adams, Benenson, Gurunathan, Tartaglia, Mc-Neil, Francis, Stablein, Birx, Chunsuttiwat, Khamboonruang, Thongcharoen, Robb, Michael, Kunasol and Kim (2009). We code this risk score by two dummy indicator variables B_{1i} and B_{2i} , where $B_{1i} = 1$ if subject i is in the Low group and $B_{2i} = 1$ if subject i is

in the Medium group. The biomarker R_i is the phase-two covariate, which can be missing for both case and non-case subjects.

For each infection type-1 $k = 1, 2$, we fit model (3) with link function $h(x) = 1 - \exp(-x)$:

$$F_k(t; X_i, Z_i) = 1 - \exp(-\{\eta_{k0}(t) + \eta_{k1}(t)R_i + \gamma_{k1}B_{1i}t + \gamma_{k2}B_{2i}t\}), \quad (25)$$

where $X_i = (1, R_i)^T$, $Z_i = (B_{1i}, B_{2i})^T$ and R_i is the phase-two covariate.

Let ξ_i be the indicator that R_i is observed, and let $\vartheta_i = P(\xi_i = 1 | \mathcal{V}_{i-1})$ be the selection probability for subject i . In the two-phase sampling, R_i was measured in 83% (34 of 41) of cases and 2.5% (205 of 8157) of controls. The sampling does not depend on the biomarker R_i , the risk score or infecting HIV-1 sequences. Therefore, we apply the method using the simple logistic model $\text{logit}(\vartheta_i) = \theta_0 + \theta_1 \mathcal{V}_{i-1}$.

Our analysis yields an estimated $\text{logit}(\hat{\vartheta}_i) = \hat{\theta}_0 + \hat{\theta}_1 \mathcal{V}_{i-1}$ with $\hat{\theta}_0 = -3.658$ and $\hat{\theta}_1 = 4.987$ and standard errors 0.0707 and 0.3815, respectively. For $k = 1$, our method yields $(\hat{\gamma}_{11}, \hat{\gamma}_{12}) = (-0.00041, -0.00045)$ with standard errors (0.00060, 0.00058), yielding 2-sided p-values (0.4983, 0.4435). For $k = 2$, we have estimates $(\hat{\gamma}_{21}, \hat{\gamma}_{22}) = (-0.00089, -0.00017)$ with standard errors (0.00051, 0.00067), yielding 2-sided p-values (0.0812, 0.7961).

Figure 3 shows the estimated coefficient functions $\hat{\eta}_{k0}(t)$ and $\hat{\eta}_{k1}(t)$ with 95% pointwise confidence intervals for $k = 1$ and 2. Figure 3 (a) shows that the estimates of $\eta_{k0}(t)$ for $k = 1$ and 2 are very similar. Figure 3 (b) plots the estimates of $\eta_{k1}(t)$ for $k = 1$ and 2. Figure 3 (b) shows that higher immune response biomarker R_i (IgG-A244V1V2) is associated with lower cumulative incidence of type-1 HIV-1 ($k = 1$). In contrast, the association of R_i with type-2 HIV-1 ($k = 2$) is estimated to be around zero over all study times t .

Figure 4 compares the cumulative incidence of HIV-1 infection at the first (Q_1), second (Q_2) and third quartiles (Q_3) of the immune response R_i at each level of the behavioral risk score: Low (a), Medium (b), and High (c). We found that $Q_1 = -0.1440$, $Q_2 = 0.3369$ and $Q_3 = 0.5465$. The estimated cumulative incidence functions for $k = 1$ show that subjects with upper quartile response (Q_3) have lower cumulative incidences of HIV-1 infection than those with medium response (Q_2), which in turn have lower cumulative incidences than those with lower quartile response (Q_1); this pattern occurs for each behavioral risk group. This is consistent with the previous analysis, supporting that a higher IgG-A244V1V2 response may be associated with a lower risk of HIV-1 infection with V1V2 sequences close to A244. In addition, also consistent with the previous analysis, the estimated cumulative incidence function for IgG-A244V1V2 with $k = 2$ has no apparent difference by strong versus weak IgG-A244V1V2 response. These results support the theory stated above, lending additional evidence for the role of A244-induced V1V2 antibodies as a cause of vaccine efficacy.

The cumulative incidence depends on the prevalence of circulating HIV-1 genetic types. For example, Figure 4 (a) shows that the estimated cumulative incidence of HIV-1 infection from $k = 1$ is larger than that of HIV-1 infection from $k = 2$, while Figure 4 (b) and (c) show different patterns. One explanation for Figure 4 (a) is that, in the Low risk group, the

circulating HIV-1 viruses with shorter genetic distances to A244 ($k = 1$) may have been more prevalent and exposed more people than the circulating HIV-1 viruses with larger genetic distances to A244 ($k = 2$).

5.2. Additional simulation based on the RV144 data

This section presents an additional simulation study based on the RV144 data set which has a very large phase-one sample size, much smaller event rate and sampling percentage. We examine the estimation methods using model (25) close to the fitted model based on the RV144 data set. Specifically, we generate R_j from the estimated density of the observed R_j of the RV144 data set and generate $Z_j = (B_{1j}, B_{2j})$ from the estimated probability mass function of the risk scores. For $k = 1$, we set $\eta_{10}(t) = 0.003t$, $\eta_{11}(t) = -0.0001t$, $\gamma_{11} = -0.0004$ and $\gamma_{12} = -0.0005$. For $k = 2$, we set $\eta_{20}(t) = 0.003t$, $\eta_{21}(t) = -0.00001t$, $\gamma_{21} = -0.0009$ and $\gamma_{22} = -0.0002$. We take $\tau = 3.5$ and $n = 8198$.

The time to right censoring C_i is generated by $C_i = \min(C_i^*, \tau)$, where C_i^* follows an exponential distribution with rate $\lambda = 0.45$. This setting generates approximately 98% censoring of the failure time for (25). For the sample size $n = 8198$, the average numbers of observed failures with $J_j = 1, 2$ and for the censored subjects are about 39, 36 and 8123, respectively. Similar to the RV144 data analysis given in Section 5.1, we consider the model $\text{logit}(\vartheta_j) = \theta_0 + \theta_1 j$ for the selection probability of R_j , where $\theta_0 = -4$ and $\theta_1 = 5$. The average numbers of observed R_j for type-1 failure, type-2 failure and the censored subjects are 29, 26 and 180, respectively. The sampling probabilities for R_j are approximately 0.74 for $k = 1$, 0.73 for $k = 2$ and 0.02 for the censored subjects.

Table 2 shows the simulation results for the estimators of γ_{k1} and γ_{k2} for the IPW, Full and CC estimators under model (25) for each type-1 $k = 1$ and $k = 2$. The biases of the IPW estimators of γ_{k1} and γ_{k2} are very small. The performance of the IPW estimators of γ_{k1} and γ_{k2} are similar to the Full estimators. The averages of the estimated standard errors of the IPW estimators are close to the empirical standard errors. The coverage probabilities are very close to the 0.95 nominal level. The biases and the standard errors of the CC estimators are large compared with the Full and IPW estimators for both $k = 1$ and $k = 2$.

Figure 5 and Figure 6 show the performance of the Full, IPW and CC estimators of $\eta_{k0}(t)$ and $\eta_{k1}(t)$ for $k = 1$ and $k = 2$, respectively. Both Figures 5 and 6 show that the biases of the IPW estimators for both $\eta_{k0}(t)$ and $\eta_{k1}(t)$ are very small comparable to those of the Full estimator. The coverage probabilities of the IPW estimators for $\eta_{k0}(t)$ are around the 0.95 nominal level, and the coverage probabilities of the IPW estimators for $\eta_{k1}(t)$ are somewhat lower than the 0.95 nominal level. We note that the coverage probabilities of the Full estimators for $\eta_{k1}(t)$ are also below 0.95. One possible explanation is that the event rate and sampling percentage are too small.

6. Concluding Remarks

In this paper, we investigated generalized semiparametric regression models for cumulative incidence functions under the competing risks model when covariates may have missing values. The semiparametric model under study is flexible where one can choose different

link functions and can model time-varying effects. The method applies to situations where covariates are missing by design such as two-phase sampling designs or by happenstance, where the method is easier to justify for the former situation because its validity requires consistent estimation of the conditional probability of observing the covariates. The general case-cohort design is a two-phase sampling design where the sampling probability can be less than 1 for the cases. Under the competing risks models, the sampling probabilities can be different for different types of events. Our simulation studies demonstrated validity of the IPW estimators and examined their relative efficiencies under different two-phase sampling designs. Additional simulation based on the RV144 data shows that the CC estimator performs extremely bad when the phase-one sample size is very large, and both event rate and sampling percentage are very small.

The developed method can have broad applications since competing risks are common and the situation of missing data arises frequently. In particular, we applied the method to analyze the RV144 vaccine efficacy trial to investigate how the probability of HIV-1 infection with HIV-1s that are genetically near to the A244 vaccine strain in the V1V2 HIV-1 envelope sequence depends on the anti-V1V2 antibody response and on vaccine recipients' behavioral risks. The research provides insights into the potential beneficial effects of this antibody response for different types of infecting HIV-1 viruses. The code implementing the methods will be made available to the public at the corresponding author's website.

The IPW methods are modifications of the full data weighted least squares equations by weighting the contributions of observed members through the inverses of estimated sampling probabilities while excluding the missing observations. Under the IPW complete-case approach, the information of subjects with missing covariates may not be fully utilized. Further investigations are warranted to improve efficiency and robustness. Although the generalized semiparametric regression model (1) is very flexible, it is still of interest to develop formal goodness-of-fit tests to examine the adequacy of the model.

Acknowledgments

The authors thank the reviewers for their comments that have improved the content and presentation of the paper. The authors would also like to thank Georgia Tomaras for generating the RV144 study antibody data and Sodsai Tovanabutra, James Mullins, and Morgane Rolland for generating and analyzing the HIV sequence data. 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 contract 792087 from the Military HIV Research Program (MHRP). Unkyung Lee was also funded in part by a grant from the National Cancer Institute (T32-CA090301). Yanqing Sun was also partially supported by the NSF Grant DMS1513072, and the Reassignment of Duties fund provided by the University of North Carolina at Charlotte. The views expressed are those of the authors and should not be construed to represent the positions of the NIH, MHRP, the U.S. Army or the Department of Defense.

Appendix A

Condition I

- I.1** The regression function $\eta(t)$ is right-continuous with left-handed limits on $[0, \tau]$.

- I.2** The link function $h(\cdot)$ is three times continuously differentiable and invertible, and $h(\cdot)/x$ is bounded away from zero. The function $w_A(t)$ is a possibly random weight and uniformly convergent in $t \in [0, \tau]$.
- I.3** $\phi(\mathcal{V}_i; \theta) = I(\tilde{J}_i = 1)\phi_1(\mathcal{V}_i; \theta_1) + I(\tilde{J}_i > 1)\phi_2(\mathcal{V}_i; \theta_2) + I(\tilde{J}_i = 0)\phi_3(\mathcal{V}_i; \theta_3)$ is twice differentiable with respect to $\theta \in \Theta$ and $\phi'(\mathcal{V}_i; \theta) = d\phi(\mathcal{V}_i; \theta)/d\theta$ is uniformly bounded and bounded away from zero.
- I.4** Each $\hat{\theta}_k$ satisfies the asymptotic linearity property $n_k^{1/2}(\hat{\theta}_k - \theta_{k0}) = n_k^{-1/2}[\mathcal{J}_{k, \theta}(\theta_{k0})]^{-1} \sum_{i \in D_k} U_k(\mathcal{V}_i; \theta_{k0}) + o_p(1)$, for $k = 1, 2, 3$, where $D_1 = \{\tilde{J}_i = 1\}$, $D_2 = \{\tilde{J}_i > 1\}$, and $D_3 = \{\tilde{J}_i = 0\}$, n_k is the number of observations in D_k , $\mathfrak{L}_{k, \theta}(\theta_k)$ is the positive definite Fisher information matrix and $U_k(\mathcal{V}_i; \theta_k)$ for $i \in D_k$ are independent identically distributed mean zero random variables.

Proof of Theorem 1

Let $D_\eta(t)$, $D_\gamma(t)$, $R(t)$, $F_1(t)$, $\mathfrak{L}_\gamma(\hat{\theta})$, $B_\gamma(\hat{\theta})$, $H(t, \hat{\theta})$, $\mathfrak{L}_\eta(t, \hat{\theta})$ all be evaluated at the true values $\{\eta_0(t), \gamma_0\}$ of $\{\eta(t), \gamma\}$, where $\mathfrak{L}_\gamma(\hat{\theta})$, $B_\gamma(\hat{\theta})$, $H(t, \hat{\theta})$, $\mathfrak{L}_\eta(t, \hat{\theta})$ are corresponding terms for (16). By (14), we have

$$n^{\frac{1}{2}}(\hat{\gamma} - \gamma_0) = \{n^{-1}\mathcal{J}_\gamma(\hat{\theta})\}^{-1} n^{-\frac{1}{2}}B_\gamma(\hat{\theta}) + o_p(1). \quad (\text{A.1})$$

Let $A(\theta) = \Psi(\theta)/\theta$. The Taylor expansion of $\Psi(\hat{\theta})$ around the true value θ_0 is

$$\Psi(\hat{\theta}) = \Psi(\theta_0) + A(\theta_0)(\hat{\theta} - \theta_0) + o_p(n^{-\frac{1}{2}}). \quad (\text{A.2})$$

Let $K(t, \hat{\theta}) = D_\gamma^T(t)W(t)\Psi(\hat{\theta})D_\eta(t)[\mathcal{J}_\eta(t, \hat{\theta})]^{-1}$ evaluated at the true values $\{\eta_0(t), \gamma_0\}$. By plugging $H(t, \hat{\theta})$ in the formula for $B_\gamma(\hat{\theta})$ in (16), we have

$$\begin{aligned} n^{-\frac{1}{2}}B_\gamma(\hat{\theta}) &= n^{-\frac{1}{2}} \int_0^\tau D_\gamma^T(t)W(t)\Psi(\hat{\theta}) \left\{ I - D_\eta(t)[\mathcal{J}_\eta(t, \hat{\theta})]^{-1} D_\eta^T(t)W(t)\Psi(\hat{\theta}) \right\} \{R(t) - F_1(t) \\ &\quad \} dt = n^{-\frac{1}{2}} \int_0^\tau \left\{ D_\gamma^T(t) - K(t, \hat{\theta})D_\eta^T(t) \right\} W(t)\Psi(\hat{\theta}) \{R(t) - F_1(t)\} dt. \end{aligned}$$

(A.3)

We use the following decomposition

$$\mathbf{R}(t) - \mathbf{F}_1(t) = \left\{ \frac{\Delta_i N_i(t)}{\widehat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} + \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\}. \quad (\text{A.4})$$

By plugging (A.2) into (A.3), it follows by (A.4) that (A.3) is

$$\begin{aligned} n^{-\frac{1}{2}} \mathbf{B}_\gamma(\hat{\theta}) &= n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} dt \\ &+ n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \mathbf{A}_i(\theta_0) (\hat{\theta} - \theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} dt \\ &+ n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{\widehat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} dt \\ &+ n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \mathbf{A}_i(\theta_0) (\hat{\theta} - \theta_0) \left\{ \frac{\Delta_i N_i(t)}{\widehat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} dt + o_p(1) \end{aligned} \quad (\text{A.5})$$

Denote the first term of (A.5) by

$$\tilde{\mathbf{B}}_\gamma(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} dt.$$

Note that

$$\mathbf{K}(t, \theta_0) = \left[\frac{1}{n} \sum_{i=1}^n \mathbf{D}_{\gamma,i}^T(t) w_i(t) \boldsymbol{\psi}_i(\theta_0) \mathbf{D}_{\eta,i}(t) \right] \left[\frac{1}{n} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) \boldsymbol{\psi}_i(\theta_0) \mathbf{D}_{\eta,i}(t) \right]^{-1}.$$

By the law of large numbers, we have $\mathbf{K}(t, \theta_0) \xrightarrow{P} \mathbf{k}(t, \theta_0)$. It follows that $\mathbf{K}(t, \hat{\theta}) \xrightarrow{P} \mathbf{k}(t, \theta_0)$. It also follows that

$$n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{K}(t, \hat{\theta}) - \mathbf{k}(t, \theta_0) \right] \mathbf{D}_{\eta,i}^T(t) w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} dt = o_p(1). \quad (\text{A.6})$$

Let

$$\xi_{\gamma,i}(t, \theta) = \left[D_{\gamma,i}^T(t) - k(t, \theta) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\}. \quad (\text{A.7})$$

It follows by (A.6) and (A.7) that

$$\begin{aligned} \tilde{B}_{\gamma}(\hat{\theta}) &= n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - k(t, \theta_0) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt \\ &\quad - n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[K(t, \hat{\theta}) - k(t, \theta_0) \right] D_{\eta,i}^T(t) w_i(t) \psi_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \xi_{\gamma,i}(t, \theta_0) dt + o_p(1). \end{aligned} \quad (\text{A.8})$$

Denote the second term of (A.5) by

$$D_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - K(t, \hat{\theta}) D_{\eta,i}^T(t) \right] w_i(t) A_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt (\hat{\theta} - \theta_0).$$

By a similar argument as used in (A.6), we have

$$D_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - k(t, \theta_0) D_{\eta,i}^T(t) \right] w_i(t) A_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt (\hat{\theta} - \theta_0) + o_p(1).$$

By the law of large numbers, we have

$$n^{-1} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - k(t, \theta_0) D_{\eta,i}^T(t) \right] w_i(t) A_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - F_{1i}(t) \right\} dt \xrightarrow{p} g(\tau, \theta_0). \quad (\text{A.9})$$

It follows by (A.9) and (I.4) in Condition I that

$$\begin{aligned} D_{\gamma}(\hat{\theta}) &= n^{\frac{1}{2}} (\hat{\theta} - \theta_0) (g(\tau, \theta_0) + o_p(1)) + o_p(1) = g(\tau, \theta_0) n^{\frac{1}{2}} (\hat{\theta} - \theta_0) + o_p(1) \\ &= g(\tau, \theta_0) \left\{ n^{-\frac{1}{2}} \mathcal{J}_{\theta}^{-1}(\mathcal{V}_i, \theta_0) \sum_{i=1}^n U(\mathcal{V}_i, \theta_0) \right\} + o_p(1). \end{aligned} \quad (\text{A.10})$$

Now consider the third term of (A.5). Let

$$\Delta_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - K(t, \hat{\theta}) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta_0) \Delta_i N_i(t) \left[\frac{G(T_i) - \hat{G}(T_i)}{\hat{G}(T_i) G(T_i)} \right] dt.$$

Using similar arguments as in (A.6), we have

$$\begin{aligned} \Delta_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^{\tau} \left[D_{\gamma,i}^T(t) - k(t, \theta_0) D_{\eta,i}^T(t) \right] w_i(t) \psi_i(\theta_0) \Delta_i N_i(t) & \quad (\text{A.11}) \\ \left[\frac{G(T_i) - \hat{G}(T_i)}{\hat{G}(T_i) G(T_i)} \right] dt + o_p(1). \end{aligned}$$

Using the martingale representation Theorem 3.2.3 and Corollary 3.2.1 (Fleming and Harrington, 1991), we have

$$n^{\frac{1}{2}} \Delta_i N_i(t) \frac{\hat{G}(T_i) - G(T_i)}{\hat{G}(T_i) G(T_i)} = n^{-\frac{1}{2}} \Delta_i N_i(t) \frac{-\mathcal{J}(\tilde{T}_i \leq t)}{G(T_i)} \sum_{j=1}^n \int_0^{\tau} \mathcal{J}(s \leq \tilde{T}_i) \frac{dM_j^c(s)}{y(s)} + o_p(1), \quad (\text{A.12})$$

12)

where $n^{-1} Y_{\bullet}(s) \xrightarrow{P} y(s)$ as $n \xrightarrow{P} \infty$ with $Y_{\bullet}(s) = \sum_{i=1}^n Y_i(s) = \sum_{i=1}^n \mathcal{J}(\tilde{T}_i \geq s)$, and $M_j^c(s) = \mathcal{J}(\tilde{T}_j \leq s, \Delta_j = 0) - \int_0^s Y_j(u) d(-\log G(u))$ is the martingale associated with the censoring time. By plugging (A.12) into (A.11), we have

$$\begin{aligned} \Delta_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{j=1}^n \int_0^{\tau} \int_0^{\tau} n^{-1} \sum_{i=1}^n \left[D_{\gamma,i}^T(t) - k(t, \theta_0) D_{\eta,i}^T(t) \right] w_i(t) & \\ \psi_i(\theta_0) \frac{\Delta_i N_i(t)}{G(T_i)} \mathcal{J}(s \leq \tilde{T}_i \leq t) dt \frac{dM_j^c(s)}{y(s)} + o_p(1). \end{aligned}$$

By the law of large numbers,

$$n^{-1} \sum_{j=1}^n \left[D_{\gamma,j}^T(t) - k(t, \theta_0) D_{\eta,j}^T(t) \right] w_j(t) \psi_j(\theta_0) \frac{\Delta_j N_j(t)}{G(T_j)} \mathcal{J}(s \leq \tilde{T}_j \leq t) \xrightarrow{P} q_{\gamma}(s, t, \theta_0).$$

Let

$$\kappa_{\gamma,i}(t, \theta) = \int_0^\tau \frac{q_{\gamma}(s, t, \theta)}{y(s)} dM_i^c(s). \quad (\text{A.13})$$

Thus, we have

$$\Delta_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \kappa_{\gamma,i}(t, \theta_0) dt + o_p(1). \quad (\text{A.14})$$

Applying a similar technique used in proving (A.8) and (A.14), we can show that the fourth term of (A.5) is

$$n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \mathbf{A}_i(\theta_0) (\hat{\theta} - \theta_0) \left\{ \frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} dt = o_p(1). \quad (\text{A.15})$$

Let

$$\mathbf{W}_{\gamma,i}(\tau, \theta) = \int_0^\tau \xi_{\gamma,i}(t, \theta) dt + \int_0^\tau \kappa_{\gamma,i}(t, \theta) dt + g(\tau, \theta) [\mathcal{J}_{\theta}(\mathcal{V}_i, \theta)]^{-1} U(\mathcal{V}_i, \theta).$$

It follows by (A.5), (A.8), (A.10), (A.14) and (A.15) that

$$n^{-\frac{1}{2}} \mathbf{B}_{\gamma}(\hat{\theta}) = n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\gamma,i}(\tau, \theta_0) + o_p(1). \quad (\text{A.16})$$

By plugging the expression of $\mathbf{H}(t, \hat{\theta})$ into $\mathfrak{Q}_{\gamma}(\hat{\theta})$ from (16), it follows that

$$n^{-1} \mathcal{J}_{\gamma}(\hat{\theta}) = n^{-1} \sum_{i=1}^n \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{K}(t, \hat{\theta}) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \boldsymbol{\psi}_i(\hat{\theta}) \mathbf{D}_{\gamma,i}(t) dt. \quad (\text{A.17})$$

Since $\hat{\theta} \xrightarrow{P} \theta_0$ and $\mathbf{K}(t, \hat{\theta}) \xrightarrow{P} \mathbf{k}(t, \theta_0)$, it follows that $n^{-1} \mathcal{J}_{\gamma}(\hat{\theta}) \xrightarrow{P} \mathcal{Q}_{\gamma}(\theta_0)$, where

$$\mathcal{Q}_{\gamma}(\theta) = E \left\{ \int_0^\tau \left[\mathbf{D}_{\gamma,i}^T(t) - \mathbf{k}(t, \theta) \mathbf{D}_{\eta,i}^T(t) \right] w_i(t) \boldsymbol{\psi}_i(\theta) \mathbf{D}_{\gamma,i}(t) dt \right\}.$$

By the above arguments on (A.1), (A.16) and (A.17), $\sqrt{n}(\hat{\gamma} - \gamma_0)$ is asymptotically equivalent to the following identically independent distributed decomposition

$$n^{\frac{1}{2}}(\hat{\gamma} - \gamma_0) = \{Q_{\gamma}(\theta_0)\}^{-1} n^{-\frac{1}{2}} \sum_{i=1}^n \{W_{\gamma,i}(\tau, \theta_0)\} + o_p(1). \quad (\text{A.18})$$

It is clear that $n^{-1} \sum_{i=1}^n W_{\gamma,i}(\tau, \theta_0)$ has mean zero. By the central limit theorem and Slutsky's theorem, we have

$$\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\gamma}), \quad (\text{A.19})$$

where $\Sigma_{\gamma} = Q_{\gamma}(\theta_0)^{-1} E \{W_{\gamma,i}(\tau, \theta_0)\}^{\otimes 2} Q_{\gamma}(\theta_0)^{-1}$.

Proof of Theorem 2

From (15), we have

$$\begin{aligned} \sqrt{n}(\hat{\eta}(t) - \eta_0(t)) &= \left[n^{-1} \mathcal{J}_{\eta}(t, \hat{\theta}) \right]^{-1} n^{-\frac{1}{2}} D_{\eta}^T(t) W(t) \Psi(\hat{\theta}) \left\{ R(t) - F_1(t) \right. \\ &\quad \left. - D_{\gamma}(t) \{ \mathcal{J}_{\gamma}(\hat{\theta}) \}^{-1} B_{\gamma}(\hat{\theta}) \right\} + o_p(1). \end{aligned} \quad (\text{A.20})$$

The expression $n^{-\frac{1}{2}} D_{\eta}^T(t) W(t) \Psi(\hat{\theta}) \left\{ R(t) - F_1(t) - D_{\gamma}(t) \{ \mathcal{J}_{\gamma}(\hat{\theta}) \}^{-1} B_{\gamma}(\hat{\theta}) \right\}$ can be decomposed into two terms

$$n^{-\frac{1}{2}} D_{\eta}^T(t) W(t) \Psi(\hat{\theta}) \{ R(t) - F_1(t) \} - n^{-\frac{1}{2}} D_{\eta}^T(t) W(t) \Psi(\hat{\theta}) D_{\gamma}(t) \{ \mathcal{J}_{\gamma}(\hat{\theta}) \}^{-1} B_{\gamma}(\hat{\theta}). \quad (\text{A.21})$$

The first term of (A.21) can be decomposed by using (A.4):

$$\begin{aligned}
n^{-\frac{1}{2}} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\hat{\theta}) \{ \mathbf{R}(t) - \mathbf{F}_1(t) \} &= n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} \\
&+ n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) \boldsymbol{\psi}_i(\theta_0) \left\{ \frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} \\
&+ n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) (\boldsymbol{\psi}_i(\hat{\theta}) - \boldsymbol{\psi}_i(\theta_0)) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} \\
&+ n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) (\boldsymbol{\psi}_i(\hat{\theta}) - \boldsymbol{\psi}_i(\theta_0)) \left\{ \frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} + o_p(1).
\end{aligned} \quad (\text{A.22})$$

Applying Lemma A.1 in Lin and Ying (2001), we have

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) (\boldsymbol{\psi}_i(\hat{\theta}) - \boldsymbol{\psi}_i(\theta_0)) \left\{ \frac{\Delta_i N_i(t)}{G(T_i)} - \mathbf{F}_{1i}(t) \right\} = o_p(1), \quad (\text{A.23})$$

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{D}_{\eta,i}^T(t) w_i(t) (\boldsymbol{\psi}_i(\hat{\theta}) - \boldsymbol{\psi}_i(\theta_0)) \left\{ \frac{\Delta_i N_i(t)}{\hat{G}(T_i)} - \frac{\Delta_i N_i(t)}{G(T_i)} \right\} = o_p(1). \quad (\text{A.24})$$

It follows that the first term of (A.21) is

$$n^{-\frac{1}{2}} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\hat{\theta}) \{ \mathbf{R}(t) - \mathbf{F}_1(t) \} = n^{-\frac{1}{2}} \sum_{i=1}^n \boldsymbol{\zeta}_{\eta,i}(t, \theta_0) + n^{-\frac{1}{2}} \sum_{i=1}^n \boldsymbol{\kappa}_{\eta,i}(t, \theta_0) + o_p(1), \quad (\text{A.25})$$

where $\boldsymbol{\zeta}_{\eta,i}(t, \theta)$ and $\boldsymbol{\kappa}_{\eta,i}(t, \theta)$ are defined in (19).

Now we consider the second term of (A.21). Note that

$$n^{-1} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\hat{\theta}) \mathbf{D}_{\gamma}(t) \xrightarrow{P} \mathbf{Q}_{\eta,\gamma}(t, \theta_0),$$

where $\mathbf{Q}_{\eta,\gamma}(t, \theta)$ is defined in (19). By the proof of Theorem 1, we have

$$n^{-\frac{1}{2}} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\hat{\theta}) \mathbf{D}_{\gamma}(t) \left\{ \mathcal{J}_{\gamma}(\hat{\theta}) \right\}^{-1} \mathbf{B}_{\gamma}(\hat{\theta}) = \mathbf{Q}_{\eta, \gamma}(t, \theta_0) \left\{ \mathbf{Q}_{\gamma}(\theta_0) \right\}^{-1} \left[n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\gamma, i}(\tau, \theta_0) \right] + o_p(1).$$

(A.26)

Let

$$\mathbf{W}_{\eta, i}(t, \theta) = \zeta_{\eta, i}(t, \theta) + \kappa_{\eta, i}(t, \theta) - \mathbf{Q}_{\eta, \gamma}(t, \theta) \left\{ \mathbf{Q}_{\gamma}(\theta) \right\}^{-1} \mathbf{W}_{\gamma, i}(\tau, \theta).$$

It follows by (A.21), (A.25) and (A.26) that

$$n^{-\frac{1}{2}} \mathbf{D}_{\eta}^T(t) \mathbf{W}(t) \Psi(\hat{\theta}) \left\{ \mathbf{R}(t) - \mathbf{F}_1(t) - \mathbf{D}_{\gamma}(t) \left\{ \mathcal{J}_{\gamma}(\hat{\theta}) \right\}^{-1} \mathbf{B}_{\gamma}(\hat{\theta}) \right\} = n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\eta, i}(t, \theta_0) + o_p(1).$$

(A.27)

By (16), we have $\mathcal{J}_{\eta}(t, \hat{\theta}) = \sum_{i=1}^n \mathbf{D}_{\eta, i}^T(t) w_i(t) \psi_i(\hat{\theta}) \mathbf{D}_{\eta, i}(t)$. With similar arguments as made for Lemma 1 of Sun et al. (2017), since $\hat{\theta} \xrightarrow{P} \theta_0$ and by applying the Glivenko-Cantelli Theorem (Theorem 19.4 of Van der Vaart (2000)), we have $n^{-1} \mathcal{J}_{\eta}(t, \hat{\theta}) \xrightarrow{P} \mathbf{Q}_{\eta}(t, \theta_0)$ uniformly in $t \in [0, \tau]$ where $\mathbf{Q}_{\eta}(t, \theta) = E\{\mathbf{D}_{\eta, i}^T(t) w_i(t) \psi_i(\theta) \mathbf{D}_{\eta, i}(t)\}$. Hence,

$$\sqrt{n}(\hat{\eta}(t) - \eta_0(t)) = \left\{ \mathbf{Q}_{\eta}(t, \theta_0) \right\}^{-1} n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\eta, i}(t, \theta_0) + o_p(1). \quad (\text{A.28})$$

By the functional central limit theorem for empirical process, $n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{W}_{\eta, i}(t, \theta_0)$ converges in distribution to a normal random vector with zero-mean and covariance matrix $E\{\mathbf{W}_{\eta, i}(t, \theta_0)\}^{\otimes 2}$. By Slutsky's theorem and an application of Theorem 19.5 of Van der Vaart (2000), $\sqrt{n}(\hat{\eta}(t) - \eta_0(t))$ converges weakly to a mean zero Gaussian process on $t \in [0, \tau]$ with covariance matrix $\Sigma_{\eta}(t) = \mathbf{Q}_{\eta}^{-1}(t, \theta_0) E\{\mathbf{W}_{\eta, i}(t, \theta_0)\}^{\otimes 2} \mathbf{Q}_{\eta}^{-1}(t, \theta_0)$.

References

- Alam S, Liao H, Tomaras G, Bonsignori M, Tsao C, Hwang K, et al. Antigenicity and immunogenicity of RV144 vaccine AIDSVAX clade E envelope immunogen is enhanced by a gp120 N-terminal deletion. *Journal of Virology*. 2013; 87:1554–1568. [PubMed: 23175357]
- 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. *Statistical Biosciences*. 2009a; 1:32–49.
- 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]
- Cheng S, Fine J, Wei L. Prediction of cumulative incidence function under the proportional hazards model. *Biometrics*. 1988; 54:219–228.
- Cox D. Partial likelihood. *Biometrika*. 1975; 62:269–276.
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999; 94:496–509.
- Fleming, T., Harrington, D. *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc; New York: 1991.
- Haneuse S, Saegusa T, Lumley T. osDesign: An R package for the analysis, evaluation, and design of two-phase and case-control studies. *Journal of Statistical Software*. 2011; 43
- Haynes B, Gilbert P, McElrath M, Zolla-Pazner S, Tomaras G, Alam S, Evans D, Montefiori D, Karnasuta C, Sutthent R, Liao H, DeVico A, Lewis G, Williams C, Pinter A, Fong Y, Janes H, deCamp A, Huang Y, Rao M, Billings E, Karasavvas N, Robb M, Ngauy V, de Souza M, Paris R, Ferrari G, Bailer R, Soderberg K, Andrews C, Berman P, Frahm N, De Rosa S, Alpert M, Yates N, Shen X, Koup R, Pitisuttithum P, Kaewkungwal J, Nitayaphan S, Rerks-Ngarm S, Michael N, Kim J. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *New England Journal of Medicine*. 2012; 366:1275–1286. [PubMed: 22475592]
- Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Kalbfleisch, J., Prentice, R. *The Statistical Analysis of Failure Time Data*. Wiley; New York: 1980.
- Kang S, Cai J. Marginal additive hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*. 2009; 94:887–901.
- 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]
- Katsahian S, Resche-Rigon M, Chevret S, Porcher R. Analysing multicentre competing risks data with a mixed proportional hazards model for the subdistribution. *Statist. Med*. 2006; 25:4267–4278.
- Klein JP, Andersen P. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrika*. 2005; 61:223–229.
- Latouche A, Boisson V, Chevret S, Porcher R. Misspecified regression model for the subdistribution hazard of a competing risk. *Statist. Med*. 2007; 26:965–974.
- Lin D, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika*. 1994; 81:61–71.
- Lin D, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*. 2001; 96:103–126.
- McKeague I, Sasieni P. A partly parametric additive risk model. *Biometrika*. 1994; 81:501–514.
- Nickle DC, L H, 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 R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Prensri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil J, Francis D, Stablein D, Birx D, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, Robb M, Michael N, Kunasol P,

- Kim J. 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.
- Scheike T, Zhang M. An additive-multiplicative Cox-Aalen regression model. *Scandinavian Journal of Statistics*. 2002; 29:75–88.
- Scheike T, Zhang M. Extensions and applications of the Cox-Aalen survival model. *Biometrics*. 2003; 59:1036–1045. [PubMed: 14969483]
- Scheike TH, Zhang MJ, Gerds T. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*. 2008; 95:205–220.
- Shen Y, Cheng S. Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics*. 1999; 55:1093–1100. [PubMed: 11315053]
- Sun Y. Generalized nonparametric test procedures for comparing multiple cause-specific hazard rates. *Journal of Nonparametric Statistics*. 2001; 13:171–207.
- Sun Y, Qian X, Shou Q, Gilbert P. Analysis of two-phase sampling data with semiparametric additive hazards models. *Lifetime Data Analysis*. 2017; 23:377–399. [PubMed: 26995733]
- Sun Y, Tiwari R, Zalkikar J. Goodness-of-fit tests for multiplicative intensity models with application to competing risks data. *Scandinavian Journal of Statistics*. 2001; 28:241–256.
- Tsiatis AA. An example of nonidentifiability in competing risks. *Scandinavian Actuarial Journal*. 1978; 1978:235–239.
- Van der Vaart, AW. *Asymptotic Statistics*. Cambridge University Press; 2000.
- Yang G, Sun Y, Qi L, Gilbert P. Estimation of stratified mark-specific proportional hazards models under two-phase sampling with application to HIV vaccine efficacy trials. *Statistics in Biosciences*. 2016; :1–25.doi: 10.1007/s12561-016-9177-5
- Yates N, Liao H, Fong Y, deCamp A, Vandergrift N, Williams W, Alam S, Ferrari G, Yang Z, Seaton K, Berman P, Alpert M, Evans D, OConnell R, Francis D, Sinangil F, Lee C, Nitayaphan S, Rerks-Ngarm S, Kaewkungwal J, Pitisuttithum P, Tartaglia J, Pinter A, Zolla-Pazner S, Gilbert P, Nabel G, Michael N, Kim J, Montefiori D, Haynes B, Tomaras G. Vaccine-induced env V1-V2 IgG3 correlates with lower HIV-1 infection risk and declines soon after vaccination. *Science Translational Medicine*. 2014; 6:228ra39.
- Zolla-Pazner S, deCamp A, Gilbert P, Williams C, Yates N, Williams W, Howington R, Fong Y, Morris D, Soderberg K, Irene C, Reichman C, Pinter A, Parks R, Pitisuttithum P, Kaewkungwal J, Rerks-Ngarm S, Nitayaphan S, Andrews C, OConnell R, Yang Z, Nabel G, Kim J, Michael N, Montefiori D, Liao H, Haynes B, Tomaras G. 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]

Figure 1.

Comparison of the Full, IPW, CC estimators for $\eta_0(t)$ under model (23) with $n = 700$ based on 1000 simulations under sampling designs I, II and III moving from the left panels to the right panels. Figures (a), (b) and (c) show biases of the estimates under designs I, II and III, respectively. Similarly, Figures (d), (e) and (f) show empirical standard errors of the estimates and Figures (g), (h) and (i) show coverage probabilities of 95% confidence intervals.

Figure 2.

Comparison of the Full, IPW, CC estimators for $\eta_0(t)$ under model (24) with $n = 700$ based on 1000 simulations under sampling designs I, II and III moving from the left panels to the right panels. Figures (a), (b) and (c) show biases of the estimates under designs I, II and III, respectively. Similarly, Figures (d), (e) and (f) show empirical standard errors of the estimates and Figures (g), (h) and (i) show coverage probabilities of 95% confidence intervals.

Figure 3.

Estimated coefficient functions under model (25) for $k=1$ (an HIV-1 near the A244 vaccine-insert) and for $k=2$ (an HIV-1 far from the A244 vaccine-insert). Figures (a) shows the IPW estimate of $\eta_{k0}(t)$ and (b) shows the IPW estimate of $\eta_{k1}(t)$ for the antibody response biomarker R_i (IgG-A244V1V2), along with 95% pointwise confidence intervals.

Figure 4.

Estimated cumulative incidence functions under model (25) for $k = 1$ (an HIV-1 near the A244 vaccine-insert) and for $k = 2$ (an HIV-1 far from the A244 vaccine-insert). Figures (a), (b) and (c) show $\hat{F}_k(t|X_i, Z_i)$ at the first (Q_1), second (Q_2) and third quartile (Q_3) of observed antibody response biomarker R_i (IgG-A244V1V2) at each level of behavioral risk score groups (Low, Medium, High), respectively.

Figure 5.

Comparison of the Full, IPW, CC estimators for $\eta_{k0}(t)$ and $\eta_{k1}(t)$ for the RV144 data simulation under (25) for $k = 1$ with $s_0 = 0.03$, $n = 8198$ and approximately 98% censoring based on 500 simulations. Figures (a), (b) show biases of the estimates, Figures (c) and (d) show empirical standard errors of the estimates and Figures (e), (f) show coverage probabilities of 95% confidence intervals for $\eta_{k0}(t)$ and $\eta_{k1}(t)$.

Figure 6.

Comparison of the Full, IPW, CC estimators for $\eta_{k0}(t)$ and $\eta_{k1}(t)$ for the RV144 data simulation under (25) for $k=2$ with $s_0 = 0.03$, $n = 8198$ and approximately 98% censoring based on 500 simulations. Figures (a), (b) show biases of the estimates, Figures (c) and (d) show empirical standard errors of the estimates and Figures (e), (f) show coverage probabilities of 95% confidence intervals for $\eta_{k0}(t)$ and $\eta_{k1}(t)$.

Table 1

Comparison of the Full, IPW and CC estimators for γ_1 , γ_2 under models (23) and (24) for three different sampling scenarios. The overall selection probability is about $s_0 = 0.50$. Each entry is based on 1000 simulations.

Sampling scenario	n	IPW(γ_1)				IPW(γ_2)				CC				Full		IPW		CC	
		Bias	ESD	ASE	CP	Bias	ESD	ASE	CP	Bias(γ_1)	Bias(γ_2)	Bias(γ_1)	Bias(γ_2)	RE(γ_1)	RE(γ_2)	RE(γ_1)	RE(γ_2)		
		Cumulative incidence function model (23)																	
I	600	0.004	0.195	0.196	0.963	0.004	0.392	0.380	0.951	0.107	-0.015	0.003	-0.006	0.984	0.839	0.992	0.981		
	700	-0.008	0.178	0.182	0.953	-0.005	0.361	0.351	0.944	0.100	-0.020	-0.008	-0.005	0.989	0.854	1.011	0.990		
	800	0.008	0.173	0.171	0.956	-0.004	0.339	0.328	0.951	0.108	-0.018	0.008	-0.007	0.989	0.869	1.005	1.000		
II	600	0.003	0.196	0.195	0.953	-0.020	0.405	0.406	0.960	0.057	-0.021			0.979	0.814	0.864	0.857		
	700	-0.008	0.179	0.183	0.954	-0.008	0.381	0.376	0.956	0.041	-0.012			0.986	0.810	0.854	0.853		
	800	0.008	0.174	0.173	0.952	-0.013	0.367	0.351	0.944	0.050	-0.019			0.982	0.804	0.869	0.853		
III	600	0.001	0.264	0.267	0.956	-0.017	0.479	0.465	0.959	0.004	-0.016			0.729	0.687	0.728	0.691		
	700	-0.010	0.251	0.248	0.953	0.001	0.421	0.430	0.959	-0.008	-0.000			0.703	0.732	0.703	0.736		
	800	0.006	0.249	0.232	0.945	-0.001	0.418	0.403	0.951	0.008	-0.003			0.689	0.705	0.688	0.710		
Cumulative incidence function model (24)																			
I	600	0.000	0.244	0.232	0.940	0.009	0.465	0.454	0.945	0.102	0.013	-0.001	0.003	0.986	0.902	0.936	0.934		
	700	0.004	0.224	0.216	0.948	-0.013	0.444	0.418	0.941	0.106	-0.015	0.001	-0.019	0.991	0.880	0.938	0.913		
	800	0.005	0.202	0.203	0.957	-0.016	0.399	0.391	0.947	0.104	-0.016	0.004	-0.020	0.983	0.899	0.903	0.923		
II	600	-0.002	0.247	0.217	0.918	0.001	0.489	0.475	0.952	0.086	0.003			0.976	0.858	0.863	0.885		
	700	0.004	0.228	0.205	0.921	-0.033	0.457	0.440	0.949	0.106	-0.029			0.974	0.854	0.859	0.877		
	800	0.005	0.201	0.193	0.946	-0.028	0.419	0.411	0.955	0.113	-0.029			0.989	0.858	0.845	0.885		
III	600	-0.008	0.356	0.338	0.945	-0.004	0.622	0.584	0.955	-0.005	-0.004			0.676	0.674	0.676	0.676		
	700	0.003	0.311	0.311	0.949	-0.022	0.548	0.537	0.943	0.005	-0.024			0.715	0.714	0.715	0.715		
	800	0.010	0.296	0.291	0.955	-0.008	0.517	0.502	0.953	0.012	0.009			0.671	0.694	0.671	0.696		

Summary results of the RV144 data simulation. Bias, empirical standard deviation (ESD), average of the estimated standard error (ASE), and empirical coverage probability (CP) of 95% confidence intervals for the Full, IPW and CC estimators of γ_{k1} , γ_{k2} under models (25) for $k = 1$ and 2. The total selection probability is about $s_0 = 0.03$ and there is about 98% censoring for model (25). Each entry is based on 500 simulations.

Table 2

Estimator	n	k	γ_{k1}					γ_{k2}				
			Bias	ESD	ASE	CP	Bias	ESD	ASE	CP	Bias	ESD
Full	8198	1	0.0000	0.0011	0.0012	0.962	0.0000	0.0012	0.0013	0.952	0.0000	0.0012
IPW			-0.0001	0.0014	0.0014	0.966	0.0000	0.0016	0.0016	0.958	0.0000	0.0016
CC			-0.0074	0.0362	0.0356	0.964	-0.0103	0.0383	0.0383	0.948	-0.0103	0.0383
Full		2	-0.0001	0.0011	0.0011	0.958	-0.0001	0.0013	0.0013	0.954	-0.0001	0.0013
IPW			-0.0001	0.0013	0.0014	0.962	-0.0001	0.0016	0.0016	0.968	-0.0001	0.0016
CC			-0.0229	0.0335	0.0339	0.936	-0.0062	0.0393	0.0389	0.952	-0.0062	0.0393