Published in final edited form as:

Lifetime Data Anal. 2017 January ; 23(1): 136–159. doi:10.1007/s10985-016-9364-1.

Joint modeling of longitudinal and survival data with the Cox model and two-phase sampling

Rong Fu and

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Peter B. Gilbert

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Abstract

A common objective of cohort studies and clinical trials is to assess time-varying longitudinal continuous biomarkers as correlates of the instantaneous hazard of a study endpoint. We consider the setting where the biomarkers are measured in a designed sub-sample (i.e., case-cohort or two-phase sampling design), as is normative for prevention trials. We address this problem via joint models, with underlying biomarker trajectories characterized by a random effects model and their relationship with instantaneous risk characterized by a Cox model. For estimation and inference we extend the conditional score method of Tsiatis and Davidian (2001) to accommodate the two-phase biomarker sampling design using augmented inverse probability weighting with nonparametric kernel regression. We present theoretical properties of the proposed estimators and finite-sample properties derived through simulations, and illustrate the methods with application to the AIDS Clinical Trials Group (ACTG) 175 antiretroviral therapy trial. We discuss how the methods are useful for evaluating a Prentice surrogate endpoint, mediation, and for generating hypotheses about biological mechanisms of treatment efficacy.

Keywords

Case-cohort; Measurement error; Proportional hazards model; Prentice surrogate endpoint evaluation; Random effects model

1 Introduction

Consider a prospective cohort study or randomized clinical trial that measures a continuous biomarker(s) longitudinally in a designed sub-sample of study participants and follows subjects for a study endpoint of interest. Our goal is to study the association of the underlying trajectory of the true biomarker over time with the instantaneous hazard rate of the study endpoint via a Cox model, with related goals to assess the biomarker as a surrogate endpoint (Prentice, 1989) and as a mediator. To allow unbiased estimation accounting for

Supplementary Material

Supplementary materials are provided in nline Resource including the methods for mediation analysis, proofs of theorems and lemmas, and additional simulation and ACTG 175 results.

measurement error in the biomarker, we use the “joint modeling” approach that models the underlying true trajectory of the time-dependent covariate and the event time simultaneously (Tsiatis et al, 1995).

A joint model consists of two sub-models: one for the inherent trajectory of the time-dependent covariate and one for the time-to-event process. For the covariate trajectory, commonly used models include linear mixed effects models (Henderson et al, 2000; Guo and Carlin, 2004) and linear random effects models (Dafni and Tsiatis, 1998; Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2001). For the time-to-event process, we focus on the Cox model because it can be a good fit to vaccine efficacy trial applications where there is major interest in understanding how the current level of an immune response biomarker affects the instantaneous risk of pathogen infection, and the proportional hazards assumption may be reasonable. Unlike the standard Cox model, the joint modeling method aims to quantify the association of the underlying true and unobserved evolution of the biomarker process with the time-to-event process.

One approach to fitting the joint model is the biased two-stage method, which first fits the time-dependent covariate models and secondly uses ordinary Cox regression with fitted covariate values from the stage one models (Self and Pawitan, 1992). Another approach is recalibration methods building on Prentice (1982), which reduce the bias by estimating the hazard function given the observed covariate values (Tsiatis et al, 1995; Wang et al, 2001). However, since it is complex to derive the analytical form of the observed-covariate hazard, such methods are generally based on strong modeling assumptions and approximations, thus not completely eliminating bias. Likelihood approaches have also been developed, which are asymptotically consistent and efficient under correct specification of the joint distribution of the random effects and the event time data (DeGruttola and Tu, 1994; Wulfsohn and Tsiatis, 1997; Rizopoulos et al, 2009). However, misspecification induces bias and the method is computationally burdensome, especially for multiple biomarkers, because the likelihood function typically does not have a closed form solution for the integral over unknown random effects.

Tsiatis and Davidian (2001) developed a conditional score method with notable innovation that it does not require any distributional assumptions for the random effects. This method is much less computationally intensive and is easy to generalize to handle multiple time-dependent covariates (Song et al, 2002). Moreover, it yields consistent and asymptotically normal estimators under regularity conditions. Wang (2006) proposed a corrected score method and provided comparison studies to the two-stage, regression calibration and conditional score methods. For reviews of more joint modeling research see Tsiatis and Davidian (2004) and Wu et al (2012). Here we favor the conditional score method, because many studies assess multiple biomarkers (e.g., vaccine efficacy trials) where the likelihood approach may computationally fail. However, to the best of our knowledge, no joint modeling approaches have focused on the common situation where the longitudinal biomarkers are measured on a designed sub-sample of the full study cohort, for example with a case-cohort sample (Prentice, 1986) or two-phase sample (Haneuse et al, 2011). We extend the full data conditional score method to handle general types of missing at random sub-sampling designs with Bernoulli sampling of subjects for measuring the biomarker

trajectories, using inverse probability weighting (IPW) or augmented IPW (AIPW) to correct for biased sampling. Our method is developed under a general two-phase sampling scheme, which includes the outcome-dependent sampling designs as special cases (e.g., case-control and case-cohort designs). For simplicity of exposition, we concentrate on the model with a single longitudinal biomarker. The theories and derivations also apply for multiple biomarkers.

Section 2 introduces the models and summarizes their application for assessing Prentice surrogate endpoints and mediators. Section 3 develops IPW and AIPW estimation and testing procedures, including theoretical results on consistency, asymptotic normality, and relative efficiency. Section 4 studies finite-sample properties of the methods in simulations where the case-control sub-sampling design is considered. Section 5 illustrates the methods with a case-control sample from the ACTG 175 trial, and Section 6 concludes with discussion. Web Appendix A in Online Resource provides results on using the method for assessing mediation. Web Appendix B provides background IPW and AIPW theory that undergirds Section 2. Web Appendix C contains proofs of all results in Section 2 (drawing on results in Web Appendix B), Web Appendix D contains additional simulation results, and Web Appendix E contains additional ACTG 175 results. R code implementing the methods is posted at the second author's website.

2 Notation, Joint Model, Sampling Design, Surrogates and Mediators

Let T and C be the event time and censoring time; we observe $V = \min(T, C)$ and $\delta = \mathbb{I}(T \leq C)$. Let $\tilde{Z} = (Z, L^T)^T$ where Z is the randomized treatment group (1 for active and 0 for placebo) and L is a $p - 1$ dimensional vector of baseline time-independent covariates of interest to control for in assessing the association of the biomarker with the hazard of T . The same set up applies for prospective cohort studies without an assigned treatment Z , with $\tilde{Z} = L$. We denote the time-dependent biomarker process— which is not observed directly—by $\tilde{X}(\tau) = \{X(u), 0 \leq u \leq \tau\}$, with τ the end of follow-up and $X(u)$ the biomarker value at time u . We assume the following random effects model representing the inherent trajectory of $X(u)$:

$$X(u) = \alpha^T f(u), \quad (1)$$

where $f(u)$ is a q -dimensional vector of known functions of time u and α is a vector of subject-specific random effects. We assume the observed longitudinal biomarker values $W(u)$ are from an additive measurement error model

$$W(u) = \alpha^T f(u) + e(u), \quad (2)$$

where the measurement errors $e(u)$ are normally distributed with mean zero and variance $\text{Cov}\{e(u), e(s)\} = \mathbb{I}(u = s)\sigma^2$. Suppose the set of measurement time points are

$T^m = (T_1^m, \dots, T_J^m)^T$ with $0 \leq T_1^m < T_2^m < \dots < T_J^m \leq V$, and J the total number of time

points. These time points and J can vary across subjects, which occurs for many studies. Let $W = (W_1, \dots, W_J)^T$ where $W_j = W(T_j^m)$ and $e = (e_1, \dots, e_J)^T$ with $e_j = e(T_j^m)$.

For any fixed time u , let $J(u)$ be the number of measurement time points up to and including time u (i. e. $0 \leq T_1^m < \dots < T_{J(u)}^m \leq u < T_{J(u)+1}^m$), $T^m(u) = (T_1^m, \dots, T_{J(u)}^m)^T$ be the corresponding vector of ordered measurement time points, and $\tilde{e}(u) = (e_1, \dots, e_{J(u)})^T$. We consider the following proportional hazards model of the event time

$$\begin{aligned} \lambda(u) &= \lim_{du \rightarrow 0} \frac{1}{du} \mathbb{P}r \left\{ u \leq T < u + du \mid T \geq u, \alpha, \tilde{Z}, T^m(u), \tilde{e}(u), C \right\} \\ &= \lim_{du \rightarrow 0} \frac{1}{du} \mathbb{P}r \left(u \leq T < u + du \mid T \geq u, \alpha, \tilde{Z} \right) \\ &= \lambda_0(u) \exp \left\{ X(u)\beta + \tilde{Z}^T \eta + X(u)Z\gamma \right\}. \end{aligned} \quad (3)$$

We assume the censoring, measurement time points, and measurement errors are non-informative given the information already provided by α and $\tilde{Z}$, i.e. $T \perp (C, T^m, e) \mid (\alpha, \tilde{Z})$. We also assume e is independent of $(T, C, \tilde{Z}, \alpha, T^m)$. More detailed discussion on the underlying assumptions of this model can be found in Tsiatis and Davidian (2004).

2.1 Two-phase sampling model

Preventive vaccine efficacy trials illustrate studies where biomarker data are not collected on all participants by design. For example, in the RV144 HIV vaccine trial, dozens of immune response biomarkers were measured using a two-phase sampling design (Haynes et al, 2012). To obtain valid inferences, appropriate techniques are needed to account for the biased sampling design that over-samples certain groups.

We consider the following sampling design. In the first phase, data

$O_i = (V_i, \Delta_i, Z_i, L_i^T, A_i^T)^T$, $i = 1, \dots, N$ are collected from N independent subjects from a study population. The vector A_i is auxiliary variables that may be predictive of the biomarker. In the second phase, a Bernoulli random sample is taken from the N subjects, with sampling probabilities given by $\pi(O_i)$. Let ξ_i be the binary indicator of being sampled ($\xi_i = 1$). Thus the longitudinal biomarker information $\{W_i, T_i^m, J_i\}$ is assessed only on subjects with $\xi_i = 1$. The observed data are $\{V_i, Z_i, L_i, A_i, \xi_i, \xi_i W_i, \xi_i T_i^m, \xi_i J_i\}$ for $i = 1, \dots, N$.

The sampling probability model can be characterized by a parametric model with finite-dimensional parameter ρ ,

$$\mathbb{P}r(\xi=1 \mid O, \alpha, W, T^m, J) = \mathbb{P}r(\xi=1 \mid O) = \pi(O; \rho), \quad (4)$$

which expresses the missing at random (MAR) assumption. We also assume positive sampling probabilities with $0 < \delta < \pi(O_i; \rho) \leq 1$ for some constant $\delta > 0$ and all $i = 1, \dots, N$. As the outcome data are collected in the first phase on all subjects, an outcome-dependent

sampling scheme could be employed with the sampling probabilities depending on the outcome variable.

For the special case of stratified Bernoulli sampling, the N subjects are divided exclusively and exhaustively into S strata based on $\mathcal{O}: \{\mathcal{O}_1, \dots, \mathcal{O}_S\}$. We use $I(O \in \mathcal{O}_s)$ to indicate whether a subject belongs to stratum $\mathcal{O}_s$. Let $N_1, \dots, N_S$ be the stratum sizes with $N_1 + \dots + N_S = N$. In the second phase, if a subject belongs to stratum $\mathcal{O}_s$, then s/he is sampled with probability $\rho_s: \pi(O; \rho) = \sum_{s=1}^S I(O \in \mathcal{O}_s) \rho_s$. We also denote strata probabilities ρ_s as $\nu_s = \Pr(O \in \mathcal{O}_s) > 0, s = 1, \dots, S$.

2.2 Prentice surrogate endpoint assessment

In addition to providing assessment of the association of the underlying time-dependent biomarker with instantaneous risk in one or more groups, in randomized trials our methods assess the underlying biomarker as a surrogate endpoint according to the Prentice (1989) definition. Prentice's main operational criteria for checking this definition in terms of model (3) are: (i) Z has an effect on T , and Z has an effect on the covariate process $\tilde{X}(\tau)$; (ii) $\tilde{X}(\tau)$ is associated with the clinical endpoint T in each treatment group; and (iii) Z has no effect on T given the biomarker $\tilde{X}(\tau)$.

Conditions (i)–(iii) can be assessed based on estimation and testing of the coefficients in different versions of model (3). The key condition (iii) may be checked by first ruling out that the treatment effect on T is modified by the biomarker in (3). If ruled out, then the model $\lambda(u) = \lambda_0(u) \exp\{X(u)\beta + Z\eta_{Z|X} + L^T \eta_L\}$ may be fitted and one would check if $\eta_{Z|X}$ is plausibly close to zero, for example based on a confidence interval. All of the model fits should include all known and putative dual predictors L for $X(u)$ and T , in order that the Prentice criteria truly check the Prentice surrogate definition.

The proportion of treatment effect explained (PTE) may be used as a summary measure for the closeness of the biomarker in meeting the Prentice definition, $\text{PTE} = 1 - \eta_{Z|X}/\eta_Z$, where η_Z is the regression coefficient of Z without adjustment for the biomarker (Lin et al, 1997). The re-parameterized PTE from Kobayashi and Kuroki (2014) may also be used, the proportion of the treatment effect captured by the candidate surrogate (PCS), which is guaranteed to be in $[0, 1]$ and is less variable: $\text{PCS} = 1 / \{1 + (1/\text{PTE} - 1)^2\}$.

2.3 Natural direct effects/mediation approach

Mediation of the treatment effect through the true biomarker process may be studied using natural direct and indirect effects. Define the counterfactual underlying biomarker history up to and including time t as $\tilde{X}^z(t) = \{X^z(u), 0 \leq u \leq t\}$ if treatment had been $Z = z$. Under the random effects model (1), this trajectory is entirely determined by the random effects. Moreover, the time to event depends on this trajectory entirely through the random effects. Therefore, we use notation X_α^z to emphasize that the whole biomarker trajectory is determined by α .

The natural direct effect (NDE_Z) contrasts a functional of the distributions of $T^{1X_\alpha^z}$ and $T^{0X_\alpha^z}$, where X_α^z is the biomarker value if treatment had been z , for $z \in \{0, 1\}$. Similarly, the

natural indirect effect (NIE_z) contrasts $T^{zX_\alpha^1}$ with $T^{zX_\alpha^0}$. The total effect (TE) can be decomposed into the sum of NDE_z and NIE_{1-z} for each $z = 0, 1$. This motivates a “proportion of the total effect mediated by the biomarker” summary measure, for example $pNIE_z = 1 - NDE_z / TE$, or the PCS_z defined in Kobayashi and Kuroki (2014) that is guaranteed to be in $[0, 1]$, $PCS_z = 1 / \{1 + (1/pNIE_z - 1)^2\}$. Web Appendix A in Online Resource describes how model (3) can be used to define estimands NDE_z and NIE_z of interest, and provides results on identifiability assumptions and estimation.

3 IPW and AIPW Conditional Score Methods

Web Appendix B summarizes properties of IPW and AIPW estimators that undergird our results in this section, including results on efficiency and robustness to outlying sampling weights. We first briefly review the conditional score method with full sampling, which specifies notation needed later.

3.1 Conditional score estimator with full sampling

Let $\theta = (\beta, \eta^T, \gamma)^T$ be the regression coefficients in model (3). Define the event counting process $N(u) = I(V = u, J(u) = q)$ and the at-risk process $Y(u) = I(V = u, J(u) = q)$, where $J(u) = q$ indicates that q measurements have been observed up to and including time u . Define the design matrix and the vector of observed longitudinal measurements for each subject up to and including time u as $\tilde{F}(u) = [f(T_1^m), \dots, f(T_{J(u)}^m)]^T$ and $\tilde{W}(u) = (W_1, \dots, W_{J(u)})^T$. Then, (2) implies $\tilde{W}(u) = \tilde{F}(u)\alpha + \tilde{\epsilon}(u)$. Let $\hat{\alpha}(u) = \{\tilde{F}^T(u)\tilde{F}(u)\}^{-1}\tilde{F}^T(u)\tilde{W}(u)$ be the least squares estimate of α using data up to and including time u . Conditional on $\{\alpha, \tilde{Z}, T^m(u), Y(u) = 1\}$, the least squares estimate of $X(u)$, $\hat{X}(u) = \hat{\alpha}^T(u)f(u)$ is normally distributed as $N\{X(u), d(u, \sigma^2)\}$, where $d(u, \sigma^2) = \sigma^2 f^T(u) \{\tilde{F}^T(u)\tilde{F}(u)\}^{-1} f(u)$. Define $Q(u, \theta, \sigma^2) = \hat{X}(u) + dN(u)(\beta + \gamma Z)d(u, \sigma^2)$. Similarly as in Tsiatis and Davidian (2001), the conditional intensity for $N(u)$ can be derived as

$$\lim_{du \rightarrow 0} du^{-1} \Pr \left\{ dN(u) = 1 \mid Q(u, \theta, \sigma^2), \tilde{Z}, T^m(u), Y(u) = 1 \right\} = \lambda_0(u) \exp \left\{ \beta Q(u, \theta, \sigma^2) + \eta^T \tilde{Z} + \gamma Z Q(u, \theta, \sigma^2) - \frac{1}{2}(\beta + \gamma Z)^2 d(u, \sigma^2) \right\}, \quad (5)$$

which does not depend on the unknown random effects α (see Web Appendix C for a proof).

For notational simplicity, we let

$$H(u, \theta, \sigma^2) = \left[Q(u, \theta, \sigma^2), \tilde{Z}^T, Q(u, \theta, \sigma^2)Z \right]^T.$$

If σ^2 were known, then similarly as in Tsiatis and Davidian (2001), unbiased estimating equations for θ could be derived as

$$U_F(\theta, \sigma^2) = \sum_{i=1}^N \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{E_F^{(1)}(u, \theta, \sigma^2)}{E_F^{(0)}(u, \theta, \sigma^2)} \right\} dN_i(u) = 0, \quad (6)$$

where

$$E_i^{(0)}(u, \theta, \sigma^2) = \exp \left\{ H_i^T(u, \theta, \sigma^2) \theta - \frac{1}{2} (\beta + \gamma Z_i)^2 d_i(u, \sigma^2) \right\}, \quad (7)$$

$$E_i^{(1)}(u, \theta, \sigma^2) = H_i(u, \theta, \sigma^2) E_i^{(0)}(u, \theta, \sigma^2), \quad (8)$$

$$E_F^{(r)}(u, \theta, \sigma^2) = N^{-1} \sum_{i=1}^N Y_i(u) E_i^{(r)}(u, \theta, \sigma^2), \quad r=0, 1. \quad (9)$$

In practice, we could estimate σ^2 as $\hat{\sigma}^2$ by solving $S_{e,F}(\sigma^2) \equiv \sum_{i=1}^N S_{e,i}(\sigma^2) = 0$, where

$$S_{e,i}(\sigma^2) = J_i I(J_i \geq q) \left[\left\{ \tilde{W}_i(V_i) - \tilde{F}_i(V_i) \hat{\alpha}_i(V_i) \right\}^T \left\{ \tilde{W}_i(V_i) - \tilde{F}_i(V_i) \hat{\alpha}_i(V_i) \right\} - \sigma^2 (J_i - q) \right]. \quad (10)$$

Then, θ is estimated by solving $U_F(\theta, \hat{\sigma}^2) = 0$.

3.2 IPW conditional score estimator

3.2.1 Prespecified sampling probabilities $\pi_i = \pi(\mathbf{O}_i)$ —We start with the case with correctly and fully specified sampling probabilities. We define the IPW conditional score estimator $\hat{\theta}_{IPW}(\pi)$ for θ as the solution to $U_{IPW}(\theta, \hat{\sigma}_{IPW}^2(\pi), \pi) = 0$ where

$$U_{IPW}(\theta, \hat{\sigma}_{IPW}^2(\pi), \pi) = \sum_{i=1}^N \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \hat{\sigma}_{IPW}^2(\pi)) - \frac{E_{IPW}^{(1)}(u, \theta, \hat{\sigma}_{IPW}^2(\pi), \pi)}{E_{IPW}^{(0)}(u, \theta, \hat{\sigma}_{IPW}^2(\pi), \pi)} \right\} dN_i(u). \quad (11)$$

Here $\hat{\sigma}_{IPW}^2(\pi)$ estimates σ^2 by solving $S_{e,IPW}(\sigma^2, \pi) \equiv \sum_{i=1}^N (\xi_i / \pi_i) S_{e,i}(\sigma^2) = 0$, and

$$E_{IPW}^{(r)}(u, \theta, \sigma^2, \pi) = \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} Y_i(u) E_i^{(r)}(u, \theta, \sigma^2), \quad r=0, 1. \quad (12)$$

The baseline hazard could be estimated as

$$d\hat{\lambda}_0^{IPW}(u) = \frac{1}{N} \sum_i dN_i(u) / \left\{ E_{IPW}^{(0)}(u, \hat{\theta}_{IPW}(\pi), \hat{\sigma}^2(\pi), \pi) \right\}.$$

To precede to theoretical results, we define

$$M_i(\theta, \sigma^2) = \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{e^{(1)}(u, \theta, \sigma^2)}{e^{(0)}(u, \theta, \sigma^2)} \right\} dD_i(u), \quad (13)$$

$$dD_i(u) = dN_i(u) - \lambda_0(u) Y_i(u) E_i^{(0)}(u, \theta, \sigma^2) du, \quad (14)$$

$$e^{(r)}(u, \theta, \sigma^2) = \left\{ Y(u) E^{(r)}(u, \theta, \sigma^2) \right\}, \quad r=0, 1. \quad (15)$$

By Lemma C.1 (Web Appendix C) and the same arguments as for (8a) in Tsiatis and Davidian (2001), we can show that $\{M(\theta_0, \sigma_0^2)\} = 0$, where θ_0 is the true parameter of (3) and σ_0^2 is the true variance of measurement errors. We show next that under regularity conditions, the estimating functions $N^{-1} U_{IPW}(\theta, \hat{\sigma}_{IPW}^2(\pi), \pi)$ are asymptotically equivalent to $N^{-1} \sum_{i=1}^N (\xi_i / \pi_i) M_i(\theta, \sigma^2)$. The latter is a mean of i.i.d. random variates for which empirical theoretical results are readily applied. Let $(\theta_0^T, \sigma_0^2, \rho_0^T)^T$ be the true parameters. For any parameter, for example θ , we use $\mathcal{N}(\theta_0)$ to denote a compact neighborhood of θ_0 and $\mathcal{N}(\tau, \theta_0)$ for $[0, \tau] \times \mathcal{N}(\theta_0)$. Let $\mathbb{E}_0(\cdot)$ and $\mathbb{V}_{\theta_0}(\cdot)$ denote the expectation and variance evaluated under the truth. We also use $\dot{f}_\theta = \partial f / \partial \theta$ to denote the derivative of a function f with respect to the parameter θ . We sometimes omit θ in the subscript and use $\dot{f}$ when f is fully parameterized by θ .

Lemma 1: Under conditions A1–A5 listed in Assumption A of Web Appendix C,

$$\sup_{(u, \theta^T, \sigma^2)^T \in \mathcal{N}(\tau, \theta_0, \sigma_0^2)} \left| \frac{E_{IPW}^{(1)}(u, \theta, \sigma^2, \pi)}{E_{IPW}^{(0)}(u, \theta, \sigma^2, \pi)} - \frac{e^{(1)}(u, \theta, \sigma^2)}{e^{(0)}(u, \theta, \sigma^2)} \right| \xrightarrow{p} 0 \text{ as } N \rightarrow \infty.$$

Lemma 2: Under conditions A1–A5, $N^{-1/2} U_{IPW}(\theta_0, \sigma_0^2, \pi)$ is asymptotically linear with

$$N^{-1/2}U_{IPW}(\theta_0, \sigma_0^2, \pi) = N^{-1/2} \sum_{i=1}^N \frac{\xi_i}{\pi_i} M_i(\theta_0, \sigma_0^2) + o_p(1).$$

Theorem 1: Under conditions A1–A8, as $N \rightarrow \infty$, (i) $\hat{\theta}_{IPW}(\pi) \xrightarrow{p} \theta_0$ and (ii) $\sqrt{N}(\hat{\theta}_{IPW}(\pi) - \theta_0)$ converges weakly to a normal random variate with mean zero and covariance matrix $A^{-1}B(A^{-1})^T$, where

$$A = \mathbb{E}_0 \left(\dot{M}_\theta \right), \quad B = \mathbb{E}_0 \left(\frac{1}{\pi} R R^T \right), \\ R = M(\theta_0, \sigma_0^2) - \mathbb{E}_0 \left(\dot{M}_{\sigma^2} \right) \left\{ \mathbb{E}_0 \left(\dot{S}_e \right) \right\}^{-1} S_e(\sigma_0^2).$$

3.2.2 Estimated sampling probabilities $\hat{\pi}_i = \pi(\mathbf{O}_i, \hat{\rho})$ —Results in Web Appendix B show that the IPW estimator with prespecified sampling probabilities π is inefficient. It is often recommended to use the estimated $\hat{\pi} = \hat{\pi}(\hat{\rho})$ to improve efficiency. To implement this we still apply the sampling probability model (4) and the resulting likelihood score functions

$$S_{\pi, F}(\rho) = \sum_{i=1}^N S_{\pi, i}(\rho) = \sum_{i=1}^N \frac{\partial}{\partial \rho} \log \left[\pi(O_i; \rho)^{\xi_i} \{1 - \pi(O_i; \rho)\}^{1 - \xi_i} \right] \\ = \sum_{i=1}^N \frac{\xi_i - \pi(O_i; \rho)}{\pi(O_i; \rho) \{1 - \pi(O_i; \rho)\}} \frac{\partial \pi(O_i; \rho)}{\partial \rho}. \quad (16)$$

The IPW estimators $\hat{\sigma}_{IPW}^2(\hat{\pi})$ and $\hat{\theta}_{IPW}(\hat{\pi})$ are obtained by solving $S_{e, IPW}(\sigma^2, \hat{\pi}) = 0$ and $U_{IPW}(\theta, \hat{\sigma}_{IPW}^2(\hat{\pi}), \hat{\pi}) = 0$, respectively.

Theorem 2: Under conditions A1–A8, as $N \rightarrow \infty$, (i) $\hat{\theta}_{IPW}(\hat{\pi}) \xrightarrow{p} \theta_0$ and (ii) $\sqrt{N}(\hat{\theta}_{IPW}(\hat{\pi}) - \theta_0)$ converges weakly to a normal random variate with mean zero and covariance matrix $A^{-1}B^*(A^{-1})^T$, where

$$B^* = B - \mathbb{E}_0 \left(R \frac{\dot{\pi}}{\pi} \right) \left\{ \mathbb{E}_0 (S_\pi S_\pi^T) \right\}^{-1} \left\{ \mathbb{E}_0 \left(R \frac{\dot{\pi}}{\pi} \right) \right\}^T,$$

p with A, B and R defined in Theorem 1.

3.3 AIPW conditional score estimator

The IPW estimator is easy to implement but is biased if the sampling model is misspecified, which could occur if some selected subjects have happenstance missing data. In Web Appendix B we state the result that the use of estimated sampling probabilities can improve the efficiency of the IPW estimator, but it does not achieve the minimal variance bound

unless it is augmented as the AIPW estimator and implemented with a correct model for the full data given O under stratified Bernoulli sampling.

If we temporarily assume the $e^{(r)}(u, \theta, \sigma^2)$ ($r = 0, 1$) in (13) are known, then for our problem, the double robustness property of the AIPW estimator is expressed as follows— if either the sampling probability model for $\pi(O; \rho)$ or the model for $\mathbb{E}\{M(\theta, \sigma^2)|O\}$ is correct, then the estimating equations are unbiased for θ . In practice, $\mathbb{E}\{M(\theta, \sigma^2)|O\}$ is estimated as accurately as possible. Given the MAR assumption in (4), we have $\mathbb{E}\{M(\theta, \sigma^2)|\xi = 1, O\} = \mathbb{E}\{M(\theta, \sigma^2)|O\}$. Therefore we could build a model to estimate it based on the second phase sample. However, since the $e^{(r)}(u, \theta, \sigma^2)$ ($r = 0, 1$) are unknown we have to estimate them, with approaches for doing so described below. We first develop the AIPW conditional score estimator and its asymptotic properties assuming π and $\mathbb{E}(\cdot|O)$ related to the estimating equations are fully and correctly specified. Then we move on to the situations where either or both of these models are estimated.

3.3.1 Prespecified sampling probabilities $\pi_i = \pi(O_i)$ and conditional expectation $\mathbb{E}(\cdot|O)$ —Similar to Qi et al (2005), we define the AIPW conditional score estimating functions for θ as

$$U_{AUG}(\theta, \sigma^2, \pi,) = \sum_{i=1}^N \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{E_{AUG}^{(1)}(u, \theta, \sigma^2, \pi)}{E_{AUG}^{(0)}(u, \theta, \sigma^2, \pi)} \right\} dN_i(u) + \sum_{i=1}^N \left(1 - \frac{\xi_i}{\pi_i}\right) \int_0^\tau \left[\{H_i(u, \theta, \sigma^2) dN_i(u)|O_i\} - \frac{E_{AUG}^{(1)}(u, \theta, \sigma^2, \pi)}{E_{AUG}^{(0)}(u, \theta, \sigma^2, \pi)} \{dN_i(u)|O_i\} \right], \quad (17)$$

where, for $r = 0, 1$,

$$E_{AUG}^{(r)}(u, \theta, \sigma^2, \pi) = \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} Y_i(u) E_i^{(r)}(u, \theta, \sigma^2) + \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{\xi_i}{\pi_i}\right) \{Y_i(u) E_i^{(r)}(u, \theta, \sigma^2)|O_i\}.$$

Unlike the partial likelihood for classic Cox regression, here we define the at-risk process as $Y_i(u) = I(V_i - u, J_i(u) - q)$, which contains the incomplete data of measurement time points.

The estimate $\hat{\theta}_{AUG}(\pi, \mathbb{E})$ for θ solves $U_{AUG}(\theta, \hat{\sigma}_{AUG}^2(\pi,), \pi,) = 0$, where $\hat{\sigma}_{AUG}^2(\pi,)$ solves

$$S_{e, AUG}(\sigma^2, \pi,) = \sum_{i=1}^N \frac{\xi_i}{\pi_i} S_{e, i}(\sigma^2) + \sum_{i=1}^N \left(1 - \frac{\xi_i}{\pi_i}\right) \{S_{e, i}(\sigma^2)|O_i\} = 0. \quad (18)$$

Similarly as in (13), we define

$$M_{AUG, i}(\theta, \sigma^2, \pi) = \frac{\xi_i}{\pi_i} M_i(\theta, \sigma^2) + \left(1 - \frac{\xi_i}{\pi_i}\right) \{M_i(\theta, \sigma^2)|O_i\}. \quad (19)$$

Evidently $\mathbb{E}\{M_{AUG}(\theta, \sigma^2, \pi)\} = \mathbb{E}\{M(\theta, \sigma^2)\}$ with correctly specified sampling probabilities. We will demonstrate that under the regularity conditions listed in Assumption A and Assumption B (stated in Web Appendix C), $N^{-1}U_{AUG}(\theta, \sigma^2, \pi, \mathbb{E})$ is asymptotically equivalent to the i.i.d. average $N^{-1}\sum_{i=1}^N M_{AUG,i}(\theta, \sigma^2, \pi)$.

Lemma 3: Under conditions listed in Assumption A and B, as $N \rightarrow \infty$,

$$\sup_{(u, \theta^T, \sigma^2)^T \in \mathcal{N}(\tau, \theta_0, \sigma_0^2)} \left| \frac{E_{AUG}^{(1)}(u, \theta, \sigma^2, \pi)}{E_{AUG}^{(0)}(u, \theta, \sigma^2, \pi)} - \frac{e^{(1)}(u, \theta, \sigma^2)}{e^{(0)}(u, \theta, \sigma^2)} \right| \xrightarrow{p} 0.$$

Lemma 4: Under conditions in Assumption A and B, $N^{-1/2}U_{AUG}(\theta_0, \sigma_0^2, \pi, \cdot)$ is asymptotically linear with

$$N^{-1/2}U_{AUG}(\theta_0, \sigma_0^2, \pi, \cdot) = N^{-1/2}\sum_{i=1}^N M_{AUG,i}(\theta_0, \sigma_0^2, \pi) + o_p(1).$$

Theorem 3: Under conditions in Assumption A and B, as $N \rightarrow \infty$, (i) $\hat{\theta}_{AUG}(\pi, \cdot) \xrightarrow{p} \theta_0$ and (ii) $\sqrt{N}(\hat{\theta}_{AUG}(\pi, \cdot) - \theta_0)$ converges weakly to a normal random variate with mean zero and covariance matrix $A^{-1}CA^{-1})^T$, where

$$C = B - \mathbb{E}_0 \left\{ \frac{1-\pi}{\pi} \mathbb{E}_0(R|O) \mathbb{E}_0(R|O)^T \right\}.$$

3.3.2 Estimated sampling probabilities $\hat{\pi}_i = \pi(O_i, \hat{\rho})$ and prespecified conditional expectation $\mathbb{E}(\cdot|O)$ —Parallel theoretical results hold when using estimated sampling probabilities, which we relegate to Theorem C.4 in Web Appendix C.

3.3.3 Prespecified sampling probabilities $\pi_i = \pi(O_i)$ and estimated $\hat{\mathbb{E}}(\cdot|O)$ —In practice it is hard to derive the analytical form of $\mathbb{E}(\cdot|O)$. A common approach replaces it with a function $h(O; \gamma)$ in terms of a finite-dimensional parameter γ . However, when $h(O; \gamma) \neq \mathbb{E}(\cdot|O)$ for any γ , the resulting estimator cannot achieve the optimal asymptotic variance given in Theorem 3 and Theorem C.4 (in Web Appendix C). Cao et al (2009) investigated how to find the optimal γ that minimizes the variance of the estimator of θ with h fixed.

Qi et al (2005) investigated the AIPW estimator for the Cox model with time-independent covariates. They proposed to estimate $\mathbb{E}(\cdot|O)$ by the non-parametric Nadaraya-Watson method, and proved that the optimal asymptotic variance can be achieved. We extend their method to our more complicated setting with a time-dependent covariate.

The scores in (17) show that we need to estimate $\mathbb{E}\{G(u, \theta, \sigma^2)|O\}$ as a continuous function of u and integrate it over u , where $G(u, \theta, \sigma^2)$ could be $Y(u)E^{(n)}(u, \theta, \sigma^2)$, $dN(u)$, or $H(u, \theta$,

$\sigma^2)dN(u)$. Suppose the predictor variables O are d continuous variables, in which case we use the Nadaraya-Watson estimator, i.e., for any $g(u, O; \theta, \sigma^2) \equiv \mathbb{E}\{G(u, \theta, \sigma^2)|O\}$,

$$\hat{g}(u, O; \theta, \sigma^2) \equiv \{G(u, \theta, \sigma^2)|O\} = \frac{\sum_{j=1}^N \xi_j G_j(u, \theta, \sigma^2) K_H(O - O_j)}{\sum_{j=1}^N \xi_j K_H(O - O_j)}, \quad (20)$$

where $K_H(\cdot)$ is an s^{th} order kernel function and H is the bandwidth that is a $d \times d$ symmetric and positive definite matrix. If any component of O is discrete, we consider kernel regression with mixed types of predictor variables (Hall et al, 2004), which can be implemented by the R package np. The resulting estimating equations for θ are $U_{AUG}(\theta, \sigma^2, \pi, \hat{E}) = 0$, where $\mathbb{E}$ in (17) is replaced with $\hat{E}$ given by (20).

However, intuitively such estimation relies on strong assumptions for the functional form of $g(u, \cdot; \cdot, \cdot)$ in u . By examining the two stochastic processes involving $dN(u)$ mentioned above, we find that they are always estimated as zero for time u if there are no events occurring at that time. This implies that, practically, the estimates of their conditional expectations given O at non-event time points are zero. Thus they only need to be estimated at event times. Alternatively, if O includes the event time $(V, \cdot)$, then by conditioning on it, $U_{AUG}(\theta, \sigma^2, \pi, \hat{E})$ reduces to

$$U_{AUG}(\theta, \sigma^2, \pi, \hat{E}) = \sum_{i=1}^N \frac{\xi_i}{\pi_i} \int_0^\tau \left\{ H_i(u, \theta, \sigma^2) - \frac{\hat{E}_{AUG}^{(1)}(u, \theta, \sigma^2, \pi)}{\hat{E}_{AUG}^{(0)}(u, \theta, \sigma^2, \pi)} \right\} dN_i(u) + \sum_{i=1}^N \left(1 - \frac{\xi_i}{\pi_i}\right) \int_0^\tau \left\{ \hat{g}^H(u, O_i; \theta, \sigma^2) - \frac{\hat{E}_{AUG}^{(1)}(u, \theta, \sigma^2, \pi)}{\hat{E}_{AUG}^{(0)}(u, \theta, \sigma^2, \pi)} \hat{g}^J(u, O_i; \theta, \sigma^2) \right\} dN_i^*(u) \quad (21)$$

where $Y_i^*(u) = I(V_i \geq u)$, $dN_i^*(u) = I(V_i = u, \Delta_i = 1)$, and, for $r = 0, 1$,

$$\begin{aligned} \hat{E}_{AUG}^{(r)}(u, \theta, \sigma^2, \pi) &= \frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} Y_i(u) E_i^{(r)}(u, \theta, \sigma^2) + \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{\xi_i}{\pi_i}\right) Y_i^*(u) \hat{g}^{(r)}(u, O_i; \theta, \sigma^2), \\ g^J(u, O; \theta, \sigma^2) &= \{I(J(u) \geq q)|O\}, \\ g^H(u, O; \theta, \sigma^2) &= \{I(J(u) \geq q)H(u, \theta, \sigma^2)|O\}, \\ g^{(r)}(u, O; \theta, \sigma^2) &= \{I(J(u) \geq q)E^{(r)}(u, \theta, \sigma^2)|O\}. \end{aligned}$$

Similarly, σ^2 can be estimated by solving $S_{e,AUG}(\sigma^2, \pi, \hat{E}) = 0$ with estimators denoted by $\hat{\theta}_{AUG}(\pi, \hat{E})$ and $\hat{\sigma}_{AUG}^2(\pi, \hat{E})$.

Theorem 4: Under conditions in Assumption A, B and C, as $N \rightarrow \infty$, (i) $\hat{\theta}_{AUG}(\pi, \hat{E}) \xrightarrow{P} \theta_0$ and (ii) $\sqrt{N}(\hat{\theta}_{AUG}(\pi, \hat{E}) - \theta_0)$ converges weakly to a normal random variate with mean zero and covariance matrix $A^{-1}CA^{-1}$.

3.3.4 Estimated sampling probabilities $\hat{\pi}_i = \pi(\mathbf{O}_i, \hat{\rho})$ and estimated $\hat{\mathbb{E}}(\cdot|\mathbf{O})$ —

When the sampling probabilities are estimated via setting the score functions in (16) to zero, results parallel to those in Section 3.3.2 hold. We estimate θ and σ^2 by solving $U_{AUG}(\theta, \sigma^2, \hat{\pi}, \hat{\mathbb{E}}) = 0$ and $S_{e,AUG}(\sigma^2, \hat{\pi}, \hat{\mathbb{E}}) = 0$. We denote the obtained estimators as $\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}})$ and $\hat{\sigma}_{AUG}^2(\hat{\pi}, \hat{\mathbb{E}})$.

Theorem 5: Under conditions in Assumptions A, B, and C, as $N \rightarrow \infty$, (i) $\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}}) \xrightarrow{P} \theta_0$ and (ii) $\sqrt{N}(\hat{\theta}_{AUG}(\hat{\pi}, \hat{\mathbb{E}}) - \theta_0)$ converges weakly to a normal random variate with mean zero and covariance matrix $A^{-1}C(A^{-1})^T$.

The above theorems yield sandwich variance estimators for the IPW and AIPW estimators based on empirical score functions, where parameters are replaced by estimates. The sandwich variance estimators are used in the simulation studies and in the application.

4 Simulation Studies of the IPW and AIPW Methods

We conduct simulation studies that mimic disease prevention trials such as vaccine efficacy trials. We generate the event time data $(\cdot, V)$ from Cox models with different combinations of the time-varying biomarker and treatment assignment to vaccine or placebo. We carry out 500 simulation runs for each scenario. Our primary goal is to evaluate the IPW and AIPW conditional score methods in two-phase sampling design cohort studies. For comparison, we also calculate the conditional score estimator based on full cohort data (Full)—the unobtainable benchmark—and conduct the naive complete-case (CC) analysis only using subjects selected into the second phase without any weighting. While bandwidth selection is important, the AIPW method requires fitting the kernel regression across time points, making it highly computationally intensive to seek the optimal bandwidth for each regression. Therefore the simulation studies use a fixed bandwidth determined from the variance of the predictor variables over time.

We conduct three main simulation studies. Study I assesses the Cox model for a biomarker in one treatment group only, with objective to compare the IPW and AIPW estimators using various sets of auxiliary variables to estimate the augmentation terms. Study II assesses the Cox model with both the biomarker and the treatment group as covariates. It also studies the impact of the number of biomarker measurements and misspecified measurement error models on estimator performance. Study III assesses the Cox model also including the interaction term for biomarker with treatment. All IPW and AIPW methods are implemented using both the prespecified true sampling probabilities (π) and the estimated sampling probabilities ($\hat{\pi}$). The results are generally similar, with the latter sometimes slightly more efficient as expected from the theory (results reported in Web Appendix D Tables 1–6. For brevity we focus on reporting the true π results. In all simulation studies, we evaluate the bias, Monte Carlo standard deviation (SD), and averages of estimated standard error (ASE) of parameter estimates. All three studies use a full cohort sample size of $N = 1500$, with phase II sample drawn with Bernoulli sampling probabilities $P(\xi = 1 | \cdot = 1) = 1$ and $P(\xi = 1 | \cdot = 0) = 0.33$, resulting in about 40% of subjects with biomarker data (Table 1). Web Appendix D Tables 1–6 show detailed results; below we report selected key results.

4.1 Simulation Study I

We consider the model $\lambda(u) = \lambda_0(u) \exp\{\beta X(u)\}$, where $X(u)$ is characterized by a linear random effects model $X(u) = \alpha_0 + \alpha_1 u$. We simulate $X(u)$ to imitate $\log_{10}$ CD4 cell count in the ACTG 175 study as described in Song et al (2002). The random effects $(\alpha_0, \alpha_1)^T$ are generated from a bivariate normal distribution with $\mathbb{E}(\alpha) = (2.5915, -0.00315)^T$ and $\text{Cov}(\alpha) = D$ with elements $(D_{11}, D_{12}, D_{22})^T = (0.02408, -0.0008, 0.000014)^T$. The variance of the measurement error e is $\sigma^2 = 0.01$. This set-up represents a noise-to-signal ratio of $\text{Var}(e)/\text{Var}\{X(0)\} \approx 0.42$, approximately the same as in ACTG 175. The scheduled visits for measuring $X(u)$ are at baseline and at weeks 2, 4, 8, 20, 32, 44, 56, 68, 80, where we create random uniform variability in visit attendance $\pm$ one-half week. Event times are generated from the Cox model with hazard ratios $e^\beta = \{1, 0.5, 0.25\}$. Censoring times follow the exponential distribution $\text{Exp}(1/180)$ and are subject to administrative censoring at Week 85. We choose the baseline hazards to yield an event rate of around 10%. Specifically, the proportions of subjects dropping off the study during follow-up and completing the study free of events are (35.5%, 54.1%), (36.0%, 54.0%) and (36.2%, 53.3%) when the hazard ratios are 1, 0.5 and 0.25, respectively. The average number of biomarker measurements per subject is around eight.

For the AIPW method, we evaluate several sets of auxiliary predictor variables $A = (A_1, A_2)^T$ that serve to estimate the augmentation terms in the nonparametric kernel regression (Nadaraya-Watson). We generate null, moderate, and strong correlation with the biomarker, quantified by correlation coefficient $R = \{0, 0.5, 0.95\}$. For $R = 0$, A is generated independently from α and $(\cdot, V)$, from $A_1 \sim N\{\mathbb{E}(\alpha_1), 1\}$ and $A_2 \sim N\{\mathbb{E}(\alpha_1 + 40\alpha_2), 1\}$. For $R = 0.5$ and 0.95 , $A_1 = \alpha_1 + e_1$, $A_2 = \alpha_1 + 40\alpha_2 + e_2$, $e_1 \sim N(0, d_1^2)$, $e_2 \sim N(0, d_2^2)$, with $d_1^2 = (1/R^2 - 1)\text{ar}(\alpha_1)$ and $d_2^2 = (1/R^2 - 1)\text{ar}(\alpha_1 + 40\alpha_2)$. In total, we consider eleven sets of predictor variables in the kernel regression: $(\cdot, V)$, A , $(\cdot, A)$ and $(\cdot, V, A)$, with A set with $R = 0, 0.5, 0.95$.

As reported in Figure 1, as expected the CC analysis generates very biased estimates when β is large. We observe slightly larger biases for the AIPW $(\cdot, V, A)$ estimators when $\beta = -\ln 4$. When the auxiliary variables are highly correlated with $X(u)$, including them provides less variable AIPW estimates than other methods, especially when $X(u)$ is strongly associated with event time. The 95% coverage probabilities are slightly below the nominal level when $\beta = -\ln 4$, suggesting that the variance is underestimated by the sandwich variance estimator.

4.2 Simulation Study II

We next consider the Cox model $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z\}$. We can assess the Prentice criteria by examining if η is plausibly close to zero. We model $X(u)$ with a linear random effects model as above and generate $Z \sim \text{Bernoulli}(0.5)$. The random effects α are simulated from a bivariate normal distribution with $\mathbb{E}(\alpha|Z=0) = (2.5915, -0.00145)^T$, $\mathbb{E}(\alpha|Z=1) = (2.5915, -0.00315)^T$ and $\text{Cov}(\alpha|Z) = D$ with elements $(D_{11}, D_{12}, D_{22})^T = (0.02408, -0.0008, 0.000014)^T$. We consider two sets of hazard ratios $(e^\beta, e^\eta)^T = \{(0.5, 1)^T, (0.5, 0.5)^T\}$.

To study the influence of i) the number of measurement time points; and ii) misspecification of the measurement error model, we consider the following four scenarios, all with random uniform variability in visit attendance $\pm$ one-half week:

II(a): $X(u)$ is measured at weeks 0, 2, 4, 8, 20, 32, 44, 56, 68, 80; $\epsilon(u) \sim N(0, 0.01)$.

II(b): $X(u)$ is measured at weeks 0, 8, 44, 80; $\epsilon(u) \sim N(0, 0.01)$.

II(c): $X(u)$ is measured at weeks 0, 2, 4, 8, 20, 32, 44, 56, 68, 80; $\epsilon(u) \sim \text{Exp}(10) - 0.1$.

II(d): $X(u)$ is measured at weeks 0, 2, 4, 8, 20, 32, 44, 56, 68, 80; measurement error depends on the value of the biomarker; $\epsilon(u) \sim N(0, 0.01)$ if $X(u) > 2.5$ and $\epsilon(u) \sim N(0, 0.05)$ if $X(u) \leq 2.5$.

Studies II(c) and II(d) assess the influence of the misspecified error models. The proportions of subjects dropping off the study during follow-up and completing the study event-free are (35.9%, 53.7%) and (36.1%, 53.4%) when the hazard ratios $(e^\beta, e^\eta)^T$ are $(0.5, 1)^T$ and $(0.5, 0.5)^T$, respectively. The average number of biomarker measurements per subject in Studies II(a), (c), and (d) is 8.3, and in Study II(b) is 3.3.

Figure 2 shows results for Studies II(a) and II(b). For II(b), AIPW does not yield an efficiency gain over IPW for $\hat{\beta}$ unless A is very strongly correlated with the biomarker, suggesting that when a limited number of biomarker measurements are available, using the more complex AIPW method can yield poorer results than the simpler IPW method.

Figure 3 shows the results of Studies II(c) and II(d). For II(c) where the error distribution has a heavy right tail, the bias for $\hat{\beta}$ is slightly greater than for II(a) where the error distribution was correctly specified. In contrast bias was near zero for $\hat{\eta}$, the same as for II(a), which occurs because Z has no measurement error. The 95% coverage probabilities are very close to the nominal level expect for the CC method. There are similar patterns of relative efficiency to those in II(a), suggesting that unless very strong predictors of the biomarker are available, the simple IPW method may be recommended. The influence of misspecified measurement error seems to center on the bias of $\hat{\beta}$, but the level of bias is acceptable, with relative bias up to 7.4%.

In II(d) where the measurement error depends on the biomarker, bias of $\hat{\beta}$ is greater than in II(c), especially in the setting with $(\beta, \eta)^T = (-\ln 2, -\ln 2)^T$, with relative bias as high as 10.7%.

4.3 Simulation study III

Lastly, we consider the Cox model $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z + \gamma X(u)Z\}$. Data are simulated as for II(a) with hazard ratios $(e^\beta, e^\eta)^T = (0.5, 0.5)^T$, and we specify no interaction, $e^\gamma = 1$. As reported in Figure 4, estimator performance for β and η is fairly similar to the other simulations, with similar conclusions. There is negligible bias for estimating the interaction coefficient γ .

5 Application to ACTG 175

We apply the methods to ACTG 175, a randomized trial comparing four antiretroviral regimens (zidovudine only, zidovudine+didanosine, zidovudine+ zalcitabine and didanosine only) in $N = 2,467$ HIV-infected subjects with CD4 cell counts 200–500 per cubic millimeter (Hammer et al, 1996). The study design measured CD4 cell counts in all subjects every 12 weeks starting at Week 8, and followed subjects for occurrence of the composite clinical endpoint of AIDS or death, whichever occurred first. A total of 308 endpoints/cases were observed. The median follow-up time was 143 weeks. Addressing the same objectives as Song et al (2002), we study treatment $Z = 1$ as three regimens pooled (zidovudine +didanosine, zidovudine+zalcitabine, and didanosine) versus $Z = 0$ as zidovudine only, and study the inherent $\log_{10}$ CD4 cell count trajectory $X(u)$ as a correlate of risk and as a Prentice surrogate endpoint for the composite clinical endpoint.

We created a case-control sample from the full study cohort with post-randomization CD4 cell data available, generating a study where the CD4 cell counts were measured only on this case-control sample. This sample included all cases and a simple random sample of 14.3% of controls, yielding 306 cases and 306 controls. The average number of CD4 measurements per subject was 9.3. Web Appendix Figure 1 shows observed $X(u)$ trajectories by case/control status and treatment group, showing that cases had a steeper drop over time than controls, and $Z = 1$ subjects had higher average $X(u)$ than $Z = 0$ subjects. Web Appendix Figure 2 shows that higher $X(u)$ is associated with lower rates of AIDS/death.

We explored the functional form of $X(u)$ over time using linear and quadratic weighted mixed effects models (fitted by maximum likelihood estimation with the `lme()` R function). Figure 5 and Web Appendix Table 7 support that the quadratic model provides a better fit. Accordingly, we select the random effects model (1) for $\log_{10}$ CD4 as $X(u) = a_0 + a_1u + a_2u^2$.

We now check the Prentice criteria using the joint model, where we adjust for the average baseline $\log_{10}$ CD4 cell count measured at the two time points before randomization, as it is a dual predictor of $X(u)$ and the clinical endpoint. Table 2 shows estimated coefficients from the IPW and AIPW methods using estimated $\hat{\pi}$ and using treatment and event status for estimating the augmentation terms. Five joint models are fit to evaluate the Prentice criteria, using five different independent variable sets. The IPW and AIPW methods yield similar point estimates, and for brevity we only interpret the results of the IPW analysis.

Model 0 shows that treatment significantly associates with AIDS/death, such that Prentice criterion (i) holds. Models 3 and 4 indicate that $\log_{10}$ CD4 cell count significantly associates with AIDS/death in the control group ($Z = 0$) but not in the treatment group ($Z = 1$), and Model 1 shows that the interaction is significant. Therefore the analysis rejects Prentice criteria (ii) and (iii). The surrogate-quality summary measures PTE and $PCS = 1/\{1 + (1/PTE - 1)^2\}$ are estimated to be near zero (-0.173 and 0.021), respectively. In addition, the mediation measure $pNIE_0$ is estimated to be near zero (0.079), indicating little evidence for mediation. In Web Appendix Table 8, we apply the conditional score method to all subjects from the ACTG 175 trial with CD4 data available. The results are generally

comparable to those from the IPW or AIPW method, except that $\log_{10}$ CD4 cell count becomes significantly associated with AIDS/death in the treatment group ($Z = 1$).

6 Discussion

Joint models of longitudinal biomarkers measured with error and an event time have been studied extensively for full cohort data. We generalized the conditional score method to case-cohort/two-phase sampling design cohort studies with Bernoulli sampling, where only subjects in the second phase sample have biomarker data. We developed IPW and AIPW techniques, providing theoretical results on consistency and asymptotic normality with resulting sandwich variance estimators, and studied their performance under correct and incorrect modeling assumptions. Focusing on nonparametric kernel regression as a component of the AIPW method, we found that AIPW outperforms IPW when there is an auxiliary variable strongly correlated with the biomarker. However, with only weak correlations, including auxiliaries can reduce efficiency and even increase bias, especially when the number of biomarker measurements per subject is limited.

Our simulation studies of misspecified measurement error models suggest that when the measurement error is non-normal but still random, the conditional score estimators are slightly biased. More serious problems arise when the measurement error violates the homoscedasticity assumption. Where deviation from the assumption of normal measurement errors is detected, it is recommended to transform the original biomarker to meet the assumption more closely. However, it is challenging to check this assumption based on observed biomarker data because of the need to first specify a correct distribution for the inherent true biomarker. Replicate samples from subjects and validation studies of the assay used to measure the biomarker can be helpful for checking and specifying a credible measurement error model.

In addition to the simulation studies of the case-control Bernoulli sampling design presented in Section 4, we also studied two-phase sampling designs that over-sample control subjects with auxiliary values that predict extreme values of the biomarker. These designs attempt to increase efficiency by increasing the variability in the longitudinal biomarker process. (See Gilbert et al (2014) for a study of this concept.) However, the resulting estimators did not provide any efficiency gains over the simpler case-control design. We conjecture that this is because we focused on scenarios with a substantial amount of measurement error and that sample all cases, which dominate the variance of the influence functions. Additional research of interest would seek scenarios where an auxiliary covariate-dependent two-phase sampling design would improve efficiency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health [Award Numbers R37AI054165, UM1AI068635], and by the Bill and Melinda Gates Foundation [Award Number OPP1110049]. The content is solely the responsibility of the authors and does not necessarily represent the

official views of the NIH or BMGF. The authors thank Ching-Yun Wang for statistical input, the AIDS Clinical Trials Group (ACTG) operations, laboratory, and statistical data management centers for generating and providing the ACTG 175 data, and the ACTG 175 study participants and investigators, in particular protocol chair Scott Hammer who reviewed and approved this work.

References

- Cao W, Tsiatis A, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*. 2009; 96(3):723–734. [PubMed: 20161511]
- Dafni U, Tsiatis A. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*. 1998; 54(4):1445–1462. [PubMed: 9883544]
- DeGruttola V, Tu X. Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*. 1994; 50(4):1003–1014. [PubMed: 7786983]
- Gilbert P, Yu X, Rotnitzky A. Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine*. 2014; 33(6):901–917. [PubMed: 24123289]
- Guo X, Carlin B. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*. 2004; 58(1):16–24.
- Hall P, Racine J, Li Q. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*. 2004; 99(468):1015–1026.
- Hammer S, Katzenstein D, Hughes M, Gundacker H, Schooley R, Haubrich R, Henry W, Lederman M, Phair J, Niu M, Hirsch M, Merigan T. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*. 1996; 335(15):1081–1090. [PubMed: 8813038]
- 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(11) pii: v43/i11/paper.
- 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 CGK, Williams Pinter A, Fong Y, Janes H, DeCamp A, Huang Y, Rao M, Billings E, Karasavvas N, Robb M, Ngaay 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. *The New England Journal of Medicine*. 2012; 366(14):1275–1286. [PubMed: 22475592]
- Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*. 2000; 1(4):465–480. [PubMed: 12933568]
- Kobayashi F, Kuroki M. A new proportion measure of the treatment effect captured by candidate surrogate endpoints. *Statistics in Medicine*. 2014; 33(19):3338–3353. [PubMed: 24782344]
- Lin D, Fleming T, DeGruttola V. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in Medicine*. 1997; 16(13):1515–1527. [PubMed: 9249922]
- Prentice R. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1982; 69(2):331–342.
- Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73(1):1–11.
- Prentice R. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*. 1989; 8(4):431–440. [PubMed: 2727467]
- Qi L, Wang C, Prentice R. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100(472):1250–1263.
- Rizopoulos D, Verbeke G, Lesaffre E. Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2009; 71(3):637–654.
- Self, S., Pawitan, Y. Modeling a marker of disease progression and onset of disease. In: Jewell, N.Dietz, K., Farewell, V., editors. *AIDS Epidemiology*. Birkhäuser; Boston: 1992. p. 231–255.

- Song X, Davidian M, Tsiatis A. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*. 2002; 3(4):511–528. [PubMed: 12933595]
- Tsiatis A, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*. 2001; 88(2):447–458.
- Tsiatis A, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*. 2004; 14:809–834.
- Tsiatis A, DeGruttola V, Wulfsohn M. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. 1995; 90(429):27–37.
- Wang C. Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*. 2006; 16:235–253.
- Wang C, Xie S, Prentice R. Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates. *Statistica Sinica*. 2001; 11:1081–1104.
- Wu L, Liu W, Yi G, Huang Y. 2012; Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*. doi: 10.1155/2012/640153
- Wulfsohn M, Tsiatis A. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997; 53(1):330–339. [PubMed: 9147598]

Fig. 1.

Simulation results for Simulation Study I: $\lambda(u) = \lambda_0(u) \exp\{\beta X(u)\}$ and correctly specified error distribution. The variables enclosed in the parentheses of AIPW() indicate which variables are used in estimating the augmentation terms, and indicate the strength of correlation with $X(u)$ ($R = 0, 0.5, 0.95$). Relative efficiency is calculated as the Monte Carlo variance of $\hat{\beta}$ from Full divided by that from IPW or AIPW method.

Fig. 2. Simulation results for Simulation Studies II(a) and II(b): $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z\}$ and correctly specified error distribution; II(b) with few measurement times. The variables enclosed in the parentheses of AIPW() indicate which variables are used in estimating the augmentation terms, and indicate the strength of correlation with $X(u)$ ($R = 0, 0.5, 0.95$).

Fig. 3. Simulation results for Simulation Studies II(c) and II(d): $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z\}$ and misspecified error distribution. The variables enclosed in the parentheses of AIPW() indicate which variables are used in estimating the augmentation terms, and indicate the strength of correlation with $X(u)$ ($R = 0, 0.5, 0.95$).

Fig. 4. Simulation results for Simulation Study III: $\lambda(u) = \lambda_0(u) \exp\{\beta X(u) + \eta Z + \gamma X(u)Z\}$ and correctly specified error distribution. The variables enclosed in the parentheses of AIPW() indicate which variables are used in estimating the augmentation terms, and indicate the strength of correlation with $X(u)$ ($R = 0, 0.5, 0.95$).

Fig. 5. ACTG 175: Spaghetti plot of observed $\log_{10}$ CD4 cell counts $X(u)$ from 10 randomly selected subjects with fitted lines from linear and quadratic mixed effects models.

Table 1

Sample sizes for Phase I (average N) and Phase II (average n) under case-control Bernoulli sampling for Simulation Studies I–III.

Simulation Study		Case ($= 1$) N (n)	Control ($= 0$) N (n)
I	β		
	0	156 (156)	1344 (443)
	$-\ln 2$	151 (151)	1349 (445)
	$-\ln 4$	158 (158)	1342 (443)
II	(β, η)		
	$(-\ln 2, 0)$	156 (156)	1344 (443)
	$(-\ln 2, -\ln 2)$	157 (157)	1343 (443)
III	(β, η, γ)		
	$(-\ln 2, -\ln 2, 0)$	157 (157)	1343 (443)

ACTG 175: Estimates and 95% confidence intervals (CIs) for the coefficients in the joint Cox model, to assess continuous log₁₀CD4 cell count as a time-dependent Prentice surrogate endpoint.

Table 2

	log ₁₀ CD4	Trt	log ₁₀ CD4 × Trt	bsl.log ₁₀ CD4
	Model 0			
	-0.37 (-0.61, -0.13)			
	IPW($\hat{\pi}$) ^a			
Model 1	-2.14 (-2.81, -1.47)	-3.33 (-5.31, -1.34)	1.42 (0.46, 2.38)	-3.60 (-5.16, -2.03)
Model 2	-1.04 (-1.93, -0.15)	-0.44 (-0.79, -0.08)		-3.61 (-5.70, -1.51)
Trt Group Z = 1	-0.73 (-1.68, 0.22)			-4.23 (-6.46, -2.00)
Control Group Z = 0	-2.89 (-3.60, -2.18)			-1.03 (-3.68, 1.62)
	AIPW($\hat{\pi}$) ^b			
Model 1	-2.35 (-3.22, -1.47)	-2.93 (-6.29, 0.43)	1.36 (-0.22, 2.94)	-1.55 (-2.90, -0.20)
Model 2	-1.17 (-1.94, -0.39)	-0.35 (-0.67, -0.02)		-2.69 (-6.61, 1.23)
Trt Group Z = 1	-0.92 (-2.24, 0.41)			-3.14 (-7.91, 1.62)
Control Group Z = 0	-3.03 (-3.84, -2.23)			-0.57 (-9.33, 8.20)

^aPTE = -0.173 (95% CI: -1.026, 0.680) (PTE = 1 - $\theta_{Trt,2} \theta_{Trt,0}$, where $\theta_{Trt,j}$ is the coefficient for Trt in Model j). PCS = 0.021 (95% CI: 0.000, 0.586) (PCS from Kobayashi and Kuroki (2014)). pNIE0 = 0.079 (95% CI: -0.836, 0.999) (pNIE0 = 1 - NDE0(26)/TE(26), where $NDE_0(26) = \{I(T^{1X^0}_\alpha \geq 26) - I(T^{0X^0}_\alpha \geq 26)\}$. We choose the Month 26 time point because it is a late follow-up time).

^bFor the AIPW method, the treatment indicator Z and the event indicator were used in estimating the augmentation terms.