Published in final edited form as:

Lifetime Data Anal. 2016 October ; 22(4): 606–625. doi:10.1007/s10985-015-9353-9.

Mark-Specific Hazard Ratio Model with Missing Multivariate Marks

Michal Juraska and

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Mail Stop M2-C200, Seattle, WA 98109, USA Tel.: +1-206-667-5568

Peter B. Gilbert

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, and Department of Biostatistics, University of Washington, 1100 Fairview Avenue North, Mail Stop M2-C200, Seattle, WA 98109, USA.

Michal Juraska: mjuraska@fredhutch.org

Abstract

An objective of randomized placebo-controlled preventive HIV vaccine efficacy trials is to assess the relationship between vaccine effects to prevent HIV acquisition and continuous genetic distances of the exposing HIVs to multiple HIV strains represented in the vaccine. The set of genetic distances, only observed in failures, is collectively termed the ‘mark.’ The objective has motivated a recent study of a multivariate mark-specific hazard ratio model in the competing risks failure time analysis framework. Marks of interest, however, are commonly subject to substantial missingness, largely due to rapid post-acquisition viral evolution. In this article, we investigate the mark-specific hazard ratio model with missing multivariate marks and develop two inferential procedures based on (i) inverse probability weighting (IPW) of the complete cases, and (ii) augmentation of the IPW estimating functions by leveraging auxiliary data predictive of the mark. Asymptotic properties and finite-sample performance of the inferential procedures are presented. This research also provides general inferential methods for semiparametric density ratio/biased sampling models with missing data. We apply the developed procedures to data from the HVTN 502 ‘Step’ HIV vaccine efficacy trial.

Keywords

Augmented inverse probability weighting; Biased sampling model; Competing risks; Cox model; Density ratio model; Missing data; Semiparametric model

Ethical statement

Institutional review board approval for the HVTN 502 Step study was obtained at all study sites. The study was undertaken in conformance with applicable local and country requirements, and participants gave written informed consent.

Supporting information

Additional supporting material referenced in this paper is available at the website of Lifetime Data Analysis:

Supplement A: Theorem proofs.

Supplement B: Additional simulation results under correctly specified missing mark models.

Supplement C: Additional robustness analysis results under mis-specified missing mark models.

Supplement D: Additional simulation results when marks are missing completely at random.

Supplement E: Additional Step trial application results.

1 Introduction

In studies with time-to-event endpoints, treatment efficacy is commonly assessed by comparing treatment-specific hazard rates of failure. In preventive HIV vaccine efficacy trials, HIV-uninfected volunteers are randomly allocated to receive vaccine or placebo, and the primary objective assesses vaccine efficacy (VE), typically defined as one minus the (vaccine/placebo) hazard ratio of HIV acquisition (Halloran et al, 1997). The vaccine trial population is exposed to circulating HIVs of extensive sequence diversity, and, in general, viral vaccines stimulate protective immune responses to HIVs with protein sequences matching those inserted in the vaccine, but fail to stimulate protective immune responses to viruses exhibiting sequence divergence from the vaccine strains. Therefore, a secondary objective assesses VE to prevent acquisition of infection with certain HIV genotypes characterized by a continuous ‘mark’ variable measuring genetic divergence of an infecting HIV to the vaccine strain(s) (a mark, defined by Cox and Oakes (1984), is only observed in failures). The mark-specific vaccine efficacy is defined as one minus the mark-specific vaccine-to-placebo hazard ratio of infection (Gilbert et al, 2008).

A biologically relevant definition of the mark variable is required to reflect the potential of the vaccine, or lack thereof, to elicit protective immune responses to divergent viruses from the vaccine strains; under an adequate mark definition, the vaccine putatively confers a higher level of protection against acquisition of HIVs with smaller distances to the vaccine strains. Multiple immunologically relevant genetic distances have been proposed, based on (i) different conceptual definitions, (ii) different HIV sequence regions, (iii) different techniques to predict HIV amino acid sites that stimulate antibody/T-cell responses, and (iv) different HIV reference sequences (Rolland et al, 2011); thus, multivariate mark variables have been considered in order to assess the dependency of VE on the joint distance distribution. We analyze bivariate genetic distances in the application in Section 7.

Considering settings with mark data observed in all failures, Gilbert et al (2004, 2008) developed estimation and testing procedures for the univariate continuous mark-specific hazard ratio. Sun et al (2009) proposed inferential procedures for the univariate continuous mark-specific proportional hazards model that allows covariate adjustment. Sun et al (2013) extended this approach to accommodate multivariate marks. Juraska and Gilbert (2013) developed estimation and testing procedures for a continuous mark-specific hazard ratio model that allows multivariate marks and improves efficiency of estimation relative to alternative approaches.

However, the aforementioned work for continuous marks did not account for missingness of the multivariate mark in failures, which commonly arises in preventive HIV vaccine trials. The goal is to characterize if and how the vaccine blocks acquisition with certain HIV strains; therefore, it is critical to measure ‘near acquisition’ or ‘early’ HIV sequences, prior to the onset of post-acquisition viral evolution, to well-approximate the transmitting strain(s). Thus, the mark of particular scientific interest is defined based on sequences measured early after acquisition, which is commonly missing in a sizeable fraction of infected subjects. The missingness rate primarily depends on the frequency of HIV testing (typically 6-monthly). Although more frequent testing is planned in future trials, the

missingness rate will remain substantial. Viral load may also be associated with missingness as levels below a limit of detection preclude sequencing of the virus. Other reasons for a missing early mark include a missing blood specimen or a technical failure in the HIV sequencing procedure. Section 7 describes an early-mark definition used in a vaccine trial.

A convenient approach in this setting is using the aforementioned ‘complete-case’ analysis methods; however, their correct application requires validity of the missing completely at random (MCAR) assumption. In addition, the complete-case analysis may be inefficient since data on subjects with missing early marks are ignored; the lack of efficiency may be severe if the missingness rate is high. Goetghebeur and Ryan (1995) studied proportional cause-specific hazards models with missing causes of failure (discrete marks). Lu and Tsiatis (2001) developed multiple imputation methods in this setting. Sun and Gilbert (2012) proposed estimation methods in the continuous mark-specific proportional hazards model with missing at random univariate marks; higher mark dimensions were not accommodated due to numerical limitations posed by the employed kernel smoothing procedure.

In this article, we extend the Juraska and Gilbert (2013) estimation and testing methods for continuous mark-specific hazard ratios to accommodate missing at random marks that are univariate or multivariate. We propose two consistent estimation procedures in semiparametric density ratio/biased sampling models, which, in the presence of complete biased sampling data, were studied by Vardi (1985), Qin (1998), Gilbert et al (1999) and Gilbert (2000). The estimation approaches are based on (i) weighting of the complete cases by the inverse of the probabilities of observing the mark of interest (Horvitz and Thompson, 1952), and (ii) augmentation of the inverse probability weighted (IPW) estimating functions by leveraging correlations between the mark and auxiliary data to “impute” the expected profile score vectors for subjects with both complete and incomplete mark data (using the general theory of Robins et al (1994)). Although motivated by the HIV vaccine trial application, this work addresses the general problem of inference in semiparametric density ratio/biased sampling models with missing data.

The remainder of the article is organized as follows. Notation, assumptions, the estimand of interest and the semiparametric model for the mark-specific hazard ratio are introduced in Section 2. We describe two estimation procedures in Section 3, followed by the presentation of the asymptotic properties of the proposed estimators in Section 4, which include double robustness. Hypothesis testing procedures are introduced in Section 5. We summarize our findings from the simulation study in Section 6 and apply the proposed inferential methods to data from the ‘Step’ HIV vaccine trial in Section 7. Theorem proofs are available in Supplement A.

2 Mark-specific hazard ratio model and missing marks

2.1 Notation

Let T and C denote the times to failure and censoring, respectively. The observed right-censored failure time is $X = \min(T, C)$ with the failure indicator $\delta = \mathbb{I}(T < C)$. Let $\mathbf{V} \in \mathbb{R}^s$ denote a continuous multivariate mark variable; without loss of generality, the support of each component of $\mathbf{V}$ is taken to be $[0, 1]$. In the competing risks framework, $\mathbf{V}$ is only

observable if $\delta = 1$. Juraska and Gilbert (2013) studied the case where $\mathbf{V}$ was assumed to be fully observed in all failures. We now allow $\mathbf{V}$ to be missing; in the HIV vaccine trial setting, we define $\mathbf{V}$ to be a multivariate distance based on a sequence measured in the early phase of HIV infection according to some operational definition. Thus, for infected subjects unobserved in the early phase, the mark $\mathbf{V}$ is missing. If $\delta = 1$, we define the indicator R of observing the mark $\mathbf{V}$ as follows: let $R = 1$ if all components of $\mathbf{V}$ are observed and let $R = 0$ otherwise (hence, we consider the ‘all-or-none’ type of missingness because a missing sequence results in each component of the multivariate distance to be missing). Let Z denote the indicator of assignment to the treatment group (in vaccine trials, $Z = 1$ indicates vaccine and $Z = 0$ indicates placebo). We denote by $\mathbf{A}$ a vector of auxiliary covariates. Our approach only requires that $\mathbf{A}$ be observed in infected subjects ($\delta = 1$) because only those observations of $\mathbf{A}$ are used by the method for predicting (i) the probability of observing $\mathbf{V}$, and (ii) the expected profile score vector. Marks $\tilde{\mathbf{V}}$ based on ‘late’ sequences, their corresponding measurement times, and/or ‘late’ viral load measurements can be included as a subset of $\mathbf{A}$. Let n be the total sample size and $(X_i, \delta_i, R_i, \delta_i R_i \mathbf{V}_i^T, Z_i, \delta_i \mathbf{A}_i^T)$, $i = 1, \dots, n$, i.i.d. replicates of $(X, \delta, \delta R, \delta R \mathbf{V}^T, Z, \delta \mathbf{A}^T)$. The observed data consist of the observations $(X_i, R_i, \delta_i \mathbf{V}_i^T, Z_i, \mathbf{A}_i^T)$ for individuals with $\delta_i = 1$ and the observations (X_i, Z_i) for those with $\delta_i = 0$. If $\delta = 1$, denote $\mathbf{W} = (Z, \mathbf{A}^T)^T$.

2.2 Assumptions

We assume C is conditionally independent of both T and $\mathbf{V}$ given Z , that is, $C \perp\!\!\!\perp T|Z$ and $C \perp\!\!\!\perp \mathbf{V}|Z$, and assume $T \perp\!\!\!\perp \mathbf{V}|Z$. Juraska and Gilbert (2013) demonstrated plausibility of the posed assumptions in the HIV vaccine trial setting, showed their necessity and sufficiency for identifiability of the Euclidean parameters in the time-independent density ratio model introduced in Section 2.4, proposed a Kolmogorov-Smirnov-type test for assessing validity of the assumption $T \perp\!\!\!\perp \mathbf{V}|Z$, and demonstrated robustness to levels of correlation between T and $\mathbf{V}$ detectable with moderate power (such correlations are of particular interest because, in these scenarios, the violation of $T \perp\!\!\!\perp \mathbf{V}|Z$ may remain undetected).

We make the following assumptions about the missing mark mechanism:

$$P(R=1|\delta=1, \mathbf{W}) = P(R=1|\mathbf{V}, \delta=1, \mathbf{W}) \quad (1)$$

and

$$\pi(\mathbf{W}) := P(R=1|\delta=1, \mathbf{W}) \geq \varepsilon \text{ with probability 1 for some } \varepsilon > 0. \quad (2)$$

Condition (1) assumes that the mark $\mathbf{V}$ is missing at random (Rubin, 1976). Condition (2) ensures that an $\frac{1}{n^2}$ -consistent estimator for ϕ in (7) exists (Robins et al, 1994). The probability of observing $\mathbf{V}$ is largely affected by the frequency of HIV testing. Viral load may also be associated with missingness as noted in the Introduction.

2.3 Estimand of interest

We define the conditional multivariate mark-specific hazard function as

$$h(t, \mathbf{v}|Z=z) = \lim_{\Delta_1, \Delta_{21}, \dots, \Delta_{2s} \searrow 0} \frac{P(T \in [t, t+\Delta_1), \mathbf{V} \in \prod_{i=1}^s [v_i, v_i+\Delta_{2i}) | T \geq t, Z=z)}{\Delta_1 \Delta_{21} \cdots \Delta_{2s}}, \quad (3)$$

which naturally extends the cause-specific hazard function in the presence of finitely many causes of failure (Prentice et al, 1978). Gilbert et al (2008) defined the mark-specific vaccine efficacy as

$$VE(t, \mathbf{v}) = 1 - h(t, \mathbf{v}|Z=1)/h(t, \mathbf{v}|Z=0), \quad (4)$$

which measures the multiplicative vaccine effect to reduce the instantaneous risk of infection by the exposing strain $\mathbf{v}$ at time t .

2.4 Mark-specific hazard ratio model

Following Juraska and Gilbert (2013), the mark-specific hazard function (3) factors as

$$h(t, \mathbf{v}|Z=z) = f(\mathbf{v}|T=t, Z=z) h(t|Z=z), \quad (5)$$

where $f(\cdot|T=t, Z=z)$ is the conditional density function of $\mathbf{V}$ given $T=t$ and $Z=z$, and $h(\cdot|Z=z)$ is the ordinary marginal hazard function ignoring mark data. Consequently, the mark-specific hazard ratio can be written as

$$\frac{h(t, \mathbf{v}|Z=1)}{h(t, \mathbf{v}|Z=0)} = \frac{f(\mathbf{v}|T=t, Z=1)}{f(\mathbf{v}|T=t, Z=0)} \frac{h(t|Z=1)}{h(t|Z=0)}, \quad (6)$$

allowing separate estimation for each of the two ratios in (6). For the mark density ratio, following the assumption $T \perp\!\!\!\perp \mathbf{V}|Z$, we consider the semiparametric density ratio model (Qin, 1998)

$$f(\mathbf{v}|Z=1)/f(\mathbf{v}|Z=0) = g(\mathbf{v}, \boldsymbol{\phi}) \quad (7)$$

with $\boldsymbol{\phi} \in \mathbb{R}^d$ the parameter of interest and $g(\mathbf{v}, \boldsymbol{\phi})$ a known time-independent weight function, continuously differentiable in $\boldsymbol{\phi}$. The nuisance parameter $f(\mathbf{v}|Z=0)$ is treated nonparametrically. The assumption $(T, C) \perp\!\!\!\perp \mathbf{V}|Z$ yields the equality $f(\mathbf{v}|Z=z) = f(\mathbf{v}|Z=z, \delta=1)$, and hence the parameter $\boldsymbol{\phi}$ in (7) is estimable using data in failures only. A popular choice of the weight function $g(\mathbf{v}, \boldsymbol{\phi})$ is $\exp\{\boldsymbol{\alpha} + \boldsymbol{g}(\mathbf{v}, \boldsymbol{\beta})\}$ where $\boldsymbol{\phi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{g}(\mathbf{v}, \boldsymbol{\beta})$ is a

polynomial function because it yields model (7) equivalent to a retrospective logistic regression model (Prentice and Pyke, 1979).

For the marginal hazard ratio in (6), we posit the Cox model $h(t|Z=1)=h(t|Z=0)=e^{\gamma}$, estimating γ by the maximum partial likelihood method or the more efficient method of Lu and Tsiatis (2008) leveraging auxiliary data predictive of the failure time (we implemented the Lu and Tsiatis method in the R `speff2trial` package). Additional, possibly time-dependent covariates, can be included in the Cox model, allowing, for example, the overall treatment effect to vary over time (further discussed in Juraska and Gilbert (2013)).

3 Estimation procedures in the density ratio model

Due to factorization (6) of the mark-specific hazard ratio, missing marks only impact estimation of the mark density ratio. Following Qin (1998) and denoting $f(v) = f(v|Z=0)$, in the presence of complete mark data in failures, the semiparametric log likelihood for the parameter of interest ϕ is $\ell(\phi) = \sum_{i \in I} \log f(V_i) + \sum_{i \in I_1} \log g(V_i, \phi)$ with the nuisance parameter $f(\cdot)$ subject to constraints $f(V_i) \geq 0$, $\sum_{i \in I} f(V_i) = 1$, and $\sum_{i \in I} f(V_i)g(V_i, \phi) = 1$ where $I = \{k : \delta_k = 1\}$ and $I_1 = \{k : \delta_k = 1 \wedge Z_k = 1\}$. To maximize $\ell(\phi)$, the Lagrange multiplier method yields the profile score functions for ϕ and the Lagrange multiplier $\lambda \in [0, 1]$ (that arises by profiling out $f(\cdot)$)

$$\begin{aligned} U_{\phi}(\phi, \lambda) &= \sum_{i \in I} \left(-\frac{\lambda g(V_i, \phi)}{1 + \lambda(g(V_i, \phi) - 1)} + Z_i \frac{g(V_i, \phi)}{g(V_i, \phi)} \right) \\ U_{\lambda}(\phi, \lambda) &= -\sum_{i \in I} \frac{g(V_i, \phi) - 1}{1 + \lambda(g(V_i, \phi) - 1)} \end{aligned}$$

where $g(v, \phi) = g(v, \phi)/\phi$. Denote $U(\phi, \lambda) = (U_{\phi}(\phi, \lambda)^T, U_{\lambda}(\phi, \lambda)^T)^T$. If the mark is completely observed in failures, the maximum profile likelihood estimator $(\hat{\phi}^T, \hat{\lambda}^T)^T$ for $(\phi^T, \lambda)^T$ is defined as the solution to the system $U(\phi, \lambda) = \mathbf{0}$. The goal is to extend the result to allow missing marks.

3.1 Inverse probability weighted complete-case estimator

The idea of the inverse probability weighted (IPW) complete-case estimator, originally proposed by Horvitz and Thompson (1952), is based on weighting of the complete cases by the inverse of the probabilities $\pi(W_i)$ or their estimates. We suppose a correctly specified parametric model $\pi(W, \psi)$ for $\pi(W)$, i.e.,

$$\pi(W) = \pi(W, \psi_0) \quad (8)$$

where ψ_0 is an unknown parameter vector and $\pi(\cdot, \psi)$ is a known smooth function taking values in $[\epsilon, 1]$ for some $\epsilon > 0$. Typically we posit a logistic model $\text{logit}\{\pi(W, \psi)\} = \psi^T c(W)$ where c is a vector function defined on the support of W . The maximum likelihood estimator $\hat{\psi}$ for ψ is then obtained by solving $\sum_{i \in I} S_{\psi, i}(\psi) = \mathbf{0}$ where $S_{\psi, i}(\psi) = (R_i - \pi(W_i, \psi)) \delta \text{logit} \pi(W_i, \psi) / \delta \psi$.

To estimate the parameter of interest ψ (and λ), we define

$U_i^{ipw}(\phi, \lambda, \psi) = R_i U_i(\phi, \lambda) / \pi(W_i, \psi)$ and $U^{ipw}(\phi, \lambda, \psi) = \sum_{i \in I} U_i^{ipw}(\phi, \lambda, \psi)$, where $U_i(\phi, \lambda)$ is the i -th individual's contribution to the profile score function $U(\phi, \lambda)$. The IPW estimator for $(\phi^T, \lambda)^T$ is defined as the solution to the IPW estimating equations $U^{ipw}(\phi, \lambda, \psi) = 0$ and denoted by $(\hat{\phi}_{ipw}^T, \hat{\lambda}_{ipw})^T$.

3.2 Augmented inverse probability weighted complete-case estimator

The IPW estimator fails to make optimal use of the available data and may be inefficient even if the model posited for $\pi(w)$ is correct. Theorem 1 shows that the estimator

$(\hat{\phi}_{ipw}^T, \hat{\lambda}_{ipw})^T$ is consistent if the probabilities $\pi(w)$ are modeled correctly; otherwise it may be biased. To improve efficiency and robustness to mis-specification of the missingness model, Robins et al (1994) proposed adding an augmented term to the IPW estimating function in a general framework. In the density ratio model (7), a more efficient estimator for $(\phi^T, \lambda)^T$ can be obtained by adding information about the conditional expectation $E[U(\phi, \lambda) | \delta = 1, W]$ into the estimation procedure (suppressing subscript i). Assumption (1) implies that $E[U(\phi, \lambda) | \delta = 1, W] = E[U(\phi, \lambda) | R = 1, \delta = 1, W]$. We posit a parametric model $q(W, \phi, \lambda, \nu)$ for $E[U(\phi, \lambda) | \delta = 1, W]$, i.e., for $(\phi^T, \lambda)^T$ in a compact set ϕ_λ^* ,

$$E[U(\phi, \lambda) | \delta = 1, W] = q(W, \phi, \lambda, \nu) \quad (9)$$

where $\nu = \nu(\phi, \lambda)$ is an unknown parameter vector and $q(\cdot, \phi, \lambda, \nu)$ is a known smooth function in $(\phi^T, \lambda)^T$. Denote $\hat{\nu} = \hat{\nu}(\phi, \lambda)$ a uniformly consistent estimator for $\nu(\phi, \lambda)$ in $(\phi, \lambda) \in \phi_\lambda^*$. We define

$$U_i^{aug}(\phi, \lambda, \psi, \nu) = U_i(\phi, \lambda) \frac{R_i}{\pi(W_i, \psi)} + q(W_i, \phi, \lambda, \nu) \left(1 - \frac{R_i}{\pi(W_i, \psi)} \right), \quad i \in I,$$

and $U^{aug}(\phi, \lambda, \psi, \nu) = \sum_{i \in I} U_i^{aug}(\phi, \lambda, \psi, \nu)$. The augmented (AUG) IPW estimator for $(\phi^T, \lambda)^T$ is defined as the solution to the augmented IPW estimating equations $U^{aug}(\phi, \lambda, \psi, \nu) = 0$ and denoted by $(\hat{\phi}_{aug}^T, \hat{\lambda}_{aug})^T$.

4 Asymptotic properties

Define $x^{\otimes 2} = xx^T$ for $x \in \mathbb{R}^p$. Let $H_i(\omega) = (U_i^{ipw}(\phi, \lambda, \psi)^T, S_{\psi, i}(\psi)^T)^T$ with $\omega = (\phi^T, \lambda, \psi^T)^T$. We assume the regularity conditions expressed as Condition A in Supplement A. In addition, let $m_0 = \sum_{i=1}^n \delta_i(1 - Z_i)$, $m_1 = \sum_{i=1}^n \delta_i Z_i$ and $m = m_0 + m_1$. Denote $\xi_{mi} = m_i/m$ and assume that $\xi_{mi} \rightarrow \xi_i > 0$, $i = 0, 1$, as $m \rightarrow \infty$.

4.1 Asymptotic properties of the inverse probability weighted complete-case estimator

The following two theorems show that $(\hat{\phi}_{ipw}^T, \hat{\lambda}_{ipw}^T)^T$ is a consistent estimator and characterize its asymptotic distribution.

Theorem 1—If (1), (2), (8), and Condition A in Supplement A are true,

$$(\phi_{ipw}^T, \lambda_{ipw}^T)^T \xrightarrow{P} (\phi^T, \lambda)^T \text{ as } m \rightarrow \infty.$$

Theorem 2—If (1), (2), (8), and Condition A in Supplement A are true, then (i) with

probability approaching 1, $(\hat{\phi}_{ipw}^T, \hat{\lambda}_{ipw}^T)^T$ exists and is unique; (ii)

$m^{\frac{1}{2}} \left((\hat{\phi}_{ipw}^T, \hat{\lambda}_{ipw}^T)^T - (\phi^T, \lambda)^T \right)$ is asymptotically normal with mean $\mathbf{0}$ and variance $\mathbf{J}^{-1} \mathbf{D} (\mathbf{J}^{-1})^T$, which can be consistently estimated by $\hat{\mathbf{J}}^{-1} \hat{\mathbf{D}} (\hat{\mathbf{J}}^{-1})^T$, where

$$\mathbf{J} = E \left\{ \partial \mathbf{U}_i^{ipw}(\phi, \lambda, \psi_0) / \partial (\phi^T, \lambda)^T \right\}, \quad \mathbf{D} = E \left\{ \text{resid}(\mathbf{U}_i^{ipw}, \mathbf{S}_{\psi,i})^{\otimes 2} \right\},$$

$$\mathbf{U}_i^{ipw} = \mathbf{U}_i^{ipw}(\phi, \lambda, \psi_0), \quad \mathbf{S}_{\psi,i} = \mathbf{S}_{\psi,i}(\psi_0),$$

with $\text{resid}(\mathbf{A}, \mathbf{B}) = \mathbf{A} - E(\mathbf{A}\mathbf{B}^T)(E(\mathbf{B}\mathbf{B}^T))^{-1}\mathbf{B}$ the residual vector from the population least squares regressions of the components of $\mathbf{A}$ on $\mathbf{B}$,

$$\hat{\mathbf{J}} = m^{-1} \sum_{i \in I} \partial \mathbf{U}_i^{ipw}(\hat{\phi}_{ipw}, \hat{\lambda}_{ipw}, \hat{\psi}) / \partial (\phi^T, \lambda)^T, \quad \hat{\mathbf{D}} = m^{-1} \sum_{i \in I} \text{Resid}(\mathbf{U}_i^{ipw}, \mathbf{S}_{\psi,i})^{\otimes 2},$$

with $\text{Resid}(\mathbf{U}_i^{ipw}, \mathbf{S}_{\psi,i})$ the residual vector for subject i from the least squares regressions of the components of $\mathbf{U}_i^{ipw}(\hat{\phi}_{ipw}, \hat{\lambda}_{ipw}, \hat{\psi})$ on $\mathbf{S}_{\psi,i}(\hat{\psi})$, $i \in I$.

4.2 Asymptotic properties of the augmented IPW complete-case estimator

In Condition A, we now consider $\mathbf{H}_i(\omega) = (\mathbf{U}_i^{aug}(\phi, \lambda, \psi, \nu)^T, \mathbf{S}_{\psi,i}(\psi)^T)^T$ with $\omega = (\phi^T, \lambda, \psi^T)^T$. The next theorem shows that $(\hat{\phi}_{aug}^T, \hat{\lambda}_{aug}^T)^T$ remains a consistent estimator if either $\pi(\mathbf{w}, \psi)$ or $q(\mathbf{w}, \phi, \lambda, \nu)$ is correctly specified, the double robustness property.

Theorem 3—Under validity of (1), (2), and Condition A in Supplement A,

$$(\phi_{aug}^T, \lambda_{aug}^T)^T \xrightarrow{P} (\phi^T, \lambda)^T \text{ as } m \rightarrow \infty \text{ if either } \pi(\mathbf{w}, \psi) \text{ or } q(\mathbf{w}, \phi, \lambda, \nu) \text{ is correctly specified.}$$

In Theorem 4, we characterize the asymptotic distribution of the AUG estimator

$$(\phi_{aug}^T, \lambda_{aug}^T)^T.$$

Theorem 4—If (1), (2), (8), (9) with $\nu = \nu_0$, and Condition A in Supplement A are true, then $m^{\frac{1}{2}} \left((\hat{\phi}_{aug}^T, \hat{\lambda}_{aug})^T - (\phi^T, \lambda)^T \right)$ is asymptotically normal with mean $\mathbf{0}$ and variance $J_*^{-1} D_* (J_*^{-1})^T$, which can be consistently estimated by $\hat{J}_*^{-1} \hat{D}_* (\hat{J}_*^{-1})^T$, where

$$J_* = E \left\{ \partial U_i^{aug}(\phi, \lambda, \psi_0, \nu_0) / \partial (\phi^T, \lambda) \right\}, \quad D_* = E \{ \text{resid}(U_i^{aug}, S_{\psi,i})^{\otimes 2} \}, \\ U_i^{aug} = U_i^{aug}(\phi, \lambda, \psi_0, \nu_0), \quad S_{\psi,i} = S_{\psi,i}(\psi_0),$$

with $\text{resid}(A, B) = A - E(AB^T)(E(BB^T))^{-1}B$ the residual vector from the population least squares regressions of the components of A on B ,

$$\hat{J}_* = m^{-1} \sum_{i \in I} \partial U_i^{aug}(\hat{\phi}_{aug}, \hat{\lambda}_{aug}, \hat{\psi}, \hat{\nu}) / \partial (\phi^T, \lambda), \quad \hat{D}_* = m^{-1} \sum_{i \in I} \text{Resid}(U_i^{aug}, S_{\psi,i})^{\otimes 2},$$

with $\text{Resid}(U_i^{aug}, S_{\psi,i})$ the residual vector for subject i from the least squares regressions of the components of $U_i^{aug}(\hat{\phi}_{aug}, \hat{\lambda}_{aug}, \hat{\psi}, \hat{\nu})$ on $S_{\psi,i}(\hat{\psi})$, $i \in I$.

The joint asymptotic distribution of the Euclidean parameter estimators in the density ratio and Cox models is required for asymptotic-based analytic inference about $VE(t, \nu)$. To this end, analogous IPW and AUG versions of Theorem 1 (and Corollary 1) in Juraska and Gilbert (2013) can be established by replacing the complete-case profile score $U_A(\phi, \lambda)$ with $U_i^{ipw}(\phi, \lambda, \hat{\psi})$ and $U_i^{aug}(\phi, \lambda, \hat{\psi}, \hat{\nu})$, respectively. The results subsequently lead to the construction of IPW and AUG versions of asymptotic pointwise Wald confidence interval for $VE(t, \nu)$.

5 Hypothesis testing

Under complete mark data in failures, Juraska and Gilbert (2013) proposed Wald-type procedures for testing whether the mark-specific hazard ratio is (i) unity or (ii) independent of the mark. Problems (i) and (ii) are equivalent to the null hypothesis of zero VE against any exposing virus and VE not varying by viral divergence, respectively. In the presence of missing marks, we consider IPW and AUG versions of the Wald test statistics that are the same as those in Juraska and Gilbert (2013) except they replace the complete-mark estimator $\hat{\phi}$ and its variance with $\hat{\phi}_{ipw}$ or $\hat{\phi}_{aug}$ and the corresponding variance.

6 Simulation study

We conducted a simulation study to investigate finite-sample performance of the proposed estimation and testing procedures in the presence of missing marks. The augmented IPW estimator (AUG) for $VE(\nu) = 1 - e^{a+\beta\nu+\gamma}$, $\nu \in [0, 1]$, i.e., assuming a monotone decline of $VE(\nu)$ in ν , is compared to the complete-case estimator (CC), which ignores information about failures with missing marks, and the IPW complete-case estimator. We additionally compare the aforementioned estimators to the (unobtainable) full data likelihood estimator

(Full) that uses the complete set of marks before a fraction of them is deleted. Finally, size and power of the associated Wald tests are examined.

The simulation setup emulates typical placebo-controlled HIV VE trials with a 2–3 fold higher infection rate to aid efficiency of the simulation study. We specify exponential failure times T with rates η and ηe^γ in the placebo and vaccine group, respectively, where $\eta = \log(0.7)/(-3)$ is chosen so that 70% of failure times in the placebo group are administratively censored at year 3. Independently of T , we specify censoring times C to be Uniform(0, 15) in each group. The observed time on study X is defined as $\min(T, C, 3)$. The vaccine-to-placebo assignment ratio is 1:1. A univariate complete mark V for placebo and vaccine failures is specified to follow distributions with density functions $f(v|Z=0) = \{2e^{-2v}/(1 - e^{-2})\} I(0 \leq v \leq 1)$ and $f(v|Z=1) = f(v|Z=0)e^{\alpha+\beta v}$, respectively, where, for a given value of β , the value of the parameter $\alpha = \alpha(\beta)$ is defined as the solution to $\int_0^1 f(v|Z=0)e^{\alpha+\beta v} dv = 1$. This mechanism ensures that the mark V follows model (7) with $g(v, \alpha, \beta) = e^{\alpha+\beta v}$.

We consider simulation scenarios (S1)–(S4) characterized, respectively, by the following values of (β, γ) : (0, 0), (0.5, -0.8), (1.2, -0.2) and (2.1, -1.3). The per-arm sample size $N = 1481$ corresponds to the expected number of $N_{PI} = 400$ observed placebo failures by year 3. In the vaccine group, the expected number of observed failures N_{VI} additionally depends on the marginal log hazard ratio, and thus, in scenarios (S1)–(S4), $N_{VI} = 400, 188, 334$, and 116. The pertaining $VE(v)$ functions in (S1)–(S4) are displayed in Supplementary Figure 1; for ease of orientation, in subsequent tables and figures, scenarios (S1)–(S4) are interchangeably labeled by their corresponding $VE(0)$ and $VE(1)$ values. The results are based on 10^4 replicated data sets.

6.1 Assessment of the IPW and AUG estimation and testing procedures under correctly specified missing mark models

We generate complete-case indicators R with conditional probabilities $\pi(W) = P(R = 1 | \delta = 1, W)$ satisfying the models

$$\begin{aligned} \text{(M1):} & \text{logit } \pi(W, \psi) = \psi_0 + \psi_1 Z + \psi_2 A + \psi_3 Z A, \\ \text{(M2):} & \text{logit } \pi(W, \psi) = \psi_0 + \psi_1 Z + \psi_2 A^* + \psi_3 Z A^*, \\ \text{(M3):} & \text{logit } \pi(W, \psi) = \psi_0 + \psi_1 Z. \end{aligned}$$

We assume a continuous auxiliary variable A representing, e.g., a distance based on a sequence measured in the post-seroconversion phase, that, conditional on V , follows the model

$$A = (1 + \kappa)^{-1} (V + \kappa U), \quad \kappa > 0, \quad (10)$$

where $U \sim \text{Uniform}(0, 1)$, independent of V . The parameter κ governs the level of association between V and A . For each of (M1)–(M3), we evaluate AUG estimators corresponding to six scenarios with varying levels of correlation between V and A : AUG-1

for $\kappa = 0.2$ corresponding to $\rho \approx 0.98$, AUG-2 for $\kappa = 0.4$ with $\rho \approx 0.92$, AUG-3 for $\kappa = 0.8$ with $\rho \approx 0.76$, AUG-4 for $\kappa = 1.6$ with $\rho \approx 0.5$, AUG-5 for $\kappa = 3.6$ with $\rho \approx 0.25$, and AUG-6 for $\kappa = 10^5$ with $\rho \approx 0$ where ρ denotes the correlation coefficient between V and A . For model (M1), we analogously study six IPW estimators, whereas each of models (M2) and (M3) evaluates a single IPW estimator. In model (M2), we consider a dichotomous auxiliary covariate A^* , e.g., a near vs. distant post-acquisition virus, or a low vs. high viral load, that is generated in two steps: first, generate $\tilde{A}$ following (10) with $\kappa = 0.4$, and second, generate A^* from Bernoulli($\tilde{A}$). We include settings with high correlations because of their feasibility in real data sets as between-subject HIV sequence diversity is considerably larger than within-subject HIV sequence diversity (Keele et al, 2008).

In (M1)–(M3), we consider the following values of ψZ with respective missing mark rates among failures in the placebo and vaccine group:

- (M1) $\psi = (-2, 0.4, 0.5, 0.8)$ resulting in $\approx 87\%$ and 73% of failures with missing marks for $\kappa = 0.4$ (minor variations for the other κ values);
- (M2) $\psi = (-2, 0.4, 0.5, 0.8)$ resulting in $\approx 86\%$ and 70% of failures with missing marks;
- (M3) $\psi = (-0.8, 0.5)$ resulting in $\approx 69\%$ and 58% of failures with missing marks.

The AUG estimator additionally requires estimation of the parameter $E[U(\phi, \lambda)|W]$. Rotnitzky and Robins (1995) advocate the use of highly parameterized regression models in this setting. We posit a linear regression model $E[U(\phi, \lambda)|W] = q(W, \nu)$ of the form

$$E[U(\phi, \lambda)|W] = \nu_0 + \nu_1 Z + \nu_2 A + \nu_3 A^2 + \nu_4 Z A; \quad (11)$$

fitted for subjects with $R_i = 1$ by the ordinary least squares method. Predicted values $\hat{E}[U(\phi, \lambda)|W_i]$ used to construct the AUG estimator are specified for subjects with $\delta_i = 1$ as $\hat{E}[U(\phi, \lambda)|W_i] = q(W_i, \hat{\nu})$. In each simulation scenario, the model (11) is exible but likely to be mis-specified. More recently, van der Laan and Rose (2011) described the targeted learning approach, which (i) utilizes cross-validation for selection of the optimal estimator, and (ii) considers targeted estimation of the data-generating distribution toward the parameter of interest. We conjecture that targeted learning would improve estimation of $E[U(\phi, \lambda)|W]$; however, its computational challenges exceed the scope of this paper.

Figures 1–2 and Supplementary Figures 2–5 present operating characteristics of four estimation procedures – Full, AUG, IPW and CC – under missing mark model (M2); results under (M1) and (M3) are included in Supplement B. Finite-sample bias of the AUG and IPW estimators for $VE(\nu)$ is small under each of (M1)–(M3) and increases in magnitude as ρ approaches 0. It tends to be smaller for the AUG than the IPW estimator, sometimes markedly so (Figure 1 and Supplementary Figure 12). Bias of the AUG-1 estimator tends to be close to that of the Full estimator representing the unachievable benchmark. Under (M1) and (M2), the CC estimator is substantially biased illustrating the inappropriateness of its use when marks are missing at random (Supplementary Figure 6 and Figure 1). The change

in bias along the mark support is similar for each estimation procedure. It is smallest for v in the central region and tends to increase in magnitude in the right tail of the mark distribution.

Median asymptotic standard error estimates of $\widehat{VE}(v)$ indicate that substantial efficiency gain of the AUG estimator may be achieved relative to the IPW estimator for $\rho > 0.7$. The most pronounced efficiency gains are observed under (M2) and (M3) for $\rho > 0.7$ (Figure 2 and Supplementary Figure 14), with the AUG-1 estimator nearing efficiency of the Full estimator. Under (M1), AUG- k proves to be more efficient than IPW- k for $k = 2$; for $k > 2$, the efficiency gain is unobserved. In each setting, the CC estimator is highly inefficient.

Coverage probabilities of 95% pointwise confidence intervals are near the nominal confidence level for scenarios (S1)–(S2) and $\rho > 0.7$. For scenarios (S3)–(S4) and $\rho = 0.5$, coverage probabilities fluctuate moderately. We verified that, in (S3)–(S4), adequate coverage probabilities were attained for lower missing mark rates if $\rho > 0.7$. Supplement B additionally summarizes operating characteristics of estimating the mark coefficient β in density ratio model (7).

Table 1 and Supplementary Tables 7–9 report size and power of three Wald tests, each based on four estimation procedures – Full, AUG, IPW and CC: two-sided Wald test and one-sided weighted Wald-type test of $H_0^0: VE(v)=0$ for all $v \in [0, 1]$ (or, equivalently, $\tilde{H}_0^0: \beta=0$ and $\gamma = 0$), and two-sided Wald test of $H_0: VE(v) = VE$ for all $v \in [0, 1]$ (or, equivalently, $\tilde{H}_0: \beta = 0$). (All three tests are characterized in detail in Juraska and Gilbert (2013) for complete mark data.) Sizes of the Wald tests induced by the AUG and IPW estimators are in accordance with the nominal significance level under each of missing mark models (M1)–(M3). In contrast, for the CC estimator, the size of each Wald test tends to be too large.

6.2 Robustness analysis of the IPW and AUG estimation and testing procedures under mis-specified missing mark models

We investigate robustness of the proposed estimators and hypothesis tests to mis-specification of $\pi(\mathbf{w}, \psi)$ and to violation of the missing at random assumption. To study robustness to mis specification of $\pi(\mathbf{w}, \psi)$, we assume validity of model (M3) while the complete-case indicator R is generated with the conditional probability $\pi(\mathbf{W}, \psi)$ satisfying

$$(M4): \text{logit } \pi(\mathbf{W}, \psi) = \psi_0 + \psi_1 Z + \psi_2 X.$$

We consider $\psi = (-0.8, 0.5, -0.5)$ such that the mark is missing in approximately 82% and 71% of failures in the placebo and vaccine group, respectively.

To examine robustness to violation of the missing at random assumption, we assume validity of model (M3) while the complete-case indicator R depends on V conditionally on W following the model

$$(M5): \text{logit } \pi(\mathbf{W}, \boldsymbol{\psi}) = \psi_0 + \psi_1 Z + \psi_2 V.$$

Here we consider $\boldsymbol{\psi} = (-0.1, 0.5, -2)$ which results in approximately 71% and 63% of failures with missing marks in the placebo and vaccine group, respectively.

The IPW estimator is moderately biased when the missing mark model is mis-specified, while the AUG estimator exhibits smaller bias approaching that of the Full estimator with an increase in ρ (Supplementary Figures 18–19). Violation of the MAR assumption induces severe bias of the IPW estimator (Supplementary Figures 22–23). On the contrary, bias of the AUG estimator is markedly smaller for $\rho > 0.7$. Under both (M4) and (M5), bias of the CC estimator is severe.

Under (M4), the IPW estimator is visibly less efficient than the AUG estimator for $\rho > 0.7$, with the AUG-1 estimator retaining efficiency close to that of the Full estimator (Supplementary Figure 20). Under (M5), the AUG estimator appears to be robust to violation of the MAR assumption for the considered scenarios with $\rho > 0.7$, with the AUG-1 estimator exhibiting minimal efficiency loss compared to the Full estimator (Supplementary Figure 24). In (M4) and (M5), all estimated weights $\pi(\mathbf{W}_i, \hat{\boldsymbol{\psi}})$ were bounded away from zero. Supplement C additionally reports robustness analysis results in estimation of the density ratio coefficient β .

Size of the Wald tests appears to be robust to model mis-specification (Supplementary Tables 14–15). Under violation of the MAR assumption, size of each Wald test tends to slightly exceed the nominal significance level.

6.3 Assessment of the IPW and AUG estimators versus a simpler complete-case single-imputation estimator when marks are missing completely at random

It is of interest to evaluate whether we lose efficiency by using the missing at random IPW and AUG methods if marks are missing completely at random (MCAR) such that simpler complete-case methods could be used. Relatedly, for applications with auxiliaries A highly predictive of V , it is of interest to compare the IPW and AUG estimators to a simpler complete-case estimator. In particular, we consider an estimator that fills in each missing V based on a parametric regression model linking the mean of V to A and then treats the imputed V 's as known values; this is a regression calibration type estimator that we refer to as the complete-case after imputation (CCIM) estimator.

To compare IPW vs. AUG vs. CC-IM, we add a simulation scenario that assumes validity of the (over-parametrized) model (M1) while the complete-case indicator R satisfies the MCAR model

$$(M6): \text{logit } \pi(\mathbf{W}, \boldsymbol{\psi}) = \psi_0.$$

We consider $\psi_0 = -0.85$ which results in approximately 70% of failures with missing marks in both the placebo and vaccine groups. For each level of correlation between V and A

specified in Section 6.1, missing V 's for the CC-IM estimator are filled in with predicted values from a beta regression model (Ferrari and Cribari-Neto, 2004) with the logit link and A as a lone predictor, fitted using complete cases. We compare the Monte-Carlo variance of all estimators as the derivation of the analytic variance of the CC-IM estimator is beyond the scope of this article.

Supplementary Figure 29 indicates that the IPW and AUG estimators do not lose efficiency if marks are MCAR and no auxiliary information about V is available (i.e., $\rho \approx 0$).

Additionally, as ρ increases, the IPW and AUG estimators gain efficiency over the CC estimator (Supplementary Figure 28). Under MCAR, finite-sample bias of the CC-IM estimator is small for highly correlated V and A and becomes larger for $\rho < 0.7$, particularly in scenario (S4) (Supplementary Figures 26–27). Overall, the CC-IM estimator has similar performance as the AUG estimator in this simulation study, making it appealing in practice given it is simpler to implement. However, as a regression calibration estimator CC-IM has the theoretical disadvantage of not being a consistent estimator.

7 Application

In the Step Phase IIb HIV vaccine trial, 1836 HIV-negative men were randomized to receive either the Merck adenovirus type 5 (Ad5) vector vaccine (MRKAd5) or placebo (Buchbinder et al, 2008). Of the 1836 subjects, 88 acquired HIV infection prior to unblinding (52 in the vaccine and 36 in the placebo group). Single-genome-amplification HIV sequences were successfully measured in 65 of the 88 infections. It is of particular scientific interest to analyze sequence data collected early upon HIV acquisition, prior to the onset of post-acquisition HIV evolution. We define the early mark as satisfying at least one of the following two conditions: (i) the sample originates from the estimated Fiebig stage I or II with negative ELISA and negative Western Blot (Keele et al, 2008); and (ii) all measured Nef sequences from an infected subject are identical (this lack of detected viral diversification suggests limited time elapsed between acquisition and virus sampling). Of the total 88 infected subjects included in the analysis, 22 (25.0%) have observed early marks (11 vaccine, 11 placebo).

The Step trial showed the estimated overall vaccine-to-placebo hazard ratio (HR) of 1.50 (95% CI, 0.95–2.41, $p = 0.06$) suggesting that the vaccine elevated the risk of HIV acquisition. Consequently, we sought to study if the elevated risk was greater against some HIV genotypes. To this end, we analyzed distances defined based on four HIV proteins, Gag, Pol, and Nef, each of which was represented in the vaccine, and the complement of the whole HIV genome outside the vaccine (Env-Rev-Tat-Vif-Vpr-Vpu) using the central HXB2 reference sequence. For each genomic region, bivariate distances (V_1 , V_2) were computed based on two bioinformatics methods, EpiPred (Heckerman et al, 2007) (defining V_1) and NetMHC (Buus et al, 2003) (defining V_2), used to predict, for each infected subject, HIV peptides that could bind to the subject's HLA molecules, and thus be part of potential T-cell epitopes. These distances, calculated as described in Rolland et al (2011), quantify the amino acid divergence in known and predicted T-cell epitopes present in the reference sequence compared to the subjects' sequences, with EpiPred based on a training set of

known T-cell epitopes and putative non-T-cell epitopes, and NetMHC based on predicted binding strength between HLA molecules and HIV peptides.

To predict the probability of observing the early mark, we used a logistic regression model with the number of days from randomization to the first HIV PCR+ test as the lone predictor (this variable is predictive because HIV diagnostic testing was more frequent early in follow-up). In addition, using the AUG estimation method, we specified a linear regression model for the expected profile scores conditional on the number of days from randomization to the first HIV PCR+ test, the vaccine group indicator, and viral load sampled at the date of the first PCR+ test. These three auxiliary covariates were selected from available demographics, host genetics and post-infection biomarker measurements using all-subsets model selection by BIC. Each of these auxiliaries was a weak predictor, which is not surprising given that heuristically none of these auxiliaries is expected to be related to HIV sequence distance. However, due to resource constraints the study did not measure the critical auxiliary information that would be highly predictive (HIV sequence data measured at a later post-acquisition sampling time point), and, given this absent auxiliary, empirical selection of the best predictive model from all input features regardless of their level of biological plausibility was deemed a reasonable approach.

We specified the time from randomization to the first HIV PCR+ test as the failure time variable T in the analysis. Each mark variable was standardized by subtracting its minimum and dividing by its range. Figure 3 shows the distribution of the observed V_1 and V_2 distances by the vaccine/placebo group for each HIV protein. We initially fit density ratio model (7) with the weight function

$$g(v, \alpha, \beta) = \exp\{\alpha + \beta_1 v_1 + \beta_2 v_2 + \beta_{12} v_1 v_2\} \quad (12)$$

and found no significant evidence for an interaction $\beta_{12} = 0$; the Wald tests yielded two-sided interaction p -values of 0.55 for Gag, 0.14 for Pol, 0.18 for Nef, and 0.50 for Env-Rev-Tat-Vif-Vpr-Vpu. Subsequently, model (12) excluding the interaction term was used for reporting inferential results for each HIV protein.

Due to the higher infection rate in the vaccine group, instead of the mark-specific VE, we present the results as the mark-specific hazard ratio $HR(v_1, v_2) = \exp\{\alpha + \beta_1 v_1 + \beta_2 v_2 + \gamma\}$ characterizing whether and how the vaccine selectively elevated susceptibility to HIV infection. The estimated $HR(v_1, v_2)$ surfaces for each protein are displayed in Supplementary Figure 30, with point and interval estimates for selected $(v_1, v_2) = (0.2, 0.2)$, $(0.5, 0.5)$ and $(0.8, 0.8)$ reported in Table 2. We observe that the standard error estimates for $\widehat{HR}(v_1, v_2)$ are markedly smaller than those obtained in the complete-case proportional hazards model fit by Sun et al (2013). We found some evidence that the vaccine selectively modified the HIV acquisition risk in the Nef reference region ($p = 0.01$, Table 2), suggesting that vaccine-enhancement of HIV acquisition risk may have been stronger against HIVs with Nef sequences more similar to the vaccine. No such evidence was indicated in the other HIV proteins.

The assumption $T \perp (V_1, V_2) | Z$ was verified to be reasonable by application of the complete-case Kolmogorov–Smirnov-type diagnostic test (Juraska and Gilbert, 2013), which, using 10^3 boot-strap iterations, yielded p -values in the 0.15–0.99 range in the vaccine and placebo groups for each HIV protein. The complete-case Qin and Zhang (1997) goodness-of-fit test of the final bivariate-mark density ratio model did not reject ($p = 0.46$ for Gag, 0.62 for Pol, 0.55 for Nef, and 0.22 for Env-Rev-Tat-Vif-Vpr-Vpu) (an extension of the diagnostic tests for incomplete mark data was beyond the scope of this paper). The proportional marginal hazards assumption was verified using the Grambsch and Therneau (1994) test ($p = 0.69$).

8 Discussion

Growing complexity of HIV vaccine designs and commonly missing early sequence data have motivated the development of inferential procedures in the multivariate mark-specific hazard ratio model accounting for missing at random marks. Although motivated by this specific scientific application, the article presents general methods of analysis of time-to-event data with continuous, possibly multivariate, marks subject to missingness at random. The presented methods also address the general problem of parameter estimation and hypothesis testing in density ratio/biased sampling models with missing data. Exponential forms of the weight function yield density ratio models that are equivalent to retrospective logistic regression models (Prentice and Pyke, 1979), suggesting an alternative approach to inference based on weighted estimating equations, developed in a general framework by Robins et al (1994) and further discussed in Zhao et al (1996).

We developed an IPW and an augmented IPW (AUG) estimator for $VE(v)$. The AUG estimator is doubly robust (Scharfstein et al, 1999), performs well under correctly specified missing mark models, shows considerable robustness and efficiency under the considered mis-specified missing mark models, and maintains efficiency under MCAR mark models with no auxiliary information about the mark. The AUG estimator is superior to the IPW estimator in each setting and is recommended for use. It is of note that the AUG estimator carries potential to be severely biased if (i) both models in (8) and (9) are mis-specified, or (ii) the estimated weights $\pi(W_i, \psi O)$ are outliers close to zero for some observations (e.g., Kang and Schafer (2007)). The approach of Tan (2006) partially addresses these shortcomings by using an alternative likelihood-based estimator for $\pi(w)$. To improve robustness and efficiency of the AUG estimator to (i) and (ii), a modified doubly robust AUG estimator following the approach of Cao et al (2009) could be constructed in this setting.

In future work, it would be of interest to compare the AUG estimators to estimators based on multiple imputation, which would constitute an extension of Lu and Tsiatis (2001) to the case of a continuous mark V . A multiple imputation estimator would potentially improve efficiency given an approximately correctly specified model for predicting V ; however, it would lack the double robustness property and hence may have more bias if it is difficult to model V . In future work, it would also be of interest to compare the AUG approach to measurement error approaches, for example based on profile likelihood or the conditional score function, and our initial work on a regression calibration estimator (Section 6.3) showed promising performance. Similar to the multiple imputation approaches, these

methods would potentially improve efficiency under sufficiently accurate modeling of V but may have more bias in some contexts due to the lack of double robustness.

To the best of our knowledge, the problem of hypothesis testing for $VE(\mathbf{v})$ [or, equivalently, $HR(\mathbf{v})$] with missing multivariate $\mathbf{v}$ has not been addressed in the existing published literature. For a univariate missing mark, the presented estimation method is an alternative to the Sun and Gilbert (2012) method, which estimates the same $VE(\mathbf{v})$ estimand under the MAR assumption. In this setting, the discussion in Juraska and Gilbert (2013) applies and provides guidance for method selection. The *R* code implementing the proposed methods is available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the participants, investigators, and sponsors of the HVTN 502 Step HIV vaccine trial. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Numbers UM1AI068635 and R37AI054165 and by the Bill and Melinda Gates Foundation (BMGF) 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.

References

- Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, Gilbert PB, Lama JR, Marmor M, del Rio C, McElrath MJ, Casimiro DR, Gottesdiener KM, Chodakewitz JA, Corey L, Robertson MN. the Step Study Protocol Team. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of- concept trial. *Lancet*. 2008; 372(9653):1881–1893. [PubMed: 19012954]
- Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S. Sensitive quantitative predictions of peptide-mhc binding by a ‘query by committee’ artificial neural network approach. *Tissue Antigens*. 2003; 62(5):378–384. [PubMed: 14617044]
- Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*. 2009; 96(3):723–734. DOI: 10.1093/biomet/asp033 [PubMed: 20161511]
- Cox, DR.; Oakes, D. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman and Hall; 1984. Analysis of Survival Data.
- Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*. 2004; 31(7):799–815.
- Gilbert PB. Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics*. 2000; 28(1):151–194.
- Gilbert PB, McKeague IW, Sun Y. The 2-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. *Biostatistics*. 2008; 9(2):263–276. [PubMed: 17704528]
- Gilbert PB, Lele SR, Vardi Y. Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*. 1999; 86(1):27–43.
- Gilbert PB, McKeague IW, Sun Y. Tests for comparing mark-specific hazards and cumulative incidence functions. *Lifetime Data Analysis*. 2004; 10:5–28. [PubMed: 15130048]
- Goetghebeur E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika*. 1995; 82(4):821–833.
- Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994; 81(3):515–526. DOI: 10.1093/biomet/81.3.515

- Halloran ME, Struchiner CJ, Longini IM. Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *American Journal of Epidemiology*. 1997; 146(10):789–803. [PubMed: 9384199]
- Heckerman D, Kadie C, Listgarten J. Leveraging information across hla alleles/supertypes improves epitope prediction. *Journal of Computational Biology*. 2007; 14:736–746. [PubMed: 17691891]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47(260):663–685.
- Juraska M, Gilbert PB. Mark-specific hazard ratio model with multivariate continuous marks: An application to vaccine efficacy. *Biometrics*. 2013; 69(2):328–337. DOI: 10.1111/biom.12016 [PubMed: 23421613]
- Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*. 2007; 22(4):523–539.
- Keele B, Giorgi E, Salazar-Gonzalez J, Decker J, Pham K, Salazar M, Sun C, Grayson T. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences*. 2008; 105:7552–7557.
- van der Laan, M.; Rose, S. *Springer Series in Statistics*. Springer; 2011. Targeted Learning.
- Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*. 2001; 57:1191–1197. [PubMed: 11764260]
- Lu X, Tsiatis AA. Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*. 2008; 95(3):679–694. DOI: 10.1093/biomet/asn003
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66(3):403–411. DOI: 10.2307/2335158
- Prentice RL, Kalbeisch JD, Peterson JAV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978; 34(4):541–554. [PubMed: 373811]
- Qin J. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*. 1998; 85(3):619–630.
- Qin J, Zhang B. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1997; 84(3):609–618.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89(427):846–866.
- Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, Heath L, Magaret CA, Bose M, Bradfield A, O'Sullivan A, Crossler J, Jones T, Nau M, Wong K, Zhao H, Raugi DN, Sorensen S, Stoddard JN, Maust BS, Deng W, Hural J, Dubey S, Michael NL, Shiver J, Corey L, Li F, Self SG, Kim J, Buchbinder S, Casimiro DR, Robertson MN, Duerr A, McElrath MJ, McCutchan FE, Mullins JI. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nature Medicine*. 2011; 17(3):366–U168. DOI: 10.1038/nm.2316
- Rotnitzky A, Robins JM. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*. 1995; 82(4):805–820. DOI: 10.1093/biomet/82.4.805
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63(3):581–590.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semi-parametric nonresponse models: Rejoinder. *Journal of the American Statistical Association*. 1999; 94(448): 1135–1146.
- Sun Y, Gilbert PB. Estimation of stratified mark-specific proportional hazards models with missing marks. *Scandinavian Journal of Statistics*. 2012; 39:34–52.
- Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. *Ann Stat*. 2009; 37(1):394–426. DOI: 10.1214/07-AOS554 [PubMed: 19838313]
- Sun Y, Li M, Gilbert PB. Mark-specific proportional hazards model with multivariate continuous marks and its application to HIV vaccine efficacy trials. *Biostatistics*. 2013; 14(1):60–74. [PubMed: 22764174]
- Tan Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*. 2006; 101(476):1619–1637.

- Vardi Y. Empirical distributions in selection bias models. *The Annals of Statistics*. 1985; 13(1):178–203.
- Zhao LP, Lipsitz S, Lew D. Regression analysis with missing covariate data using estimating equations. *Biometrics*. 1996; 52(4):1165–1182. DOI: 10.2307/2532833 [PubMed: 8962448]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Bias of estimation for $VE(v)$ in simulation scenarios (S1)–(S4) using four procedures – Full, IPW, CC and AUG – for the missing mark model (M2). Estimators AUG-1 corresponding to $\rho \approx 0.98$, AUG-2 for $\rho \approx 0.92$, and AUG-3 for $\rho \approx 0.76$ are evaluated here, while estimators AUG-4 for $\rho \approx 0.5$, AUG-5 for $\rho \approx 0.25$, and AUG-6 for $\rho \approx 0$ are evaluated in Supplementary Figure 2.

Fig. 2.

Median asymptotic standard error estimates of $\widehat{VE}(v)$ in simulation scenarios (S1)–(S4) using four procedures – Full, IPW, CC and AUG – for the missing mark model (M2). Estimators AUG-1 corresponding to $\rho \approx 0.98$, AUG-2 for $\rho \approx 0.92$, and AUG-3 for $\rho \approx 0.76$ are evaluated here, while estimators AUG-4 for $\rho \approx 0.5$, AUG-5 for $\rho \approx 0.25$, and AUG-6 for $\rho \approx 0$ are evaluated in Supplementary Figure 3.

Fig. 3. Distribution of the Step early distances to the vaccine or HXB2 reference sequence based on Gag, Pol, Nef, and Env-Rev-Tat-Vif-Vpr-Vpu sites identified using EpiPred and NetMHC epitope prediction methods.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Size of Wald tests of $\{H_0^0: VE(v)=0 \text{ for all } v \in [0, 1]\}$ and $\{H_0: VE(v) = VE \text{ for all } v \in [0, 1]\}$ using four procedures – Full, IPW, CC and AUG – under missing mark models (M1)–(M3). Estimators AUG-1 corresponding to $\rho \approx 0.98$, AUG-2 for $\rho \approx 0.92$, and AUG-3 for $\rho \approx 0.76$ are evaluated here, while estimators AUG-4 for $\rho \approx 0.5$, AUG-5 for $\rho \approx 0.25$, and AUG-6 for $\rho \approx 0$ are evaluated in Supplementary Table 7. For model (M1), the IPW estimator considers the case $\rho \approx 0.98$.

Table 1

Missingness model	Estimator					
	Full	AUG-1	AUG-2	AUG-3	IPW	CC
Size of Wald test of $H_0^0 (\alpha=0.05)$						
(M1)	0.049	0.049	0.054	0.055	0.046	0.081
(M2)	0.049	0.049	0.050	0.050	0.050	0.075
(M3)	0.049	0.048	0.055	0.056	0.048	0.048
Size of weighted Wald-type test of $H_0^0 (\alpha=0.025)$						
(M1)	0.023	0.023	0.025	0.028	0.025	0.028
(M2)	0.023	0.023	0.026	0.023	0.024	0.029
(M3)	0.023	0.023	0.025	0.029	0.024	0.026
Size of Wald test of $H_0 (\alpha = 0.05)$						
(M1)	0.050	0.048	0.054	0.056	0.048	0.102
(M2)	0.050	0.050	0.054	0.050	0.054	0.088
(M3)	0.050	0.050	0.052	0.054	0.047	0.046

Inference about bivariate mark-specific hazard ratios with marks based on EpiPred (V_1) and NetMHC (V_2) epitope prediction methods in 4 HIV genomic regions in the Step trial.

Table 2

Gene	HR(0.2, 0.2)			HR(0.5, 0.5)			HR(0.8, 0.8)			P-value $H_0^{\dagger}$	P-value $H_0^{\ddagger}$
	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI		
Gag	0.15	0.16	(0.02, 1.24)	0.85	0.44	(0.31, 2.34)	4.94	4.18	(0.94, 25.93)	0.05	0.10
Pol	1.46	0.58	(0.67, 3.19)	1.46	0.61	(0.65, 3.30)	1.46	1.44	(0.21, 10.07)	0.20	0.46
Nef	3.29	2.10	(0.94, 11.52)	2.30	0.89	(1.07, 4.93)	1.61	1.22	(0.36, 7.16)	0.01	0.01
Env [§]	2.11	1.58	(0.48, 9.18)	1.47	0.49	(0.77, 2.84)	1.03	0.43	(0.46, 2.34)	0.07	0.15

[§] Env-Rev-Tat-Vif-Vpr-Vpu

[†] $H_0^0: HR(v_1, v_2) = 1$ for all $(v_1, v_2) \in [0, 1]^2$

[‡] $H_0: HR(v_1, v_2) = HR$ for all $(v_1, v_2) \in [0, 1]^2$