Published in final edited form as:
J R Stat Soc Ser C Appl Stat. 2015 January 1; 64(1): 49–73. doi:10.1111/rssc.12067.

Inferences on relative failure rates in stratified mark-specific proportional hazards models with missing marks, with application to HIV vaccine efficacy trials

Peter B. Gilbert[*] and
Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

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

Abstract

This article develops hypothesis testing procedures for the stratified mark-specific proportional hazards model in the presence of missing marks. The motivating application is preventive HIV vaccine efficacy trials, where the mark is the genetic distance of an infecting HIV sequence to an HIV sequence represented inside the vaccine. The test statistics are constructed based on two-stage efficient estimators, which utilize auxiliary predictors of the missing marks. The asymptotic properties and finite-sample performances of the testing procedures are investigated, demonstrating double-robustness and effectiveness of the predictive auxiliaries to recover efficiency. The methods are applied to the RV144 vaccine trial.

Keywords

Auxiliary marks; competing risks failure time data; proportional hazards model; genetic data; augmented inverse probability weighting; semiparametric model

1 Introduction

The primary objective of a preventive HIV vaccine efficacy trial is to assess vaccine efficacy (VE) to prevent HIV infection, where typically VE is defined as one minus the hazard ratio (vaccine/placebo) of HIV infection diagnosis. However, the great genetic variability of HIV poses a central challenge to developing a highly efficacious vaccine (Fauci et al., 2008). The trial population is exposed to many HIV genotypes but the vaccine only contains a few, and the vaccine is less likely to protect against HIVs with greater genetic distance from the sequences inside the vaccine (Gilbert et al., 1999). The trial has objectives to assess whether and how the vaccine impacts the infection rate with any HIV genotype and whether and how the vaccine effect varies by HIV genotype; assessment of

[*]Corresponding author's contact information: pgilbert@scharp.org, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, PO Box 19024, Seattle, WA, 98109, USA.

these objectives has been named ‘sieve analysis’ (Gilbert et al., 1998).Gilbert et al. (2008),Sun et al. (2009), and Sun and Gilbert (2012) developed sieve analysis methods using the competing risks failure time framework (Prentice et al., 1978), which attach a continuous ‘mark’ variable to HIV infected subjects that measures the genetic distance of an infecting HIV sequence to a sequence inside the vaccine. The goal of the sieve analysis methods is evaluation of mark-specific vaccine efficacy, here defined as one minus the mark-specific hazard ratio (vaccine/placebo) of infection. Beyond HIV, the methods apply generally to any preventative vaccine efficacy trial for which the pathogen targeted by the vaccine is genetically diverse, which include influenza, malaria, tuberculosis, dengue, streptococcus pneumoniae, human papilloma virus, and hepatitis C virus.

Gilbert et al. (2008) and Sun et al. (2009) assumed no missing mark data in infected subjects, whereas Sun and Gilbert (2012) allowed missing at random (MAR) marks. In practice there are missing marks, for example in the Vax004 trial 32 of 368 infected subjects had no HIV sequence data (Gilbert et al., 2008), due to drop-out or to inability of the HIV sequencing technology to measure the infecting HIV sequence, and in the ‘Step’ trial 22 of 88 infected subjects had no HIV sequence data (Rolland et al., 2011). While it is of scientific interest to evaluate amark defined based on the earliest available HIV sequence, a mark of particular scientific interest is defined based on an HIV sequence measured near the time of acquisition, which is missing in a much larger fraction of infected subjects due to the periodic (typically 6-monthly) diagnostic tests for HIV infection. Specifically, HIV sequences are measured from the earliest available post-infection blood sample, and a ‘near acquisition’ or ‘early’ sample may be defined as one documented to be sufficiently near acquisition. In the Step trial, only 23 of the 66 infected subjects with sequence data had an early mark measured, defined as sampling within 3 weeks. Sun and Gilbert (2012) provide details on the HIV testing algorithm that is used to define an early mark.

Sun and Gilbert (2012) is currently the only paper on sieve analysis that accommodates missing continuous marks. It develops two valid estimation approaches based on the stratified mark-specific proportional hazards model. The first uses inverse probability weighting (IPW) of the complete-case estimator, which leverages auxiliary predictors of whether the mark is observed, whereas the second, adapting Robins et al. (1994), augments the IPW complete-case estimator with auxiliary predictors of the missing marks. Sun and Gilbert (2012) restricted attention to estimation methods, and this article is a sequel that develops corresponding inferential/hypothesis testing methods based on the augmented IPW estimator. An important new component of this work compared to the previous work is to center it around the sieve analysis of the RV144 Thai trial, which recently delivered the landmark result that a prime-boost HIV vaccine appeared to provide partial protection against HIV infection (estimated VE = 31%, 95% CI 1% to 51%, Rerks-Ngarm et al., 2009). This result has stimulated intense interest in the sieve analysis, for two reasons. First, there is controversy about whether the vaccine is really partially working versus a false positive result (Gilbert et al., 2011), and the sieve analysis of HIV sequences can help resolve this question. In particular, if evidence is found that the vaccine efficacy declines with genetic distance, and the distance is defined based on known parts of HIV that contain putatively protective antibody epitopes, then an interpretation of real vaccine efficacy is supported. Secondly, the HIV vaccine field is grappling with how to modify the tested vaccine to increase its potential vaccine efficacy for the next efficacy trial, and understanding the relationship between vaccine efficacy and the genetic distance provides direct guidance on which HIV sequences to put inside of the next generation vaccines.

This article is organized as follows. Notations, assumptions, and the stratified mark-specific proportional hazards model are introduced in Section 2. Background on the estimation procedures needed for the testing procedures are described in Section 3. The testing procedures are developed, and asymptotic properties described, in Section 4. The finite-sample performances of the tests are evaluated via simulations in Section 5. The application to the Thai trial is given in Section 6, and the asymptotic results and their proofs are placed in the Appendix.

2 Model and missing mark data

2.1 Stratified mark-specific proportional hazards (PH) model

Let T be the failure time, V a continuous mark variable with bounded support [0, 1], and $\mathbf{Z}(t)$ a possibly time-dependent p-dimensional covariate. The mark V is only observable when T is observed. Suppose that the conditional mark-specific hazard function at time t given the covariate history $\mathbf{Z}(s)$, for $s \le t$, only depends on the current value $\mathbf{Z}(t)$. We consider the stratified mark-specific proportional hazards (PH) model

$$\lambda_k(t,v|\mathbf{z}(t))=\lambda_{0k}(t,v)\exp\left\{\boldsymbol{\beta}(v)^T\mathbf{z}(t)\right\},k=1,\ldots,K, \quad (1)$$

where $\lambda_k(t, v|\mathbf{z}(t))$ is the conditional mark-specific hazard function given covariate $\mathbf{z}(t)$ for an individual in the kth stratum, $\lambda_{0k}(\cdot, v) = \lambda_k(t, v|\mathbf{z}(t) = 0)$ is the unspecified baseline hazard function for the kth stratum, $\beta(v)$ is the p-dimensional unknown regression coefficient function of v, and K is the number of strata. Model (1) allows different baseline functions for different strata and flexibly allows for arbitrary mark-specific infection hazards over time in the placebo group. In practice, different key subgroups (e.g., men and women in the Thai trial) are assigned different baseline mark-specific hazards of HIV infection.

Arranging $\boldsymbol{\beta}(v)=(\beta_1(v),\boldsymbol{\beta}_2^T(v))^T$, so that $\beta_1(v)$ is the coefficient for vaccination status and $\beta_2(v)$ for other covariates, the covariate and stratum adjusted mark-specific vaccine efficacy VE(v) equals $1 - \exp(\beta_1(v))$.Sun et al. (2009) developed some statistical procedures for model (1) with $K = 1$ based on observations of the random variables $(X,\mathbf{Z}(\cdot), V)$ for $\delta = 1$ and $(X,\mathbf{Z}(\cdot))$ for $\delta = 0$, where $X = \min\{T,C\}$, $\lambda = I(T \le C)$, and C is a censoring random variable. Sun and Gilbert (2012) developed estimation procedures for model (1) with general K allowing V to be missing for some subjects with $\delta = 1$; these methods incorporate auxiliary covariates and/or auxiliary mark variables that inform about the probability V is observed and about the distribution of V. This article develops parallel hypothesis testing procedures for assessing VE(v). As summarized in the Introduction, the two objectives are to assess if the vaccine efficacy ever deviates from 0 [i.e., test VE(v) = 0] and to assess if the vaccine efficacy changes with the mark [i.e., test VE(v) = VE].

2.2 Missing data assumptions

Let R be the indicator of whether all possible data are observed for a subject; $R = 1$ if either $\delta = 0$ (right-censored) or if $\delta = 1$ and V is observed; and $R = 0$ otherwise. Auxiliary variables $\mathbf{A}$ may be helpful for predicting missing marks. Since the mark can only be missing for failures, supplemental information is potentially useful only for failures, for predicting missingness and for informing about the distribution of missing marks. For example, if V is defined based on the early virus, then $\mathbf{V}^*$, the auxiliary mark information, may include sequences of later sampled viruses, and can be considered a subset of $\mathbf{A}$. In general, $\mathbf{A}$ could include multiple viral sequences per infected subject at multiple time-points, giving information on intra-subject HIV evolution. The relationship between $\mathbf{A}$ and V can be modelled to help predict V (see Section 5 for a simulated example).

We assume C is conditionally independent of (T, V) given $\mathbf{Z}(\cdot)$ and the stratum. We also assume V is MAR (Rubin, 1976); that is, given $\delta = 1$ and $\mathbf{W} = (T, \mathbf{Z}(T), \mathbf{A})$, the probability V is missing depends only on the observed $\mathbf{W}$, not on the value of V; this assumption is expressed as

$$r_k(W) \equiv P(R{=}1\,|\delta{=}1, \mathbf{W}){=}P(R{=}1|\,V, \delta{=}1, \mathbf{W}). \quad (2)$$

Let $\pi_k(\mathbf{Q}) = P(R = 1|\mathbf{Q})$ where $\mathbf{Q} = (\delta, \mathbf{W})$. Then $\pi_k(\mathbf{Q}) = \delta r_k(\mathbf{W}) + (1 - \delta)$. The MAR assumption (2) also implies that V is independent of R given $\mathbf{Q}$:

$$\rho_k(v, \mathbf{W}) \equiv P(V \le v\,|\delta{=}1, \mathbf{W}){=}P(V \le v|\,R{=}1, \delta{=}1, \mathbf{W}). \quad (3)$$

Define $r_k(\mathbf{w}) = P(R = 1|\delta = 1, \mathbf{W} = \mathbf{w})$ and $\rho_k(v, \mathbf{w}) = P(V \le v|\lambda = 1, \mathbf{W} = \mathbf{w})$. The stratum-specific definitions of $r_k(\mathbf{w})$ and $\rho_k(v, \mathbf{w})$ allow the models of the probability of complete-case and of the mark distribution to differ across strata.

Let τ be the end of the follow-up period, and n_k be the number of subjects in the kth stratum; the total sample size is $n{=}\sum_{k=1}^{K} n_k$. Let $\{X_{ki}, \mathbf{Z}_{ki}(\cdot), \delta_{ki}, R_{ki}, V_{ki}, \mathbf{A}_{ki}\,; i = 1,\dots, n_k\}$ be iid replicates of $\{X, \mathbf{Z}(\cdot), \delta, R, V, A\}$ from the kth stratum. The observed data are $\{\mathbf{O}_{ki}\,; i = 1,\dots, n_k, k = 1,\dots, K\}$, where $\mathbf{O}_{ki} = \{X_{ki}, \mathbf{Z}_{ki}(\cdot), R_{ki}, R_{ki} V_{ki}, \mathbf{A}_{ki}\}$ for $\delta_{ki} = 1$ and $\mathbf{O}_{ki} = \{X_{ki}, \mathbf{Z}_{ki}(\cdot), R_{ki} = 1\}$ for $\delta_{ki} = 0$. We assume the $\mathbf{O}_{ki}$ are independent for all subjects.

2.3 Hypotheses to test

We develop procedures for testing the following two sets of hypotheses. Let $[a, b] \subset (0, 1)$. The first set of hypotheses is

H_{10} : VE(v) = 0 for $v \in [a, b]$

versus H_{1a} : VE(v) ≠ 0 for some v (general alternative)

or H_{1m} : VE(v) ≥ 0 with strict inequality for some v (monotone alternative).

The second set of hypotheses is

H_{20} : VE(v) does not depend on $v \in [a, b]$

versus H_{2a} : VE(v) depends on v (general alternative)

or H_{2m} : VE(v) decreases as v increases (monotone alternative).

The null hypothesis H_{10} implies the vaccine affords no protection (nor increased risk) against any HIV genotype. The ordered alternative H_{1m} indicates that the vaccine provides protection for at least some of the HIV genotypes, while H_{1a} indicates that the vaccine provides protection and/or increased risk for some HIV genotypes. The null hypothesis H_{20} implies there is no difference in vaccine protection against different HIV genotypes. The ordered alternative H_{2m} indicates that vaccine efficacy decreases with v and H_{2a} indicates that the vaccine efficacy changes with v. With β_1 (v) the first component of $\beta(v)$, the first set of hypotheses is equivalent to $H_{10} : \beta_1 (v) = 0$ for $v \in [a, b]$ versus $H_{1a} : \beta_1 (v) \neq 0$ for some v or $H_{1m} : \beta_1 (v) \leq 0$ with strict inequality for some v. The second set of hypotheses is equivalent to $H_{20} : \beta_1 (v)$ does not depend on $v \in [a, b]$ versus $H_{2a} : \beta_1 (v)$ depends on v or $H_{2m} : \beta_1 (v)$ increases as v increases. We develop testing procedures for detecting departures from H_{10} in the direction of H_{1a} and H_{1m} and for detecting departures from H_{20} in the direction of H_{2a} and H_{2m}. The procedures are developed based on the augmented IPW complete-case estimator developed by Sun and Gilbert (2012).

3 Estimation procedure with missing marks

The augmented IPW estimator for model (1) is obtained in two stages. First the IPW complete-case estimator is derived and second the augmented IPW estimator is obtained, which improves efficiency by accounting for information in the conditional distribution of V given the auxiliaries.

Let r_k ($\mathbf{W}_{ki}$, ψ_k) be the parametric model for the probability of complete-case, r_k ($\mathbf{W}_{ki}$) defined in (2), where $\mathbf{W}_{ki} = (T_{ki}, \mathbf{Z}_{ki} (T_{ki}), \mathbf{A}_{ki})$ and ψ_k is a q-dimensional parameter. For example, one can assume the logistic model with $\text{logit}(r_k(\mathbf{W}_{ki}, \psi_k)) = \psi_k^T \mathbf{W}_{ki} ==$ for those with $\lambda_{ki} = 1$, where $\mathbf{W}_{ki} = (T_{ki}, \mathbf{Z}_{ki} (T_{ki}), \mathbf{A}_{ki})$. By (2), the maximum likelihood estimator $\hat{\psi} = (\hat{\psi}_1, \ldots, \hat{\psi}_K)^T$ of $\psi = (\psi_1, \ldots, \psi_K)^T$ is obtained by maximizing the observed data likelihood,

$$\prod_{k,i} \{r_k(W_{ki}, \psi_k)\}^{R_{ki}\delta_{ki}} \{1 - r_k(W_{ki}, \psi_k)\}^{(1-R_{ki})\delta_{ki}}. \quad (4)$$

Let $K(x)$ be a kernel function with support [−1, 1] and let $h = h_n$ be a bandwidth. Let N_{ki} (t, v) = $I(X_{ki} \leq t, \delta_{ki} = 1, V_{ki} \leq v)$ and Y_{ki} (t) = $I(X_{ki} \geq t)$. Let $\mathbf{Q}_{ki} = (\delta_{ki}, \mathbf{W}_{ki})$ and π_k ($\mathbf{Q}_{ki}$, ψ_k) = δ_{ki} r_k ($\mathbf{W}_{ki}$, ψ_k) + $(1 - \delta_{ki})$. The first-stage estimator is the IPW estimator $\hat{\beta}^{ipw}$ (v), which solves the following estimating equation for β: U_{ipw} (v, β, $\hat{\psi}$) = 0, where

$$U_{ipw}(v, \boldsymbol{\beta}, \hat{\psi}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u-v)(\boldsymbol{Z}_{ki}(t) - \tilde{\boldsymbol{Z}}_k(t, \boldsymbol{\beta}, \hat{\psi}_k)) \frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki}, \hat{\psi}_k)} N_{ki}(dt, du), \quad (5)$$

where $K_h(x)=K(x/h)/h, \tilde{\boldsymbol{Z}}_k(t,\boldsymbol{\beta},\psi_k) = \tilde{\mathbf{S}}_k^{(1)}(t,\boldsymbol{\beta},\psi_k)/\tilde{\mathbf{S}}_k^{(0)}(t,\boldsymbol{\beta},\psi_k) \text{and} \tilde{\mathbf{S}}_k^{(j)}(t,\boldsymbol{\beta},\psi_k) = n_k^{-1}\sum_{i=1}^{n_k} R_{ki}(\pi_k(\mathbf{Q}_{ki},\psi_k))^{-1} Y_{ki}(t)\exp\left\{\beta^T \boldsymbol{Z}_{ki}(t)\right\} \boldsymbol{Z}_{ki}(t)^{\otimes j}$ for j = 0, 1, where $\mathbf{z}^{\otimes 0} = 1$ and $\mathbf{z}^{\otimes 1} = \mathbf{z}$ for any $\mathbf{z} \in \mathbb{R}^p$. The score function (5) can be viewed as an extension of the score function used for the cause-specific Cox model (Prentice et al., 1978) for a particular failure cause $J = j$, for which the counting process only counts events of type j. It borrows strength from observations having marks in the neighborhood of v. The kernel function is designed to give greater weight to observations with marks near v than those further away.

The baseline function λ_{0k} (t, v) can be estimated by $\hat{\lambda}_{0k}^{ipw}(t,v)$, obtained by smoothing the increments of the following estimator of the doubly cumulative baseline function $\Lambda_{0k}(t,v)=\int_0^t\int_0^v \lambda_{0k}(s,u)dsdu$:

$$\hat{\Lambda}_{0k}^{ipw}(t,v)=\sum_{i=1}^{n_k}\int_0^t\int_0^v \frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki},\hat{\psi}_k)} \frac{N_{ki}(ds,du)}{n_k\tilde{S}_k^{(0)}(s,\hat{\boldsymbol{\beta}}^{ipw}(u),\hat{\psi}_k)}. \quad (6)$$

For example, one can use the following kernel smoothing

$$\hat{\lambda}_{0k}^{ipw}(t,v)=\int_0^\tau\int_0^1 K_{h1}^{(1)}(t-s)K_{h2}^{(2)}(v-u)\hat{\Lambda}_{0k}^{ipw}(ds,du), \quad (7)$$

where $K_{h_1}^{(1)}(x)=K^{(1)}(x/h_1)/h_1 \text{and} K_{h_2}^{(2)}(x)=K^{(2)}(x/h_2)/h_2$, with $K^{(1)}$ $(\cdot)$ and $K^{(2)}$ $(\cdot)$ the kernel functions and h_1 and h_2 the bandwidths.

Following Robins et al. (1994), Sun and Gilbert (2012) proposed a more efficient procedure for estimating (1) by incorporating the knowledge of ρ_k $(\mathbf{w}, v)$ into the estimation procedure. Let $\mathbf{w} = (t, \mathbf{z}, \mathbf{a})$ and g_k $(\mathbf{a}|t, v, \mathbf{z}) = P(\mathbf{A}_{ki} = \mathbf{a}|T_{ki} = t, V_{ki} = v, \mathbf{Z}_{ki} = z, \delta_{ki} = 1)$. Then

$$\rho_k(\mathbf{w},v)=\int_0^v \lambda_k(t,u\left|\mathbf{z}\right.)g_k(\mathbf{a}\left|t,u,\mathbf{z}\right.)du/\int_0^1 \lambda_k(t,u|\mathbf{z})g_k(\mathbf{a}\left|\,t,u,\mathbf{z}\right.)du. \quad (8)$$

If no auxiliary variables are available or if $\mathbf{A}_{ki}$ is conditionally independent of V_{ki} given $(T_{ki}$, $\mathbf{Z}_{ki}$, $\delta_{ki})$, then $\rho_k(\mathbf{w},v)=\int_0^v \lambda_k(t,u\left|\mathbf{z}\right.)du/\int_0^1 \lambda_k(t,u\left|\,\mathbf{z}\right.)du$. In this case, $\rho_k(\mathbf{w}, v)$ can be estimated by

$\hat{\rho}_k^{ipw}(w,v)=\int_0^v \hat{\lambda}_k^{ipw}(t,u\left|\mathbf{z}\right.)du/\int_0^1 \hat{\lambda}_k^{ipw}(t,u|\mathbf{z})du, \text{where} \hat{\lambda}_k^{ipw}(t,u\left|\,\mathbf{z}\right.)=\hat{\lambda}_k^{ipw}(t,u)\exp\left\{(\hat{\beta}^{ipw}(u))^T\mathbf{z}\right\}$. When the auxiliary marks $\mathbf{A}_{ki}$ are correlated with V_{ki} conditional on T_{ki}, $\mathbf{Z}_{ki}$ and $\delta_{ki} = 1$, the conditional distribution $\rho_k(\mathbf{w}, v)$ involves the function g_k $(\mathbf{a}|t, u, \mathbf{z})$, for which a parametric or semiparametric model may be developed to describe the dependence between $\mathbf{A}_{ki}$ and V_{ki}. Let $\hat{g}_k$ $(\mathbf{a}|t, u, \mathbf{z})$ be an estimator of g_k $(\mathbf{a}|t, u, \mathbf{z})$ with a convergence rate of at least $(nh)^{-1/2}$. Then $\rho_k(\mathbf{w}, v)$ can be estimated by

$$\hat{\rho}_k^{ipw}(\mathbf{w}, v)=\int_0^v \hat{\lambda}_k^{ipw}(t, u \left| \mathbf{z})\hat{g}_k(\mathbf{a} \right| t, u, \mathbf{z})\hat{g}_k(\mathbf{a}|t, u, \mathbf{z})du / \int_0^1 \hat{\lambda}_k^{ipw}(t, u|\mathbf{z})\hat{g}_k(a \left| t, u, \mathbf{z})du. \right. \quad (9)$$

Let $N_{ki}^x(t)=I(X_{ki} \leq t, \delta_{ki}=1) \text{and} N_{ki}^v(v)=I(V_{ki} \leq v)$. The augmented IPW (AIPW) estimating equation for β is U_{aug} $(v, \beta, \hat{\psi}, \hat{\rho}(\cdot)) = 0$, where

$$U_{aug}(v, \boldsymbol{\beta}, \hat{\psi}, \hat{\rho}(\cdot))=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)(\boldsymbol{Z}_{ki}(t)-\overline{\boldsymbol{Z}}_k(t, \boldsymbol{\beta})) \left\{\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki}, \hat{\psi}_k)} N_{ki}(dt, du)+\left(1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki}, \hat{\psi}_k)}\right) N_{ki}^x(dt)d(\hat{\rho}_k^{ipw}(\mathbf{W}_{ki}, u))\right\}, \quad (10)$$

and $\overline{\boldsymbol{Z}}_k(t, \boldsymbol{\beta})=\mathbf{S}_k^{(1)}(t, \boldsymbol{\beta})/\mathbf{S}_k^{(0)}(t, \boldsymbol{\beta}), \mathbf{S}_k^{(j)}(t, \boldsymbol{\beta})=n_k^{-1}\sum_{i=1}^{n_k} Y_{ki}(t)\exp\left\{\boldsymbol{\beta}^T \boldsymbol{Z}_{ki}(t)\right\} \boldsymbol{Z}_{ki}(t)^{\otimes j}$ for $j = 0, 1$. The AIPW estimator of β(v) solves the above equation and is denoted by $\hat{\beta}^{aug}$ (v). The estimator of the cumulative function $\mathbf{B}(v)=\int_0^v \boldsymbol{\beta}(u)du \text{is given by} \hat{\mathbf{B}}^{aug}(v)=\int_0^v \hat{\boldsymbol{\beta}}^{aug}(u)du$. Note that there is no $\hat{\psi}_k$ in $\bar{\mathbf{Z}}_k$ (t, β); this is a difference between the IPW and AIPW estimators.

To implement the estimation procedures in practice, one can use arbitrary auxiliaries for estimating $\hat{\psi}_k$; these auxiliaries may include covariates and marks at multiple time-points pre-infection and post-infection, respectively. In contrast, while in principle arbitrary auxiliaries may also be used for the terms $\hat{g}_k$ ($\mathbf{a}|t, u, \mathbf{z}$) in (9), due to the curse of dimensionality the method is expected to perform best in practice with a univariate auxiliary, where semiparametric or fully parametric models for g_k ($\mathbf{a}|t, u, \mathbf{z}$) would be required to include multivariate auxiliaries.

Sun and Gilbert (2012) proved that the estimators $\hat{\beta}_{ipw}$ (t, v) and $\hat{\beta}^{aug}$ (t, v) are consistent and that $\hat{\beta}^{aug}$ (v) is more efficient than $\hat{\beta}^{ipw}$ (v). In the next section, we develop some hypothesis testing procedures for assessing mark-specific vaccine efficacy based on $\hat{\mathbf{B}}^{aug}$ (v).

4 Testing of mark-specific vaccine efficacy

The covariate-adjusted vaccine efficacy VE(v) is defined through the first component of β(v). Let B_1 (v) be the first component of the cumulative coefficient function **B**(v). The hypothesis tests concerning VE(v) are constructed based on the first component $\hat{B}_1^{aug}(v)$ of the AIPW estimator $\hat{\mathbf{B}}^{aug}$ (v). The cumulative estimator $\hat{\mathbf{B}}^{aug}$ (v) has more stable large-sample behavior and a faster convergence rate than β^{aug} (v).

Let $\mathbf{W}_B$ (v) = $n^{1/2}$ {$\hat{\mathbf{B}}^{aug}$ (v) − $\hat{\mathbf{B}}^{aug}$ (a)} − $n^{1/2}$ {B(v) − **B**(a)} for $v \in [a, b]$. In the Appendix we show that $\mathbf{W}_B$ (v), $v \in [a, b]$, converges weakly to a p-dimensional mean-zero Gaussian process with continuous sample paths on $v \in [a, b]$. Further, the distribution of $\mathbf{W}_B$ (v), for $v \in [a, b]$, can be approximated using the Gaussian multipliers resampling method [of Lin et al. (1993)] based on $\mathbf{W}_B^*(v)=n^{-1/2}\sum_{k=1}^K\sum_{i=1}^{n_k}\xi_{ki}\hat{\mathbf{H}}_k i(v)$ $v \in [a, b]$, where {ξ_{ki}, i = 1,…, n_k, k = 1,…,K} are iid standard normal random variables and $\hat{\mathbf{H}}_k$ i(v) is defined in (22) in the

Appendix. Let $\mathbf{W}_{B_1}(v)$ and $W^*_{B1}(v)$ be the first component of $\mathbf{W}_B(v)$ and $\mathbf{W}^*_B(v)$, respectively. With the Gaussian multipliers method, the variance $\mathrm{Var}\left\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\right\}$ can be consistently estimated by $\hat{\mathrm{Var}}\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\}=n^{-1}\mathrm{Var}*(W^*_{B_1}(v)),\mathrm{where}\ \ \mathrm{Var}*(W^*_{B_1}(v))$ is the first component on the diagonal of the covariance given in (23) in the Appendix.

4.1 Testing the null hypothesis H_{10}

Consider the test process $Q^{(1)}(v)=n^{1/2}\left\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\right\},v\in[a,b]$. Then $Q^{(1)}(v)=W_{B_1}(v)+n^{1/2}\{B_1(v)-B_1(a)\}$, $v\in[a,b]$. Under H_{10}, $B_1(v)-B_1(a)=0$ for $v\in[a,b]$, which motivates the following test statistics for testing H_{10}:

$$T^{(1)}_{a1}=\sup_{v\in[a,b]}\left|Q^{(1)}(v)\right|,T^{(1)}_{a2}=\int_a^b\left\{Q^{(1)}(v)\right\}^2d\mathrm{Var}*\left\{W^*_{B1}(v)\right\},\ T^{(1)}_{m1}=\inf_{v\in[a,b]}\left|Q^{(1)}(v)\right|,T^{(1)}_{m2}=\int_a^bQ^{(2)}(v)d\mathrm{Var}*\left\{W^*_{B1}(v)\right\},$$

The test statistics $T^{(1)}_{a1}\mathrm{and}T^{(1)}_{a2}$ capture general departures H_{1a}, while the test statistics $T^{(1)}_{m1}\mathrm{and}T^{(1)}_{m2}$ are sensitive to the monotone departures H_{1m}. It is easy to derive that all the test statistics $T^{(1)}_{a1},T^{(1)}_{a2},T^{(1)}_{m1}\mathrm{and}T^{(1)}_{m2}$ are consistent against their respective alternative hypotheses, and the Appendix derives their limiting distributions under H_{10}.

Under H_{10}, the distribution of $Q^{(1)}(v)$, $v\in[a,b]$, can be approximated by the conditional distribution of $W^*_{B_1}(\cdot)$, $v\in[a,b]$, given the observed data sequence. Hence, the distributions of $T^{(2)}_{a1},T^{(2)}_{a2},T^{(2)}_{m1}\mathrm{and}T^{(2)}_{m2}$ under H_{10} can be approximated by the conditional distributions of $T^{*(1)}_{a1}=\sup_{v\in[a,b]}\left|W^*_{B_1}(v)\right|,T^{*(1)}_{a2}=\int_a^b\left\{W^*_{B_1}(v)\right\}^2d\mathrm{Var}*\left\{W^*_{B_1}(v)\right\},T^{*(1)}_{m1}=\inf_{v\in[a,b]}W^*_{B_1}(v)\mathrm{and}T^{*(1)}_{m2}=\int_a^bW^*_{B_1}(v)d\mathrm{Var}*\left\{W^*_{B_1}(v)\right\}$, given the observed data sequence, respectively. The critical values, $c^{(1)}_{a1}\mathrm{and}c^{(1)}_{a2}$, of the test statistics $T^{(1)}_{a1}\mathrm{and}T^{(1)}_{a2}$ can be approximated by the (1 − α)-quantile of $T^{*(1)}_{a1}\mathrm{and}T^{*(1)}_{a2}$, which can be obtained by repeatedly generating a large number, say 500, of independent sets of normal samples $\{\lambda_{ki}, i=1,\ldots,n_k, k=1,\ldots,K\}$ while holding the observed data sequence fixed. Similarly, the critical values, $c^{(1)}_{m1}\mathrm{and}c^{(1)}_{m2}$, of the test statistics $T^{(1)}_{m1}\mathrm{and}T^{(1)}_{m2}$ can be approximated by the α-quantile of $T^{*(1)}_{m1}\mathrm{and}T^{*(1)}_{m2}$, which again can be obtained by repeatedly generating independent sets of normal samples $\{\xi_{ki}, i=1,\ldots,n_k, k=1,\ldots,K\}$. At significance level α, the tests based on $T^{(1)}_{a1}\mathrm{and}T^{(1)}_{a2}$ reject H_{10} in favor of H_{1a} if $T^{(1)}_{a1}>c^{(1)}_{a1}\mathrm{and}T^{(1)}_{a2}>c^{(1)}_{a2}$, respectively, and the tests based on $T^{(1)}_{m1}\mathrm{and}T^{(1)}_{m2}$ reject H_{10} in favor of H_{1m} if $T^{(2)}_{m1}<c^{(2)}_{m1}\mathrm{and}T^{(2)}_{m2}<c^{(2)}_{m2}$, respectively.

4.2 Testing the null hypothesis H_{20}

Let

$Q^{(2)}(v)=(v-a)^{-1}n^{1/2}\left\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\right\}-(b-a)^{-1}n^{1/2}\left\{\hat{B}_1^{aug}(b)-\hat{B}_1^{aug}(a)\right\}$.Then

$$Q^{(2)}(v)=\Gamma(v,W_{B_1})+n^{1/2}\Gamma(v,B_1)\text{for}a<v\leq b, \quad (11)$$

where $\Gamma(v, F_1) = (v - a)^{-1} \{F_1 (v) - F_1 (a)\} - (b - a)^{-1} \{F_1 (b) - F_1 (a)\}$ is a transformation of $F_1 (\cdot)$. We note that $\Gamma(\cdot, B_1) = 0$ under H_{20} and $\Gamma(\cdot, B_1) \neq 0$ under the alternatives, motivating $Q^{(2)} (v)$ as the test process and the following test statistics for testing H_{20}:

$$T_{a1}^{(2)}=\sup_{v\in[a',b]}\left|Q^{(2)}(v)\right|,T_{a2}^{(2)}=\int_{a'}^{b}\left\{Q^{(2)}(v)\right\}^2d\text{Var}*\left\{W_{B1}^{*}(v)\right\},$$
$$T_{m1}^{(2)}=\inf_{v\in[a',b]}\left|Q^{(2)}(v),T_{m2}^{(2)}=\int_{a'}^{b}Q^{(2)}(v)d\text{Var}*\left\{W_{B1}^{*}(v)\right\},\right.$$

where $a < a' < b$. We choose $a' > a$ to avoid zero in the denominator of $Q^{(2)} (v)$. In practice, one can choose a'close to a to make use of available data and to ensure the tests are consistent.

By the asymptotic results shown in the Appendix and the continuous mapping theorem, under H_{20} the distribution of $Q^{(2)} (v)$, $v \in [a, b]$, can be approximated by the conditional distribution of $\Gamma(v,W_{B_1}^{*})$, $v \in [a, b]$, given the observed data sequence. Hence, the distributions of $T_{a1}^{(2)},T_{a2}^{(2)},T_{m1}^{(2)}\text{and}T_{m2}^{(2)}$ under H_{20} can be approximated by the conditional distributions of

$$T_{a1}^{*(2)}=\sup_{v\in[a',b]}\Gamma(v,W_{B_1}^{*})\big|,T_{a2}^{*(2)}=\int_{a'}^{b}\left\{\Gamma(v,W_{B_1}^{*})\right\}^2d\text{Var}*\left\{W_{B_1}^{*}(v)\right\},T_{m1}^{*(2)}=\inf_{v\in[a',b]}\Gamma(v,W_{B_1}^{*}),T_{m2}^{*(2)}=\int_{a'}^{b}\Gamma(v,W_{B_1}^{*})\text{and}T_{m2}^{*(2)}=\int_{a'}^{b}\Gamma(v,W_{B_1}^{*})d\text{Var}*\left\{W_{B_1}^{*}(v)\right\}$$

, given the observed data sequence, respectively. Similar to Section 4.1, the respective critical values $c_{a1}^{(2)}\text{and}c_{a2}^{(2)}$ of the test statistics $T_{a1}^{(2)}\text{and}T_{a2}^{(2)}$ can be approximated by the (1 − α)-quantiles of the conditional distributions of $T_{a1}^{*(2)}\text{and}T_{a2}^{*(2)}$ obtained through repeatedly generating independent sets of normal samples $\{\xi_{ki}, i = 1,\ldots, n_k, k = 1,\ldots, K\}$ while holding the observed data sequence fixed. The critical values $c_{m1}^{(2)}\text{and}c_{m2}^{(2)}\text{for}T_{m1}^{(2)}\text{and}T_{m2}^{(2)}$ can be approximated similarly. At the significance level α, the tests based on $T_{a1}^{(2)}\text{and}T_{a2}^{(2)}$ reject H_{20} in favor of H_{2a} if $T_{a1}^{(2)}>c_{a1}^{(2)}\text{and}T_{a2}^{(2)}>c_{a2}^{(2)}$, respectively, and the tests based on $T_{m1}^{(2)}\text{and}T_{m2}^{(2)}$ reject H_{20} in favor of H_{2m} if $T_{m1}^{(2)}<c_{m1}^{(2)}\text{and},T_{m2}^{(2)}<c_{m2}^{(2)}$, respectively.

The tests $T_{a1}^{(2)}\text{and}T_{a2}^{(2)}$ capture general departures H_{2a} while the tests $T_{m1}^{(2)}\text{and}T_{m2}^{(2)}$ are sensitive to the monotone departure H_{2m}. Note that the derivative $d\Gamma(v, B_1)/dv = (v - a)^{-1} [\beta_1 (v) - (v - a)^{-1} B_1 (v)] \geq 0$ under H_{2m} with strict inequality for at least some $v \in [a, b]$. This plus the fact that $\Gamma(v,B_1)$ is non-decreasing with $\Gamma(b,B_1) = 0$ lead to the results that the tests based on $T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ are consistent against H_{2m} and the tests based on $T_{a1}^{(2)} \text{and} T_{a2}^{(2)}$ are consistent against H_{2a}. The proofs are given in the second paragraph following Theorem 1 in the Appendix.

In Sections 4.1 and 4.2, we considered two types of test statistics, namely the integration-based test statistics and the supremum-based test statistics, for each pair of hypotheses. The former are generalizations of the Cramér-von Mises test statistic, and involve integration of deviations over the whole range of the mark, whereas the latter are extensions of the classic Kolmogorov-Smirnov test statistic for testing goodness-of-fit of a distribution function, and take the supremum of such deviations. As demonstrated in a comprehensive analysis of the relative powers of the classic Kolmogorov-Smirnov test and the Cramér-von Mises test by Stephens (1974), we expect that the two types of test statistics have different powers for different true alternative distributions. The integration-based test statistics are best-suited for situations where the true alternative distribution deviates a little over the whole support of the mark and the supremum-based test statistics may have more power against situations where the true alternative has large deviations over a small section of the support. For example, for testing differential $VE(v)$, H_{20}, the supremum-based tests will tend to be relatively more powerful if $V\hat{E}(v)$ is very high for a small range of marks near a and declines sharply to zero and is constant at zero for all other marks.

5 Simulation study

5.1 Numerical assessment of the tests under correctly specified models

We conduct a simulation study to evaluate the finite-sample performance of the proposed testing procedures. The empirical sizes and powers of the test statistics are assessed for various models, sample sizes (500 and 800) and choices of bandwidths. The powers of the tests are evaluated in both situations where a correlated auxiliary variable is used and where it is absent.

We consider $K = 1$ stratum. Let Z_{ki} be the treatment indicator with $P(Z_{ki} = 1) = 0.5$. The (T_{ki}, V_{ki}) are generated from the following mark-specific proportional hazards model:

$$\lambda(t, v|z) = \exp\{\gamma v + (\alpha + \beta v) z\}, t \geq 0, 0 \leq v \leq 1, \quad (12)$$

where α, β and γ are constants. Under model (12), λ_0 $(t, v) = \exp(v)$ and $\text{VE}(v) = 1 - \exp(\alpha + \beta v)$. For α = 0 and β = 0, VE(v) = 0, indicating no vaccine efficacy, and for β = 0, VE(v) = VE, indicating mark-invariant vaccine efficacy; whereas β > 0 indicates VE(v) decreasing in v. We examine the hypothesis testing procedures for the following specific models:

- (M1) (α, β, γ) = (0, 0, 0.3), implying VE(v) = 0;
- (M2) (α, β, γ) = (−0.69, 0, 0.3), implying VE(v) does not depend on v;
- (M3) (α, β, γ) = (−0.6, 0.6, 0.3), implying VE(v) decreases;
- (M4) (α, β, γ) = (−1.2, 1.2, 0.3), implying VE(v) decreases;
- (M5) (α, β, γ) = (−1.5, 1.5, 0.3), implying VE(v) decreases.

We generate the censoring times from an exponential distribution, independent of (*T*, *V*), with censoring rates ranging from 20% to 30%. We take τ = 2.0. The complete-case indicator R_{ki} is generated with conditional probability $r_k(W_{ki}) = P(R_{ki} = 1|\delta_{ki} = 1, W_{ki})$, where

$$\text{logit}(r_k(W_{ki}))=\psi_k 0+\psi_{k1}Z_{ki}, i=1,\ldots,n_k, k=1,\ldots,K. \quad (13)$$

With $\psi_{k0} = 0.2$ and $\psi_{k1} = -0.2$ about 50% of observed failures are missing marks.

Conditional on (T_{ki}, Z_{ki}, V_{ki}), we assume that the auxiliary marks follow the model

$$A_{ki}=(0+1)^{-1}(V_{ki}+\theta U_{ki}), \theta>0, \quad (14)$$

for $i = 1,\ldots, n_k$, $k = 1,\ldots, K$, where V_{ki} are the possibly missing marks, U_{ki} is uniformly distributed on [0, 1] independent of V_{ki}, and θ > 0 is an association parameter between A_{ki} and V_{ki}. The correlation coefficient ρ between A_{ki} and V_{ki} is 1 for θ = 0. Since A_{ki} is observed for all observed failure times, the AIPW estimator in this case is the full data estimator. The A_{ki} and V_{ki} are independent for θ = ∞, yielding ρ = 0. In addition, the θ values of 0.8, 0.4 and 0.2 correspond to ρ = 0.78, 0.92 and 0.98.

Under model (14), the conditional density of A_{ki} given (T_{ki}, Z_{ki}, V_{ki}) is

$$g_k(a|t,v,z;\theta)=\frac{1+\theta}{\theta}I\left\{\frac{v}{1+\theta}\leq a\leq\frac{v+\theta}{1+\theta}\right\}, 0\leq a\leq 1, 0\leq v\leq 1. \quad (15)$$

The likelihood function for θ is

$$L(\theta)=\prod_{\delta_{ki}=1,R_{ki}=1}\left(\frac{1+\theta}{\theta}I\left\{\frac{V_{ki}}{1+\theta}\leq A_{ki}\leq\frac{V_{ki}+\theta}{1+\theta}\right\}\right)\text{for}\theta>0.$$

It is easy to show that the maximum likelihood estimator equals

$$\hat{\theta}\max_{\delta_{ki}=1,R_{ki}=1}\{V_{ki}/A_{ki},(1-V_{ki})/(1-A_{ki})\}-1.$$

The density estimator $g_k(a|t, v, z; \hat{\theta})$ is plugged into (9) to obtain $\hat{\rho}_k^{ipw}(w,v)$, which is used to construct the AIPW estimator of β in (10).

The performances of the proposed test procedures are evaluated through simulations for the models described in (12), (13) and (14) under the settings (M1)–(M5), where (M1) is a setting under the null hypothesis H_{10} and (M2) is a setting under the null hypothesis H_{20}. We consider the situations where no auxiliary information is provided and where the correlation between the auxiliary mark and the mark of interest is ρ = 0.92 [under model (14) with θ = 0.4]. Table 1 presents the empirical sizes and powers of the tests $T_{a1}^{(1)},T_{a2}^{(1)},T_{m1}^{(1)}\text{and}T_{m2}^{(1)}$ for testing H_{10} at the nominal level 0.05. Table 2 presents the empirical sizes and powers of the tests $T_{a1}^{(2)}, T_{a2}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for testing H_{20} at the nominal level 0.05. The results are presented for $n = 500$ with $h_1 = 0.1$ and $h = h_2 = 0.15$ and 0.2, and for $n = 800$ with $h_1 = 0.1$ and $h = h_2 = 0.1$ and 0.15. We take $a = 0$, $b = 1$ and $a' = 0.5$ for the tests. The Epanechnikov kernel $K(x) = .75(1 - x^2)I\{|x| \le 1\}$ is used throughout the numerical analysis.

Tables 1 and 2 show that all of the tests have satisfactory empirical sizes close to the nominal level 0.05. The powers of the tests increase with sample size and they are not overly sensitive to the selected bandwidths. The powers of the tests for testing H_{10} increase as the model moves in the direction $M1 \rightarrow M3 \rightarrow M4 \rightarrow M2$, representing increased departure from the null hypothesis H_{10}. The powers of the tests for testing H_{20} increase as the model moves in the direction $M2 \rightarrow M3 \rightarrow M4 \rightarrow M5$, representing increased departure from the null hypothesis H_{20}. The tests utilizing the auxiliary marks have higher power than those without using the auxiliary marks.

As with any nonparametric smoothing procedure, one needs to carefully select bandwidths. In practice, the appropriate bandwidth selection can be based on a $\mathscr{K}$-fold cross-validation method [e.g., Efron and Tibshirani (1993), Hoover et al. (1998), Cai et al. (2000) and Tian et al. (2005)].

The proposed testing procedures properly handles missing marks under MAR with asymptotically correct significance levels. However, if only the observations with complete information are used, i.e., the complete-case analysis, then the testing procedures are expected to often not provide correct type I error control. We conduct a simulation study to evaluate the observed sizes of the proposed tests using the complete cases under two different models for missing the indicator R_{ki} – model (13) and the following model:

$$\text{logit}(r_k(W_{ki})) = 0.8 - Z_{ki} - 0.3T_{ki}, i = 1, \ldots, n_k, k = 1, \ldots, K. \quad (16)$$

For $K = 1$ both models (13) and (16) yield about 50% missing marks among the observed failures. The sizes of $T_{a1}^{(1)}, T_{a2}^{(1)}, T_{m1}^{(1)} \text{and} T_{m2}^{(1)}$ for testing H_{10} are evaluated under model (M1) and the sizes of $T_{a1}^{(2)}, T_{a2}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for testing H_{20} are evaluated under model (M2) (Table 3). Under model (13), the observed sizes for testing H_{10} are elevated (around 7–15%), whereas those for testing H_{20} remain around 5%. Under model (16), the observed sizes for testing H_{10} exceed 37% for all tests, whereas those for testing H_{20} reach 12% and 14% for the tests $T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ when $n = 800$.

These simulation results verify that the testing procedures applied to complete cases generally do not have nominal size, although for some of the scenarios the sizes are nominal. To explain this, it can be shown that, under MAR, $\lambda_k(t, v|z, R_{ki} = 1) = \lambda_k(t, v|z)h_k(t, z)$, where $h_k(t, z) = P(R_{ki} = 1|T_{ki} = t, Z_{ki} = z)/P(R_{ki} = 1|T_{ki} \ge t, Z_{ki} = z)$. If $h_k(t, z)$ does not depend on z and MAR holds, then the observations for individuals with the observed marks only can be viewed as a random sample from a mark-specific proportional hazards model with a different baseline hazard function but the same regression function $\beta(v)$. In this case, the tests for both H_{10} and H_{20} based on the complete cases are valid. If h_k (t, z) depends on z but not on t and MAR holds, then h_k (t, z) can be expressed as $h_k(t,z)=\exp(\vartheta_k' z)$ [the scenario under model (13)], and the tests of H_{10} based on the complete cases will be biased. However, the tests of H_{20} remain unbiased since the biases in the estimation of β(*v*) that do not depend on *v*, such that the test process $Q^{(2)}$ (v) is still asymptotically a mean zero process. In general, if h_k (t, z) depends on both z and t and MAR holds, which is the scenario under the missing model (16), then the test process $Q^{(2)}$ (v) is not an asymptotically mean zero process. The magnitude of departure of the asymptotic sizes of the test statistics of H_{20} from the nominal level depends on h_k (t, z) in a complicated manner.

5.2 Numerical assessment of the tests under mis-specified models

This subsection evaluates robustness of the proposed test procedures to mis-specifications of r_k (**w**) and/or g_k $(a|t, v, z)$, and to violation of the MAR assumption. The Z_{ki}, (T_{ki}, V_{ki}), and C_{ki} are generated using the same models as above, again with approximately 30% censoring.

Robustness of the tests to mis-specification of r_k (**w**) is examined by assuming model (13) while the actual complete-case indicator R_{ki} is generated with the conditional probability r_k $(\mathbf{W}_{ki}) = P(R_{ki} = 1|\delta_{ki} = 1, \mathbf{W}_{ki})$, where

$$\text{logit}(r_k(\mathbf{W}_{ki}))=1.1+Z_{ki}-2T_{ki}, i=1,\ldots,n_k. \quad (17)$$

This model yields approximately 50% missing marks among observed failures under (M1)–(M5).

Robustness of the tests is also examined when g_k $(a|t, v, z)$ is mis-specified. This is carried out by assuming model (14) for the auxiliary mark, or, equivalently, model (15) for g_k $(a|t, v, z)$, while the actual mark for $\lambda_{ki} = 1$ is generated from

$$A_{ki}=(1.4+2\tau)^{-1}(V_{ki}+0.4U_{ki}+2X_{ki}), \quad (18)$$

for $i = 1,\ldots, n_k$. Here U_{ki} is uniformly distributed on [0, 1] and is independent of V_{ki}.

Robustness of the tests to violation of the MAR assumption (2) is examined by assuming model (13), while the actual R_{ki} depends on V_{ki} through the model

$$\text{logit}(r_k(W_{ki}))=0.6+Z_{ki}-2V_{ki}, i=1,\ldots,n_k. \quad (19)$$

The proportion of missing marks among the observed failures is kept around 50% in all scenarios.

The models (17), (18) and (19) are similar to those used in Sun and Gilbert (2012) for examining robustness of the AIPW estimator. However, instead of examining biases and standard errors of the estimators, here we check whether the empirical sizes of the tests are close to their nominal level 0.05 and how the powers of the tests are affected by these mis-specifications. For sample size $n = 500$ and bandwidths $h_1 = 0.1$ and $h = h_2 = 0.20$, Table 4 shows the empirical sizes and powers of the tests of H_{10} and Table 5 shows the empirical sizes and powers of the tests of H_{20}. In both tables, the first block shows the results when r_k (w) is mis-specified following (17) and g_k $(a|t, v, z)$ is correctly specified by (15) with $\lambda = 0.4$; the second block shows the results when g_k $(a|t, v, z)$ is mis-specified following (18) and r_k (w) is correctly specified by (13) with $\psi_{k1} = 0.2$ and $\psi_{k1} = -0.2$; the third block shows the results when r_k (w) is mis-specified following (17) and g_k $(a|t, v, z)$ is mis-specified following (18); and the fourth block shows the results when r_k $(\mathbf{w})$ depends on V_{ki} following (19) and g_k $(a|t, v, z)$ is correctly specified by (15) with $\lambda = 0.4$.

Tables 4 and 5 show that the empirical sizes of the tests are very close to the nominal level 0.05 when one of r_k (w) and g_k $(a|t, v, z)$ is mis-specified, reflecting the double robustness property of the AIPW estimator. The empirical sizes are also close to 0.05 when both r_k (w) and g_k $(a|t, v, z)$ are mis-specified and when the MAR assumption is violated, which is intriguing. When only r_k (w) is mis-specified and MAR holds, the empirical powers in Tables 4 and 5 closely track the corresponding powers in Tables 1 and 2 under correct model specifications. The empirical powers are lower than those observed in Table 1 and 2 when g_k $(a|t, v, z)$ is mis-specified or when both r_k (w) and g_k $(a|t, v, z)$ are mis-specified, whereas the empirical powers in Tables 4 and 5 are very close to those in Tables 1 and 2 when MAR is violated. Apparently for our particular data simulation, the bias due to the MAR violation counter-balances the bias due to mis-specification of both r_k (w) and g_k $(a|t, v, z)$; however, in general these violations could distort sizes and powers.

5.3 Simulation study for the Thai trial

We conduct a simulation of the Thai trial, to gain insight about the power available for this real trial. Specifically, we simulated data to yield about the numbers of infections observed (74 in the placebo group and 51 in the vaccine group), the overall vaccine efficacy from the proportional hazards model is about 31%, and the true VE(v) curve decreases with v to be around 65–70% for v close to zero and around 0% for v close to 1. The actual infection rate was only 0.3% over 3.5 years; to speed the simulations we use a 20% placebo infection rate and retain 74 infections on average.

Again with $K = 1$ stratum, the (T_{ki}, V_{ki}) are generated from the following model:

$$\lambda(t, v|z)=\gamma\exp\{(\alpha+\beta v)z\}, t \geq 0, 0 \leq v \leq 1, \quad (20)$$

where α, β and γ are constants. Under model (20), VE(v) = $1 - \exp(\lambda + \beta v)$, the marginal hazards are $\lambda_0(t) = \sigma$ for $z = 0$, and $\lambda_1(t) = \gamma \exp(\alpha)(\exp(\beta)-1)/\beta$ for $z = 1$, and the Cox proportional hazards vaccine efficacy equals $VE_C = 1 - \lambda_1(t)/\lambda_0(t) = 1 - \exp(\alpha)(\exp(\beta) - 1)/\beta$. We choose $(\alpha, \beta, \gamma) = (-1.1, 1.3, 0.068)$, yielding $VE_C = 0.32$, VE(0) = 0.67, and VE(0.85) = 0. We study 400 subjects each in the vaccine and placebo groups. Matching the actual trial, the censoring rate before τ is kept very low, just under 5%. The missing mark indicator is generated from model (13), with (ψ_{k0}, ψ_{k1}) set to yield about 0%, 25% $(-1.2, -0.2)$, 50% $(0.2, -0.2)$, and 75% $(-1.0, -0.2)$ missing marks among observed failures. We assume the auxiliary variable A_{ki} follows the model (14) given in Section 5.1, where the θ values of ∞, 0.8, 0.4 and 0.2 correspond to $\lambda = 0$, 0.78, 0.92 and 0.98 for the correlation coefficient between A_{ki} and V_{ki}.

Because of lost information on the mark, we choose larger bandwidths for higher percentages of missing marks. We use $h = 0.4$ for the case with 75% missing marks; $h = 0.3$ for the case with 50% missing marks; $h = 0.2$ for the case with 25% missing marks; and $h = 0.15$ for the case with 0% missing marks. The bandwidths h_1 and h_2 in (7) in the estimation of $\hat{\lambda}_{0k}^{ipw}(t, v)$ are taken to be 0.50 and $h_2 = h$ in each case. Power of the proposed tests $T_{a1}^{(1)}, T_{a2}^{(1)}, T_{m1}^{(1)}, T_{m2}^{(1)}, T_{a1}^{(2)}, T_{a1}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for the simulations based on the Thai trial at the nominal level 0.05 are reported in Table 6. The tests show similar performance as was found in the simulation study of Section 5.1. As only 10% of infected subjects had missing marks in RV144 and the auxiliary was very weakly predictive, we focus on the entries with 0% or 25% missing marks and $\rho = 0$. There is 67%–95% power to reject H_{10}, and 33%–60% power to reject H_{20}. These results show that a fairly strong sieve effect with $VE(v)$ declining from 67% to 0% could readily be missed in the Thai trial due to limited power. The only slightly improved power with an excellent auxiliary $\rho = 0.98$ shows that greater numbers of events would be needed to achieve high power for testing H_{20}.

6 Analysis of the RV144 Thai trial

In the RV144 Thai trial, 125 subjects (51 of 8197 in the vaccine group and 74 of 8198 in the placebo group) were diagnosed with HIV infection over a 42 month follow-up period, from whom full-length HIV genomes were measured from 121; 3 missed data because their HIV viral load was too low for the Sanger sequencing technology to work, and 1 dropped out [Rerks-Ngarm et al. (2009), Rolland et al. (2012)]. We focus on the gp120 region of the HIV Env protein, because this region stimulates anti-HIV antibody responses which are the putative cause of the observed partial vaccine efficacy. Three gp120 sequences were included in the vaccine: 92TH023 in the ALVAC canarypox vector prime component; and CM244, MN in the AIDSVAX gp120 protein boost component. 92TH023 and CM244 are subtype E HIVs where as MN is subtype B, and 110 of the 121 subjects were infected with subtype E sequences. The subtype E vaccine-insert sequences are much closer genetically to the infecting (and regional circulating) sequences than MN, and thus are more likely to stimulate protective immune responses. Accordingly, the analysis focuses on the 92TH023 and CM244 reference sequences, and right-censors the 15 subjects HIV infected with subtype B or with unknown subtype. One subject who acquired HIV infection during the trial was documented to have acquired HIV from another trial participant who had previously become HIV infected; the analysis excludes this subject because his/her inclusion would violate the independent observations assumption. In the context of our model set-up, T is the time to HIV infection diagnosis with subtype E HIV. The time to HIV infection diagnosis with subtype B or with unknown HIV subtype is treated as censoring.

We define V based on HIV sequence data measured from a blood sample drawn at or before the HIV diagnosis date. (The trial documented acute-phase/pre-seroconversion infection in only a few subjects, prohibiting defining the mark based on acute-phase sequences.) Eleven of the 109 (11%) subtype E infected subjects have sequences measured from a post-diagnosis sample and hence are missing V. To maximize biological relevance and statistical power, we restrict the gp120 distances to the published set of gp120 sites in contact with known broadly neutralizing monoclonal antibodies (Moore et al., 2009; Wei et al., 2003).

For each HIV sequence from a subject and each of the two reference vaccine sequences, V is computed as a weighted Hamming distance using the PAM-between scoring matrix (Nickle et al., 2007). Between 2 and 13 sequences (total 1030) sequences) were measured per infected subject, and V is defined as the subject's sequence closest to his or her consensus sequence (the consensus sequence is comprised of the majority amino acids at each site, one site at a time). Finally, the distances are re-scaled to values between 0 and 1. In total, 109 infected subjects (43 vaccine, 66 placebo) are included in the analysis, of which 98 (39 vaccine, 59 placebo) have an observed mark V ; Figure 1 displays the observed V's.

To predict the probability of observing V among the 109 infected subjects, we use all-subsets logistic regression model selection considering demographics, host genetics, and biomarker data post-infection. The best model by BIC includes only the years from entry until HIV infection diagnosis (X_1), with model fit $\mathrm{logit}(\hat{P}(R = 1|\delta = 1, X_1)) = 1.17 + 0.70X_1$ for the CM244 reference sequence. The model was very similar for the 92TH023 reference sequence (not shown). In addition, we consider linear and logistic regression models for relating the mean of various potential auxiliary variables (A) to V, X_1, and treatment indicator Z. Model selection did not reveal any significantly predictive auxiliary variables; we expect that HIV sequence information measured after V is defined would be a good predictor, but these data were not collected. Nevertheless, to implement the AIPW method we select the best available auxiliary variable, gender ($A = X_2$, 1=male; 0=female), and use the logistic regression model that results; for CM244 the fitted model $\hat{g}(A = a|V, X_1, Z)$ is $\mathrm{logit}(\hat{P}(X_2 = 1|\delta = 1, V, X_1, Z) = 0.24 - 0.33V + 0.16X_1 + 0.38Z$, and the model was very similar for 92TH023 (not shown).

The AIPW estimation and testing procedures are applied to the Thai trial data set with bandwidths $h_1 = 0.5$ and $h_2 = h = 0.3$, $a = 0.05$, $b = 1$ and $a' = a + 0.01$ (a and a' are near the minimum observed marks). As in the simulation study, 500 simulated Gaussian multipliers are used. Because the results are nearly identical with and without the auxiliary variable, only the latter results are presented. Figure 2 shows the estimated VE(v) along with 95% pointwise confidence bands, indicating that vaccine efficacy appears to be high against HIVs near to the 92TH023 reference sequence [estimated VE(0.01) = 56%], and declines to zero against HIVs farthest from the 92TH023 reference sequence [estimated VE(1.0) = 2.4%]. The decline is similar for the CM244 reference sequence, with estimated VE(0.01) = 45% and estimated VE(0.95) = −9.1%.

Figure 3 (a) and (b) shows the test processes $Q^{(1)}(v)$ versus 20 realizations from the Gaussian multiplier process $W^{*}_{B_1}(v)$ given the observed data, and Figure 3 (c) and (d) shows the parallel results for the test process $Q^{(2)}(v)$, each suggesting departures from the null hypothesis H_{10} and from the null hypothesis H_{20} for each reference sequence. The p-values of the tests based on the test statistics $T^{(1)}_{m1} \mathrm{and} T^{(1)}_{m2}$ for testing H_{10} against the monotone alternative over $v \in [0, 1]$ are 0.032 and 0.008 for 92TH023, and 0.014 and 0.010 for CM244. The p-values of the test statistics $T^{(1)}_{a1} \mathrm{and} T^{(1)}_{a2}$ for testing H_{10} against the general alternative are 0.054 and 0.018 for 92TH023 and 0.030 and 0.010 for CM244. For testing H_{20} over $v \in [0, 1]$, the p-values of the supremum-type tests based on the test statistics

$T_{a1}^{(2)}\,\mathrm{and}\,T_{m1}^{(2)}$ are 0.53 and 0.27 for 92TH023 and 0.37 and 0.18 for CM244. The *p*-values of the integrated square type tests based on the test statistics $T_{a2}^{(2)}\,\mathrm{and}\,T_{m2}^{(2)}$ are 0.35 and 0.14 for 92TH023 and 0.44 and 0.19 for CM244.

These analyses provide more evidence that the vaccine had some protective efficacy than the original primary analysis that did not account for the mark information (Rerks-Ngarm et al., 2009): the primary analysis test for any vaccine efficacy yielded p=0.04 whereas the tests for any vaccine efficacy against any mark reported here yielded median p-value of 0.016 across the four test statistics and two reference sequences. The analyses also showed a nonsignificant trend (p-values around 0.14–0.19) that the vaccine protected better against HIVs closely matched to the vaccine strain HIVs in the monoclonal antibody contact sites, but had less or absent protection against HIVs with many mismatches in these sites. While the significance levels are not compelling, the simulation study presented in Section 5.3 of the power available for detecting a vaccine sieve effect in the Thai trial showed that the study is well-powered only to detect large sieve effects [with greater decline of $VE(v)$ in v than what was observed in the estimated $VE(v)$ curves]; thus a moderate-to-large sieve effect is consistent with the observed results. These results may guide future vaccine research by suggesting modifications of future vaccine candidates to include HIV sequences more closely matched to circulating HIVs in the monoclonal antibody contact sites. They may also motivate the design of future experiments to understand functional effects of amino acid mutations at the monoclonal antibody contact sites.

Acknowledgements

The authors thank Hasan Ahmed and Paul Edlefsen for generating the HIV sequence distances, and thank the participants, investigators, and sponsors of the RV144 Thai trial, including the U.S. Military HIV Research Program (MHRP); U.S. Army Medical Research and Materiel Command; NIAID; U.S. and Thai Components, Armed Forces Research Institute of Medical Science Ministry of Public Health, Thailand; Mahidol University; SanofiPasteur; and Global Solutions for Infectious Diseases. The authors thank the Editor, Associate Editor, and two referees for their helpful suggestions. The research of Yanqing Sun was partially supported by NSF grants DMS-0905777 and DMS-1208978, and the research of Drs. Sun and Peter Gilbert was partially supported by NIH NIAID grant R37AI054165. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix: Asymptotic results

The following regularity conditions from Sun and Gilbert (2012) are assumed.

Condition A

(A.1) $\beta(v)$ has component wise continuous second derivatives on [0, 1]. For each $k = 1, \ldots, K$,the second partial derivative of $\lambda_{0k}\,(t, v)$ with respect to v exists and is continuous on $[0, \tau]\times[0, 1]$. The covariate process $\mathbf{Z}_k\,(t)$ has paths that are left continuous and of bounded variation, and satisfies the moment condition $E[\|\mathbf{Z}_k\,(t)\|^4 \exp(2M\|\mathbf{Z}_k\,(t)\|)] < \infty$, where M is a constant such that $(v, \beta(v)) \in [0, 1] \times (-M, M)^p$ for all v and $\|A\| = \max_{k,l} |a_{kl}|$ for a matrix $A = (a_{kl})$.

(A.2) Each component of $\mathbf{s}_k^{(j)}(t,\theta)$ is continuous on [0, τ] × [−*M*, *M*]^p, $\tilde{\mathbf{s}}_k^{(j)}(t,\theta,\psi)$ is continuous on [0, τ] × [−*M*, *M*]^p × [−*L*, *L*]^q for some *M*, *L* > 0 and *j* = 0, 1, 2.

$\sup_{t\in[0,\tau],\theta\in[-M,M]^p}\|\mathbf{S}_k^{(j)}(t,\theta)\|=O_p(n^{-1/2})\text{and}\sup_{t\in[0,\tau],\theta\in[-M,M]^p,\psi_k\in}\|\tilde{\mathbf{S}}_k^{(j)}(t,\theta,\psi_k)-\tilde{\mathbf{s}}_k^{(j)}(t,\theta,\psi)\|=O_p$

(A.3) The limit $p_k = \lim_{n\to\infty} n_k/n$ exists and $0 < p_k < \infty$. $\mathbf{s}_k^{(0)}(t,\theta)>0$ on [0, τ] × [−*M*, *M*]^p and the matrix $\Sigma(v)=\sum_{k=1}^{K}p_k\Sigma_k(v)$ is positive definite, where

$\Sigma_k(v)=\sum_{k=1}^{K}\int_0^\tau I_k(t,\beta(v))\lambda_{0k}(t,v)\mathbf{s}_k^{(0)}(t,\boldsymbol{\beta}(v))dt\text{and}I_k(t,\boldsymbol{\beta})=\mathbf{s}_k^{(2)}(t,\boldsymbol{\beta})/\mathbf{s}_k^{(0)}(t,\boldsymbol{\beta})-(\overline{\mathbf{z}}_k(t,\beta))^{\otimes 2}$
.

(A.4) The kernel function *K*(·) is symmetric with support [−1, 1] and of bounded variation. The bandwidth *h* satisfies $nh^2 \to \infty$ and $nh^4 \to 0$ as $n \to \infty$.

(A.5) There is a σ > 0 such that $r_k(\mathbf{W}_{ki}) \geq \sigma$ for all *k*, *i* with $\delta_{ki} = 1$.

Let $\mathscr{F}_t = \sigma\{I(X_{ki} \leq s, \delta_{ki} = 1), I(X_{ki} \leq s, \delta_{ki} = 0), V_{ki}\, I(X_{ki} \leq s, \delta_{ki} = 1), \mathbf{Z}_{ki}(s); 0 \leq s \leq t, i = 1, \ldots, n_k, k = 1,\ldots, K\}$ be the (right-continuous) filtration generated by the full data processes $\{N_{ki}(s, v), Y_{ki}(s), \mathbf{Z}_{ki}(s); 0 \leq s \leq t, 0 \leq v \leq 1, i = 1,\ldots, n_k, k = 1,\ldots, K\}$. Assume $E(N_{ki}(dt, dv)|\mathscr{F}_{t-}) = E(N_{ki}(dt, dv)|Y_{ki}(t), \mathbf{Z}_{ki}(t))$, that is, the mark-specific instantaneous failure rate at time *t* given the observed information up to time *t* only depends on the failure status and the current covariate value. By the definition of the conditional mark-specific hazard function, $E(N_{ki}(dt, dv)|\mathscr{F}_{t-}) = Y_{ki}(t)\lambda_k(t, v|\mathbf{Z}_{ki}(t))\, dtdv$. Hence, the mark-specific intensity of $N_{ki}(t, v)$ with respect to $\mathscr{F}_t$ equals $Y_{ki}(t)\lambda_{ki}(t, v|\mathbf{Z}_{ki}(t))$. Let $M_{ki}(t,u)=\int_0^t\int_0^u[N_{ki}(ds,dx)-Y_{ki}(s)\lambda_k(s,x|\boldsymbol{Z}_{ki}(s))dsdx]$. By Aalen and Johansen (1978), $M_{ki}(\cdot, v_1)$ and $M_{ki}(\cdot, v_2) - M_{ki}(\cdot, v_1)$ are orthogonal square integrable martingales with respect to $\mathscr{F}_t$ for any $0 \leq v_1 \leq v_2 \leq 1$.

The weak convergence of $\mathbf{W}_B(v) = n^{1/2}\{\mathbf{B}^{\hat{aug}}(v) - \mathbf{B}(v)\} - n^{1/2}\{\mathbf{B}^{\hat{aug}}(a) - \mathbf{B}(a)\}$ for $v \in [a, b]$ is given in Theorem 1 below.

Theorem 1. *Under conditions (A.1)–(A.5),* $\mathbf{W}_B(v)=n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbf{H}_ki(v)+o_p(1)$, *uniformly in* $v \in [a, b]$, *where*

$$\mathbf{H}_ki(v)=\int_a^v\int_0^\tau\{\Sigma(u)\}^{-1}[\boldsymbol{Z}_{ki}(t)-\overline{\mathbf{z}}_k\{t,\boldsymbol{\beta}(u)\}]\left[\frac{R_{ki}}{\pi_k(Q_{ki})}M_{ki}(dt,du)+\left\{1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}\right\}E\{M_{ki}\}(dt,du)|\mathbf{Q}_{ki}\right\}. \quad (21)$$

The processes $\mathbf{W}_B(v)$ *converges weakly to a p-dimensional mean-zero Gaussian process with continuous sample paths on* $v \in [a, b]$, *where*

$\overline{\mathbf{z}}_k(t,\boldsymbol{\beta})=\mathbf{s}_k^{(1)}(t,\boldsymbol{\beta})/\mathbf{s}_k^{(0)}(t,\boldsymbol{\beta})\text{and}\mathbf{s}_k^{(j)}(t,\boldsymbol{\beta})=E\mathbf{S}_k^{(j)}(t,\boldsymbol{\beta})$.

Theorem 1 provides the basis for obtaining asymptotically correct critical values for the testing procedures for H_{10} and for H_{20}. In particular, let $G(v)$ be the limiting Gaussian process of$W_{B_1}(v)$, $v \in [a, b]$, as $n \to \infty$. Then under H_{10}, $Q^{(1)}(v)\xrightarrow{\mathscr{D}}G(v)$, $v \in [a, b]$, as $n \to \infty$. By Theorem 1 and the continuous mapping theorem,

$T_{a1}^{(1)} \xrightarrow{\mathscr{D}} \sup_{v\in p[a,b]} |G(v)|, T_{a2}^{(1)} \xrightarrow{\mathscr{D}} \int_a^b \{G(v)\}^2 dVar\{G(v)\}, T_{m1}^{(1)} \xrightarrow{\mathscr{D}} \inf_{v\in[a,b]} G(v) \text{and} T_{m2}^{(1)} \xrightarrow{\mathscr{D}} \int_a^b G(v) d\text{Var}\{G($

under H_{10} as $n \to \infty$. Under H_{20}, $Q^{(2)}(v) = \Gamma(v, W_{B_1}) \xrightarrow{\mathscr{D}} \Gamma(v, G)$, $v \in [a, b]$, as $n \to \infty$. Applying the continuous mapping theorem, under H_{20},

$T_{a1}^{(2)} \xrightarrow{\mathscr{D}} \sup_{v\in[a',b]} |\Gamma(v,G)|, T_{a2}^{(2)} \xrightarrow{\mathscr{D}} \int_{a'}^b \{\Gamma(v,G)\}^2 d\text{Var}\{G(v)\}, T_{m1}^{(2)} \xrightarrow{\mathscr{D}} \inf_{v\in[a',b]} \Gamma(v,G) \text{and} T_{m2}^{(2)} \xrightarrow{\mathscr{D}} \int_{a'}^b \Gamma(v, G$

, as $n \to \infty$.

The proof of the consistency of the tests for testing H_{10} are straightforward. To show the consistency of the tests for testing H_{20}, we note that the derivative $d\Gamma(v, B_1)/dv = (v - a)^{-1}$ $[\beta_1(v) - (v - a)^{-1} B_1(v)] \geq 0$ under H_{2m} with strict inequality for at least some $v \in [a, b]$. The function $\Gamma(v, B_1)$ is non-decreasing with $\Gamma(b, B_1) = 0$. We have, under H_{2m}, $\Gamma(v, B_1) \leq 0$ with strict inequality for at least some $v \in [a, b]$. Let $v_0 \in [a, b]$ be such that $\Gamma(v_0, B_1) < 0$. Then $\Gamma(v, B_1) < 0$ for $v \leq v_0$. Now defining $v_m^* = \sup\{v : \Gamma(v, B_1) < 0, a \leq v \leq b\}$, we have $\Gamma(v, B_1) < 0$ for $v < v_m^*$ and $\Gamma(v, B_1) = 0$ for $v^* \leq v < b$. It follows from (11) and Theorem 1 that $T_{m1}^{(2)} \xrightarrow{P} -\infty \text{and} T_{m2}^{(2)} \xrightarrow{P} -\infty$ under H_{2m} as $n \to \infty$ for $a' < v_m^*$. Thus the tests based on $T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ are consistent against H_{2m}. Similarly, let $v_a^* = \sup\left\{v : \sup_{u\in[v,b]} |\Gamma(u, B_1)| > 0, a \leq v \leq b\right\}$. Then under H_{2a}, $|\Gamma(v, B_1)| > 0$ for $v < v_a^*$, and $|\Gamma(v, B_1)| = 0$ for $v_a^* \leq v \leq b$. Hence $T_{a1}^{(2)} \xrightarrow{P} \infty \text{and} T_{a2}^{(2)} \xrightarrow{P} \infty$ under H_{2a} as $n \to \infty$ for $a' < v_a^*$, resulting in the consistent tests against H_{2a}.

We use the Gaussian multiplier resampling method [Lin et al. (1993)] to approximate the distribution of $\mathbf{W}_B(v)$, $v \in [a, b]$. Let $\{\xi_{ki}, i = 1,\ldots, n_k, k = 1,\ldots, K\}$ be iid standard normal random variables. Replacing each term of (26), which is asymptotically equivalent to (21), by its empirical counterpart and multiplying by ξ_{ki}, we obtain

$\mathbf{W}_B^*(v) = n^{-1/2} \sum_{k=1}^K \sum_{i=1}^{n_k} \xi_{ki} \hat{\mathbf{H}}_k i(v)$, where

$$\begin{aligned}\hat{\mathbf{H}}_k i(v) = \int_0^1 \int_0^\tau H(v,u) \left\{\boldsymbol{Z}_{ki}(t) - \overline{\boldsymbol{Z}}_k(t, \hat{\boldsymbol{\beta}}^{aug}(u))\right\} \\ \left\{\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki}, \hat{\psi}_k)} N_{ki}(dt, du) + (1 - \frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki}, \hat{\psi}_k)}) N_{ki}^x(dt) d(\hat{\rho}_k^{aug}(\mathbf{W}_{ki}, u)) \right. \\ \left. - Y_{ki}(t) \exp((\hat{\boldsymbol{\beta}}^{aug}(u)^T \boldsymbol{Z}_{ki}(t)) \hat{\Lambda}_{0k}^{aug}(dt, du) \right\},\end{aligned} \quad (22)$$

where $H(v,u) = \int_a^v (\hat{\Sigma}_{aug}(x))^{-1} K_h(u - x) dx$..

Following an application of Lemma 1 of Sun and Wu (2005), the distribution of $\mathbf{W}_B(v)$, $v \in [a, b]$, can be approximated by the conditional distribution of $\mathbf{W}_B^*(v)$, $v \in [a, b]$, given the observed data sequence, which can be obtained through repeatedly generating independent sets of 23 $\{\xi_{ki}, i = 1,\ldots, n_k, k = 1,\ldots, K\}$. Hence, the distribution of $Q^{(1)}(v)$, $v \in [a, b]$, under H_{10}, can be approximated by the conditional distribution of $W_{B_1}^*$, $v \in [a, b]$, given the observed data sequence. By the continuous mapping theorem, the distribution of $Q^{(2)}(v)$, v

∈ [a, b], under H_{20}, can be approximated by the conditional distribution of Γ (v, $W^*_{B_1}$), v ∈ [a, b], given the observed data sequence.

With the Gaussian multiplier method, the variance $\mathrm{Var}\left\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\right\}$ can be consistently estimated by $\hat{\mathrm{Var}}\left\{\hat{B}_1^{aug}(v)-\hat{B}_1^{aug}(a)\right\}=n^{-1}\mathrm{Var}*(W^*_{B_1}(v))\mathrm{where}\ \ \mathrm{Var}*(W^*_{B_1}(v))$ is the first component on the diagonal of

$$\begin{aligned}&\mathrm{Cov}*(\mathbf{W}^*_B(v))\\ &\quad=\mathrm{Cov}(\mathbf{W}^*_B(v)|\mathrm{observed\ \ data})\\ &\quad=n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left[\int_0^1\int_0^\tau H(v,u)\left\{\overline{Z}_{ki}(t)-\overline{\boldsymbol{Z}}_k(t,\hat{\boldsymbol{\beta}}^{aug}(u))\right\}\left\{\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki},\hat{\psi}_k)}N_{ki}(dt,du)\right.\right.\\ &\quad+\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki},\hat{\psi}_k)}N^x_{ki}(dt)d(\hat{\rho}_k(\mathbf{W}_{ki},u))\\ &\qquad\quad-Y_{ki}(t)\exp((\hat{\boldsymbol{\beta}}^{aug}(u))^T\boldsymbol{Z}_{ki}(t))\hat{\Lambda}^{aug}_{0k}(dt,du)\}]^{\otimes 2}.\end{aligned}\tag{23}$$

Proof of Theorem 1

Let

$$\begin{aligned}&\mathscr{A}_{ki}(v)=\int_0^1\int_0^\tau K_h(u-v)(\boldsymbol{Z}_{ki}(t))-\overline{\mathbf{z}}_k(t,\boldsymbol{\beta}(u)))\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}M_{ki}(dt,du),\\ &\mathscr{B}_{ki}(v)=\int_0^1\int_0^\tau K_h(u-v)(\boldsymbol{Z}_{ki}(t)-\overline{\mathbf{z}}_k(t,\boldsymbol{\beta}(u)))\left(1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}\right)E\left\{M_{ki}(dt,du)|\mathbf{Q}_{ki}\right\}.\end{aligned}\tag{24}$$

Following the proof of Theorem 4 of Sun and Gilbert (2012, the web Appendix (W.19)) and under $nh_4 \to 0$,

$$n^{1/2}\left\{\hat{\boldsymbol{\beta}}^{aug}(v)-\boldsymbol{\beta}(v)\right\}=-(\Sigma(v))^{-1}n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathscr{A}_{ki}(v)+\mathscr{B}_{ki}(v))+o_p(1).\tag{25}$$

Hence

$$n^{1/2}(\hat{\mathbf{B}}^{aug}(v)-\mathbf{B}(v))=-n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^v(\Sigma(u))^{-1}(\mathscr{A}_{ki}(u)+\mathscr{B}_{ki}(u))du+o_p(1),$$

which, by exchanging the order of integrations, equals to

$$
\begin{aligned}
&n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\textstyle\int_0^1\int_0^\tau[\int_0^v K_h(u-x)\{\Sigma(x)\}^{-1}dx][\boldsymbol{Z}_{ki}(t)-\overline{\mathbf{z}}_k\{t,\boldsymbol{\beta}(u)\}]\\
&\left[\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}M_{ki}(dt,du)+\left\{1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}\right\}E\left\{M_{ki}(dt,du)\middle|\mathbf{Q}_{ki}\right\}\right]).
\end{aligned}
\tag{26}
$$

Let

$$
\tilde{\mathbf{J}}_n(v)=n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\textstyle\int_0^v\int_0^\tau[\boldsymbol{Z}_{ki}(t)-\overline{\mathbf{z}}_k(t,\boldsymbol{\beta}(u))]\ [\displaystyle\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}M_{ki}(dt,du)+\left\{1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}\right\}E\left\{M_{ki}(dt,du)|\mathbf{Q}_{ki}\right\}].
$$

It follows that

$$
n^{1/2}(\hat{\mathbf{B}}^{aug}(v)-\mathbf{B}(v))=-\textstyle\int_0^v(\Sigma(u))^{-1}\int_0^1 K_h(x-u)\tilde{\mathbf{J}}_n(dx)du+o_p(1)=-\int_0^1[\int_0^1(\Sigma(u))^{-1}K_h(x-u)du]\tilde{\mathbf{J}}_n(dx)+o_p(1).
\tag{27}
$$

Since the kernel function *K*(·) has compact support on [−1, 1], (27) equals to

$$
\begin{aligned}
&-\textstyle\int_h^{v-h}[\int_0^v(\Sigma(u))^{-1}K_h(x-u)du]\tilde{\mathbf{J}}_n(dx)\\
&\quad-\textstyle\int_{-h}^{h}[\int_0^v(\Sigma(u))^{-1}K_h(x-u)du]\tilde{\mathbf{J}}_n(dx)\\
&\quad-\textstyle\int_{v-h}^{v+h}[\int_0^v(\Sigma(u))^{-1}K_h(x-u)du]\tilde{\mathbf{J}}_n(dx)+o_p(1).
\end{aligned}
\tag{28}
$$

It can be shown that $\tilde{\mathbf{J}}_n$ (*x*) converges weakly to a mean-zero Gaussian process with continuous paths. Under the assumption (A.4), $\int_0^v(\Sigma(u))^{-1}K_h(x-u)du$ has bounded variation and converges uniformly to Σ(*x*)$^{-1}$ for *x* ∈ (*h*, *v* − *h*). By Lemma 2 of Gilbert et al. (2008), the first term in (28) is equal to $-\int_0^v(\Sigma(u))^{-1}\tilde{\mathbf{J}}_n(dx)+o_p(1)$. Similar arguments lead to the second and the third terms in (28) to be o_p (1). Hence,

$$
\begin{aligned}
&n^{1/2}(\hat{\mathbf{B}}^{a}(v)-\mathbf{B}(v))\\
&=n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{nk}\left(\textstyle\int_0^v\int_0^\tau\{\Sigma(u)\}^{-1}[\boldsymbol{Z}_{ki}(t)-\overline{\mathbf{z}}_k\{t,\boldsymbol{\beta}(u)\}]\right.\\
&\left.\left[\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}M_{ki}(dt,du)\middle|\left\{1-\frac{R_{ki}}{\pi_k(\mathbf{Q}_{ki})}\right\}E\left\{M_{ki}(dt,du)|\,\mathbf{Q}_{ki}\right\}\right]\right)+o_p(1),
\end{aligned}
$$

which converges weakly to a *p*-dimensional mean-zero Gaussian process on *v* ∈ [*a*, *b*] with continuous sample paths by Lemma 1 of Sun and Wu (2005). Theorem 1 follows since $\mathbf{W}_B$ (*v*) = $n^{1/2}$ {$\hat{\mathbf{B}}^{aug}$ (*v*) − **B**(*v*)} − $n^{1/2}$ {$\hat{\mathbf{B}}^{aug}$ (*a*) − **B**(*a*)} is a linear transformation of $n^{1/2}$ ($\hat{\mathbf{B}}^{aug}$ (·) − **B**(·)).

References

Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics. 1978; 5:141–150.

Cai Z, Fan J, Runze Li. Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association. 2000; 95:888–902.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association. 1979; 74:829–836.

Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.

Fauci AS, Johnston MI, Dieffenbach CW, Burton DR, Hammer SM, Hoxie JA, Martin M, Overbaugh J, Watkins DI, Mahmoud A, Greene WC. HIV vaccine research: the way forward. Science. 2008; 321:530–532. [PubMed: 18653883]

Gilbert PB, Self SG, Ashby MA. Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. Biometrics. 1998; 54:799–814. [PubMed: 9750238]

Gilbert PB, Lele S, Vardi Y. Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. Biometrika. 1999; 86:27–43.

Gilbert PB, McKeague IW, Sun Y. The two-sample problem for failure rates depending on a continuous mark: An application to vaccine efficacy. Biostatistics. 2008; 9:263–276. [PubMed: 17704528]

Gilbert PB, Berger JO, Stablein D, Becker S, Essex M, Hammer SM, Kim JH, Degruttola VG. Statistical interpretation of the RV144 HIV vaccine efficacy trial in Thailand: A case study for statistical issues in efficacy trials. Journal of Infectious Diseases. 2011; 203:969–975. [PubMed: 21402548]

Hoover DR, Rice JA, Wu CO, Yang P-L. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85:809–822.

Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 1993; 80:557–572.

Moore PL, Ranchobe N, Lambson BE, Gray ES, Cave E, Abrahams M-R, Bandawe G, Mlisana K, Abdool Karim SS, Williamson C, Morris L. the CAPRISA 002 study and the NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI). Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. PLoS Pathogens. 2009; 5:e1000598. [PubMed: 19763271]

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. HIV-specific probabilistic models of protein evolution. PLoS ONE. 2007; 2(6):e503. [PubMed: 17551583]

Prentice RL, Kalbfleisch JD, Peterson AV Jr, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978; 34:541–554. [PubMed: 373811]

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, de Souza M, Adams E, Benenson M, Gurunathan S, Tartaglia J, McNeil JG, Francis DP, Stablein D, Birx DL, Chunsuttiwat S, Khamboonruang C, Thongcharoen P, Robb ML, Michael NL, Kunasol P, Kim JH. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. New England Journal of Medicine. 2009; 361:2209–2220. [PubMed: 19843557]

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 B, 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; 17:366–371.

Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, deCamp AC, Carrico C, Menis S, Magaret CA, Ahmed H, Juraska M, Chen L, Konopa P, Nariya S, Stoddard JN, Wong K, Zhao H, Deng W, Maust BS, Bose M, Howell S, Bates A, Lazzaro M, O'Sullivan A, Lei E, Bradfield A, Ibitamuno G, Assawadarachai V, O'Connell RJ, deSouza MS, Nitayaphan S, Rerks-Ngarm S, Robb ML, McLellan JS, Georgiev I, Kwong PD, Carlson JM, Michael NL, Schief WR, Gilbert PB, Mullins JI, Kim JH. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature. 2012; 490:417–420. [PubMed: 22960785]

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.

Stephen MA. Edf statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. 1974; 69:730–737.

Sun Y, Gilbert PB. Estimation of stratified mark-specific proportional hazards models with missing marks. Scandinavian Journal of Statistics. 2012; 39:34–52.

Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. The Annals of Statistics. 2009; 37:394–426.

Sun Y, Wu H. Semiparametric time-varying coefficients regression model for longitudinal data. Scandinavian Journal of Statistics. 2005; 32:21–47.

Tian L, Zucker D, Wei LJ. On the Cox model with time-varying regression coefficients. Journal of the American Statistical Association. 2005; 100:172–183.

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD, Shaw GM. Antibody neutralization and escape by HIV-1. Nature. 2003; 422:307–312. [PubMed: 12646921]

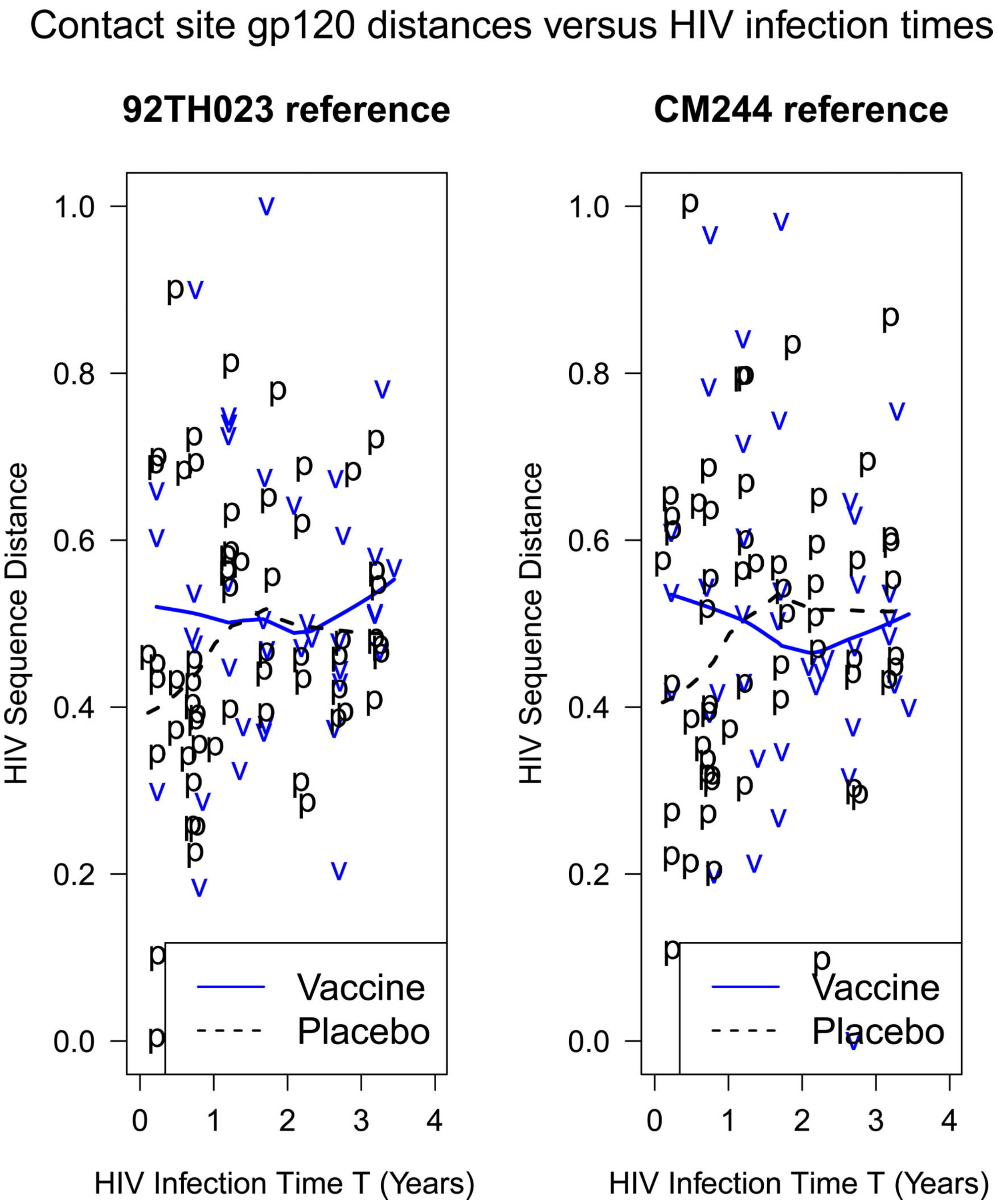

Figure 1.

Scatterplots of the marks *V* versus the HIV infection time *T* for the 98 HIV infected subjects in the Thai trial with an observed mark. The mark *V* is the HIV-specific PAM-matrix (Nickle et al., 2007) weighted Hamming distances between a subject's HIV Envelope gp120 amino acid sequence (nearest to his/her consensus sequence) and the 92TH023 or CM244 vaccine reference sequence; the distances restrict to the 172 amino acid sites in gp120 documented to contact broadly neutralizing monoclonal antibodies. The lines are lowess smooth fits (Cleveland, 1979).

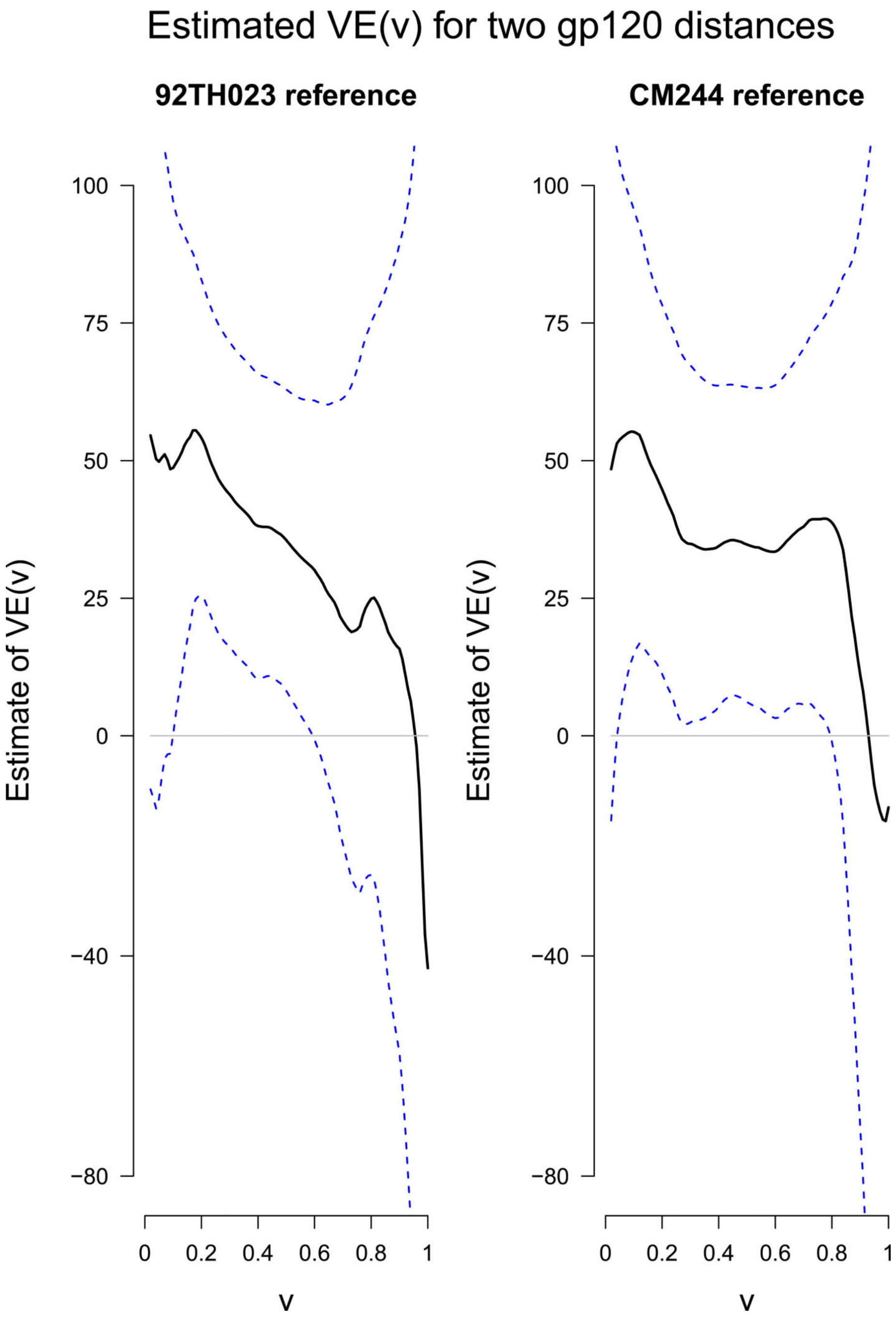

Figure 2.

AIPW estimation of VE(v) and 95% pointwise confidence bands without using auxiliary variables for the Thai trial with bandwidths $h_1 = 0.5$, $h_2 = h = 0.3$, for the monoclonal antibody contact site distances to the 92TH023 and CM244 reference sequences.

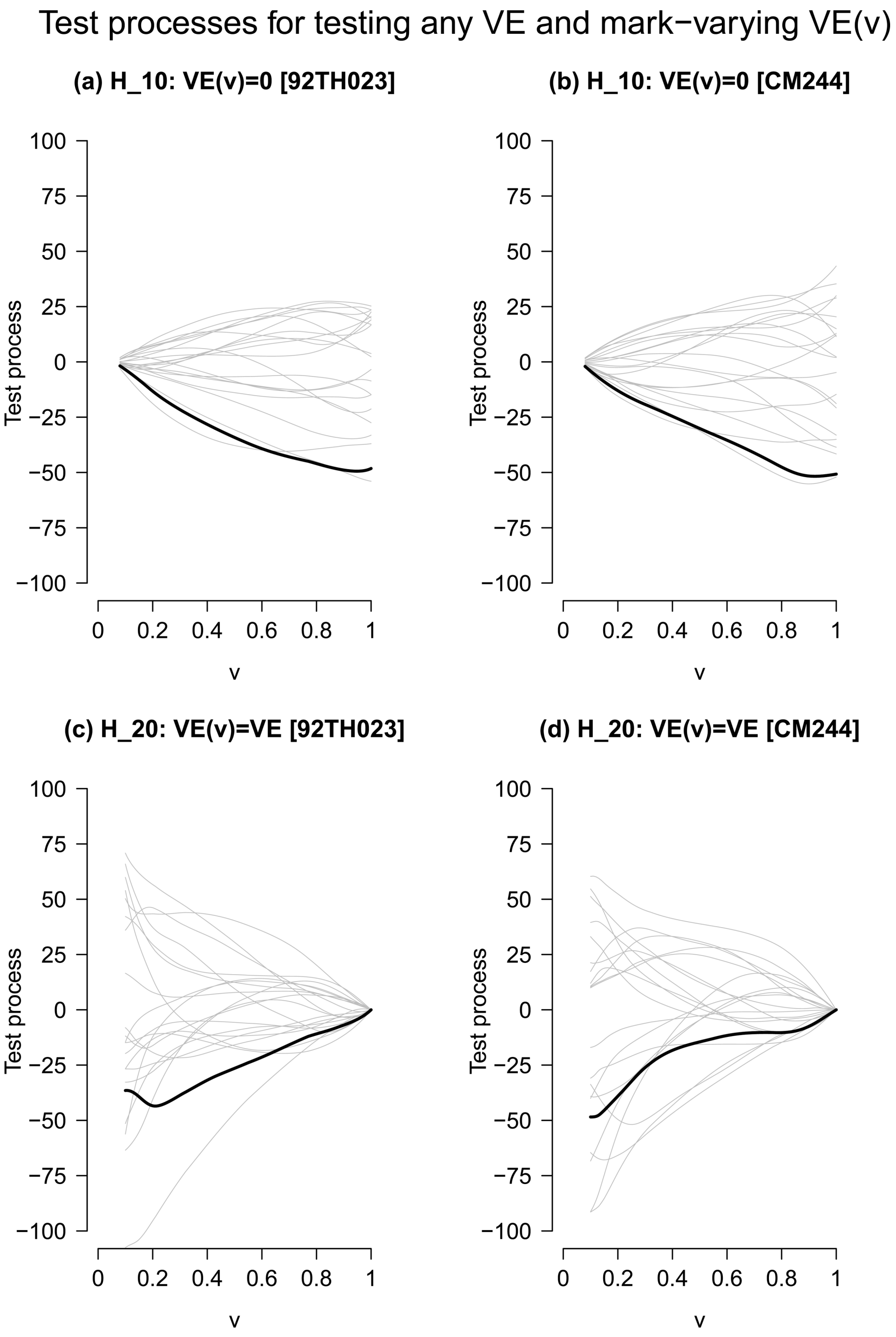

Figure 3.

Diagnostic plots of the test processes for the Thai trial data set with bandwidths h_1 = 0.5, h_2 = h = 0.3 and a = 0.05, b = 1 and a'= a + 0.01 without using auxiliary variables. (a) and (b) Plots of $Q^{(1)}(v)$ (solid dark line) versus 20 realizations (grey lines) from the Gaussian multiplier process $W^*_{B_1}(v)$ (92THf023, CM244 reference). (c) and (d) Plots of $Q^{(2)}(v)$ (solid dark line) versus 20 realizations (grey lines) from the Gaussian multiplier process $\Gamma(v, W^*_{B_1})$ (92TH023, CM244 reference).

Table 1

Empirical sizes and powers of the tests $T_{a1}^{(1)}, T_{a2}^{(1)}, T_{m1}^{(1)} \text{and} T_{m2}^{(1)}$ for testing H_{10} at the nominal level 0.05 for ρ = 0 and 0.92 when 50% of the marks are missing. The bandwidths are $h_1 = 0.1$ and $h_2 = h$. Each entry is based on 500 Gaussian multipliers samples and 500 repetitions.

				Size/Power							
				ρ = 0				ρ = 0.92			
Model	**(α, β, γ)**	***n***	**h**	$T_{a1}^{(1)}$	$T_{a2}^{(1)}$	$T_{m1}^{(1)}$	$T_{m2}^{(1)}$	$T_{a1}^{(1)}$	$T_{a2}^{(1)}$	$T_{m1}^{(1)}$	$T_{m2}^{(1)}$
M1	(0, 0, 0.3)	500	0.15	5.4	4.0	4.0	5.0	4.6	4.2	3.8	4.2
			0.20	5.0	4.4	4.6	5.2	4.8	4.0	4.2	3.6
		800	0.10	3.8	3.6	4.2	4.2	3.8	3.8	5.4	4.8
			0.15	4.0	3.8	4.6	4.6	5.0	4.4	5.4	5.6
M3	(−0.6, 0.6, 0.3)	500	0.15	68.2	67.0	79.4	76.0	73.2	74.6	83.2	85.4
			0.20	63.2	65.0	75.8	74.2	69.2	71.4	79.8	82.6
		800	0.10	88.2	86.2	94.6	90.4	92.0	93.0	95.0	97.2
			0.15	87.4	86.6	92.8	90.8	89.2	90.6	93.4	95.2
M4	(−1.2, 1.2, 0.3)	500	0.15	99.6	99.4	99.8	99.8	99.8	100	99.8	100
			0.20	99.4	99.0	99.6	99.8	99.6	99.8	99.8	100
		800	0.10	100	100	100	100	100	100	100	100
			0.15	100	100	100	100	100	100	100	100
M2	(−0.69, 0, 0.3)	500	0.15	100	100	100	99.8	100	100	100	100
			0.20	100	100	100	100	100	100	100	100
		800	0.10	100	99.8	100	100	99.8	99.8	99.8	99.8
			0.15	100	100	100	100	100	100	100	100

Table 2

Empirical sizes and powers of the tests $T_{a1}^{(2)}, T_{a2}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for testing H_{20} at the nominal level 0.05 for ρ = 0 and 0.92 when 50% of the marks are missing. The bandwidths are $h_1 = 0.1$ and $h_2 = h$. Each entry is based on 500 Gaussian multipliers samples and 500 repetitions.

Model	**(α, β, γ)**	***n***	**h**	**Size/Power**							
				ρ = 0				**ρ = 0.92**			
				$T_{a1}^{(2)}$	$T_{a2}^{(2)}$	$T_{m1}^{(2)}$	$T_{m2}^{(2)}$	$T_{a1}^{(2)}$	$T_{a2}^{(2)}$	$T_{m1}^{(2)}$	$T_{m2}^{(2)}$
M2	(−0.69, 0, 0.3)	500	0.15	5.6	4.8	5.8	5.8	7.6	7.2	7.4	7.0
			0.20	5.8	4.8	5.4	5.2	6.6	6.6	6.6	7.4
		800	0.10	6.4	5.0	5.6	5.8	6.2	5.8	7.2	7.0
			0.15	6.6	5.2	5.8	5.6	6.0	5.6	6.0	6.6
M3	(−0.6, 0.6, 0.3)	500	0.15	16.8	17.0	22.4	25.2	20.6	25.8	32.6	37.4
			0.20	14.2	15.8	22.2	24.8	19.4	24.2	31.8	34.6
		800	0.10	26.0	25.8	35.2	36.4	36.0	38.0	46.0	49.2
			0.15	25.4	25.8	34.8	35.6	34.0	36.0	45.4	47.4
M4	(−1.2, 1.2, 0.3)	500	0.15	44.4	46.2	59.0	63.2	63.6	68.4	76.4	80.2
			0.20	42.2	44.0	57.2	59.6	61.4	65.8	73.2	75.8
		800	0.10	66.2	67.6	75.2	78.0	82.8	86.6	90.6	91.8
			0.15	64.6	66.2	74.0	77.0	80.6	84.4	88.4	91.2
M5	(−1.5, 1.5, 0.3)	500	0.15	64.5	66.5	75.0	76.5	81.0	85.6	88.8	90.4
			0.20	61.0	62.6	72.2	72.2	77.8	82.4	86.8	89.4
		800	0.10	80.8	85.6	87.6	91.4	94.6	96.2	97.6	98.4
			0.15	78.6	84.8	87.8	91.4	94.4	95.6	95.8	97.8

Table 3

Empirical sizes of the tests for H_{10} and H_{20} at the nominal level 0.05 using the complete cases under MCAR when 50% of the marks are missing. The bandwidths are $h_1 = 0.1$ and $h_2 = h$. Each entry is based on 100 Gaussian multipliers samples and 100 repetitions.

Model	**Missing Model**	***n***	***h***	**Size**			
				testing H_{10}			
				$T_{a1}^{(1)}$	$T_{a2}^{(1)}$	$T_{m1}^{(1)}$	$T_{m2}^{(1)}$
M1	(13)	500	0.20	0.14	0.10	0.12	0.15
		800	0.15	0.10	0.07	0.11	0.11
	(16)	500	0.20	0.39	0.37	0.50	0.42
		800	0.15	0.50	0.46	0.63	0.55
				testing H_{20}			
				$T_{a1}^{(2)}$	$T_{a2}^{(2)}$	$T_{m1}^{(2)}$	$T_{m2}^{(2)}$
M2	(13)	500	0.20	0.08	0.04	0.08	0.05
		800	0.15	0.06	0.09	0.06	0.10
	(16)	500	0.20	0.08	0.07	0.08	0.05
		800	0.15	0.07	0.06	0.12	0.14

Table 4

Robustness of the tests for H_{10}. Empirical sizes and powers of the tests $T_{a1}^{(1)}, T_{a2}^{(1)}, T_{m1}^{(1)} \text{and} T_{m2}^{(1)}$ for testing H_{10} at the nominal level 0.05 for n = 500 and h = 0.2 when 50% of the marks are missing. The bandwidths are h_1 = 0.1 and h_2 = h. Each entry is based on 500 Gaussian multipliers samples and 500 repetitions.

Model	**(α, β, γ)**	**Size/Power**			
		$T_{a1}^{(1)}$	$T_{a2}^{(1)}$	$T_{m1}^{(1)}$	$T_{m2}^{(1)}$
		r$r_k(w)$ is misspecified			
M1	(0, 0, 0.3)	4.2	5.2	3.6	4.2
M3	(−0.6, 0.6, 0.3)	62.0	74.4	74.0	81.8
M4	(−1.2, 1.2, 0.3)	99.6	99.8	99.8	99.8
M2	(−0.69, 0, 0.3)	100	100	100	100
		$gk(a\|t, v, z)$ is misspecified			
M1	(0, 0, 0.3)	3.4	4.2	5.8	4.6
M3	(−0.6, 0.6, 0.3)	59.6	64.4	72.8	74.4
M4	(−1.2, 1.2, 0.3)	99.2	99.4	99.6	99.6
M2	(−0.69, 0, 0.3)	100	99.8	100	99.8
		$r_k(w)$ and $g_k(a\|t, v, z)$ are misspecified			
M1	(0, 0, 0.3)	4.0	4.0	3.8	3.4
M3	(−0.6, 0.6, 0.3)	61.8	61.8	71.8	73.8
M4	(−1.2, 1.2, 0.3)	99.6	98.6	99.8	99.8
M2	(−0.69, 0, 0.3)	100	100	100	100
		missing-at-random assumption is violated			
M1	(0, 0, 0.3)	3.4	3.8	3.6	5.0
M3	(−0.6, 0.6, 0.3)	60.6	67.0	73.0	77.8
M4	(−1.2, 1.2, 0.3)	99.2	99.6	99.8	99.6
M2	(−0.69, 0, 0.3)	100	100	100	100

Table 5

Robustness of the tests for H_{20}. Empirical sizes and powers of the tests $T_{a1}^{(2)}, T_{a2}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for testing H_{20} at the nominal level 0.05 for n = 500 and h = 0.2 when 50% of the marks are missing. The bandwidths are h_1 = 0.1 and h_2 = h. Each entry is based on 500 Gaussian multipliers samples and 500 repetitions.

Model	**(α, β, γ)**	**Size/Power**			
		$T_{a1}^{(2)}$	$T_{a2}^{(2)}$	$T_{m1}^{(2)}$	$T_{m2}^{(2)}$
		r_k(**w**) is misspecified			
M2	(−0.69, 0, 0.3)	5.0	3.8	5.4	6.2
M3	(−0.6, 0.6, 0.3)	24.0	25.2	34.2	36.0
M4	(−1.2, 1.2, 0.3)	60.8	66.6	72.8	78.4
M5	(−1.5, 1.5, 0.3)	76.6	82.0	85.8	88.8
		g_k (**a**|t, v, **z**) is misspecified			
M2	(−0.69, 0, 0.3)	4.8	6.6	6.0	5.8
M3	(−0.6, 0.6, 0.3)	17.2	18.0	28.4	28.2
M4	(−1.2, 1.2, 0.3)	44.8	47.2	56.4	61.0
M5	(−1.5, 1.5, 0.3)	58.0	60.4	68.6	73.2
		r_k (**w**) and g_k (**a**|t, v, **z**) are misspecified			
M2	(−0.69, 0, 0.3)	4.0	4.8	4.4	4.4
M3	(−0.6, 0.6, 0.3)	16.6	19.6	26.8	26.6
M4	(−1.2, 1.2, 0.3)	43.2	46.6	55.6	60.6
M5	(−1.5, 1.5, 0.3)	53.8	58.8	67.4	71.4
		missing-at-random assumption is violated			
M2	(−0.69, 0, 0.3)	6.8	6.0	7.6	7.8
M3	(−0.6, 0.6, 0.3)	28.6	33.6	39.6	42.0
M4	(−1.2, 1.2, 0.3)	61.8	67.0	74.0	78.4
M5	(−1.5, 1.5, 0.3)	77.4	81.6	85.4	89.2

Table 6

Power of the tests $T_{a1}^{(1)}, T_{a2}^{(1)}, T_{m1}^{(1)}, T_{m2}^{(1)}, T_{a1}^{(2)}, T_{a1}^{(2)}, T_{m1}^{(2)} \text{and} T_{m2}^{(2)}$ for the Thai trial at the nominal level 0.05. Each entry is based on 100 Gaussian multipliers samples and 100 repetitions.

			Power							
			testing *H*$_{10}$				**testing *H*$_{20}$**			
ρ	**% missing marks**	***h***	$T_{a1}^{(1)}$	$T_{a2}^{(1)}$	$T_{m1}^{(1)}$	$T_{m2}^{(1)}$	$T_{a1}^{(2)}$	$T_{a2}^{(2)}$	$T_{m1}^{(2)}$	$T_{m2}^{(2)}$
	0	0.15	77	85	86	95	48	48	59	60
0	25	0.2	67	76	79	85	36	33	50	47
	50	0.3	63	71	71	82	29	27	37	42
	75	0.4	41	51	59	58	21	18	35	31
0.78	25	0.2	67	79	82	89	36	39	46	50
	50	0.3	60	71	74	84	28	28	41	39
	75	0.4	49	53	63	65	25	25	34	34
0.92	25	0.2	70	80	84	91	37	41	50	56
	50	0.3	61	71	73	87	35	39	50	51
	75	0.4	54	58	62	71	30	33	40	44
0.98	25	0.2	71	81	82	91	39	47	53	55
	50	0.3	66	76	75	86	44	42	50	52
	75	0.4	56	66	68	76	41	43	51	49