Biostatistics (2008), **9**, 2, *pp.* 263–276
doi:10.1093/biostatistics/kxm028
Advance Access publication on August 17, 2007

The 2-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy

PETER B. GILBERT*

Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA
pgilbert@scharp.org

IAN W. MCKEAGUE

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, 6th Floor, New York, NY 10032, USA

YANQING SUN

Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC 28223, USA

SUMMARY

The efficacy of an HIV vaccine to prevent infection is likely to depend on the genetic variation of the exposing virus. This paper addresses the problem of using data on the HIV sequences that infect vaccine efficacy trial participants to (1) test for vaccine efficacy more powerfully than procedures that ignore the sequence data and (2) evaluate the dependence of vaccine efficacy on the divergence of infecting HIV strains from the HIV strain that is contained in the vaccine. Because hundreds of amino acid sites in each HIV genome are sequenced, it is natural to treat the genetic divergence as a continuous mark variable that accompanies each failure (infection) time. Problems (1) and (2) can then be approached by testing whether the ratio of the mark-specific hazard functions for the vaccine and placebo groups is unity or independent of the mark. We develop nonparametric and semiparametric tests for these null hypotheses and nonparametric techniques for estimating the mark-specific relative risks. The asymptotic properties of the procedures are established. In addition, the methods are studied in simulations and are applied to HIV genetic sequence data collected in the first HIV vaccine efficacy trial.

Keywords: Competing risks; Genetic data; Mark variable; Nonparametric statistics; Proportional hazards; Survival analysis.

1. INTRODUCTION

Competing risks failure time data consist of survival times and a mark variable that describes a feature of the failure. Often such data are available for 2 treatment groups, and it is of interest to account for

*To whom correspondence should be addressed.

this mark in comparing the failure experience. In this article, we develop procedures to assess continuous mark-specific relative risks. This departs from our recent work (Gilbert *and others*, 2004) in which we developed a test for the dependence of a single continuous mark-specific hazard rate on the mark variable (i.e. the "1-sample" problem). We expand the scope of research to include estimation as well as testing and semiparametric as well as nonparametric hypothesis testing.

We are motivated by applications in preventive HIV vaccine efficacy trials. The extensive genetic diversity of HIV poses one of the greatest challenges to developing an AIDS vaccine. Vaccine efficacy to prevent infection, usually defined in terms of the hazard ratio between vaccine and placebo recipients, may decrease with the viral genetic sequence divergence of a challenge HIV from the virus or viruses represented in the vaccine construct (Gilbert *and others*, 1999). Detecting that the vaccine protects against some strains but not others, and quantifying the relationship between vaccine efficacy and viral divergence, is useful for guiding vaccine deployment decisions and for designing new vaccines that provide greater breadth of protection.

From 1998 to 2003, VaxGen Inc. conducted the first HIV vaccine efficacy trial (Flynn *and others*, 2005). HIV-uninfected volunteers at high risk for acquiring HIV were randomized to receive the vaccine AIDSVAX ($n_1 = 3598$) or placebo ($n_2 = 1805$). Subjects were monitored for 3 years for the primary study end point HIV infection. For each subject who became HIV infected, the envelope glycoprotein (gp120) region of the infecting virus was sequenced. Of the 368 subjects who acquired HIV, the sequence data were collected for 336 subjects (217 of 241 vaccine and 119 of 127 placebo). VaxGen hypothesized that the level of vaccine efficacy would be higher against HIVs with gp120 amino acid sequences that were relatively similar to either of the 2 HIV strains (named MN and GNE8) that were represented in the vaccine. The distance of each infecting virus to MN and GNE8 was measured by the percent mismatch in the aligned amino acid sequences (i.e. Hamming distance) for 3 sets of positions hypothesized to be important for neutralizing HIV (Wyatt *and others*, 1998): (1) the neutralizing face core of gp120 that was crystalized, (2) the neutralizing face core plus the variable loop V2/V3 regions, and (3) the V3 loop. For each metric and infecting virus, the mark is defined as the minimum of the 2 distances to the MN and GNE8 reference sequences.

Gilbert *and others* (1999) and Gilbert (2000) developed a semiparametric biased sampling model as a tool for studying vaccine efficacy as a function of a continuous mark. This model parametrically specifies the relationship between the vaccine efficacy and the mark and leaves the distribution of the mark in the infected placebo group unspecified. However, there are no data available for suggesting the correct parametric model, so nonparametric methods are desirable. Furthermore, the earlier work is limited by conditioning on infection, so odds ratios but not relative risks of infection can be estimated, and the model treats HIV infection as a binary outcome, ignoring the time to HIV infection. The methods presented here were developed because they are free from these limitations, as they are non- or semi-parametric, are prospective, and incorporate the failure times.

We introduce tests for the hypothesis that the continuous mark-specific risks in the 2 groups coincide and for the hypothesis that the relative mark-specific risk between the groups is independent of the mark. Let T_k be the time to end point and V_k be the mark variable for a representative individual in group k (in vaccine trials, $k = 1$ indicates vaccine and $k = 2$ indicates placebo). In the study, we observe $(X_k, \delta_k, \delta_k V_k)$, where $X_k = \min\{T_k, C_k\}$, $\delta_k = I(T_k \leqslant C_k)$, and C_k is a censoring time assumed to be independent of both T_k and V_k, $k = 1, 2$. The notation $\delta_k V_k$ indicates that the mark V_k is only observed if the failure time is observed ($\delta_k = 1$); if $\delta_k = 0$, V_k is undefined and not relevant. We assume that V_k has known and bounded support; rescaling V_k if necessary, this support is taken to be $[0, 1]$. The mark-specific hazard rate in group k is

$$\lambda_k(t, v) = \lim_{h_1, h_2 \to 0} P\{T_k \in [t, t + h_1), V_k \in [v, v + h_2) | T_k \geqslant t\} / h_1 h_2 \tag{1.1}$$

and the mark-specific cumulative incidence function is

$$F_k(t, v) = \lim_{h_2 \to 0} P\{T_k \leqslant t, V_k \in [v, v + h_2)\}/h_2, \tag{1.2}$$

$k = 1, 2$, with t ranging over a fixed interval $[0, \tau]$. These functions are natural extensions of the cause-specific hazard function and cumulative incidence function that have been extensively studied for a discrete mark variable (e.g. Prentice *and others*, 1978). Similar to the discrete mark case, the functions (1.1) and (1.2) are related by the equation $F_k(t, v) = \int_0^t \lambda_k(s, v) S_k(s) \mathrm{d}s$, where $S_k(t)$ is the survival function for group k, and are estimable from the observed group k competing risks failure time data.

A standard measure of vaccine efficacy to prevent infection at time t is the relative reduction in hazard due to vaccination: $\mathrm{VE}(t) = 1 - \lambda_1(t)/\lambda_2(t)$, see Halloran *and others* (1997). It is natural to extend this definition to allow the vaccine efficacy to depend on viral divergence: $\mathrm{VE}(t, v) = 1 - \lambda_1(t, v)/\lambda_2(t, v)$. To interpret $\mathrm{VE}(t, v)$, consider that $\lambda_k(t, v)$ aggregates many parameters that are not identifiable due to the absence of data on sexual and needle contacts. These parameters include (i) the participant distribution of per-contact susceptibility to strain v, (ii) the distribution of infectiousness of contacts infected with strain v, (iii) the mechanism of vaccine protection (Halloran *and others*, 1992), (iv) the density of V at v in the population of HIV-infected contacts, and (v) the participant distribution of contact rates. Despite these complicating factors, randomization, double blinding, and the fact that HIV infection is a rare event in HIV vaccine efficacy trials imply that $\mathrm{VE}(t, v)$ should approximately measure the vaccine effect to reduce susceptibility to HIV acquisition given exposure to strain v at time t. Strong assumptions about the parameters in (i)–(v) confer specific meaningful interpretations to $\mathrm{VE}(t, v)$, as explored in the supplementary material available at *Biostatistics* online.

The $\mathrm{VE}(t)$ and $\mathrm{VE}(t, v)$ measures must be used with care given that they do not account for the network structure of sexual and/or needle contacts. Alternative concepts of VE could be considered that explicitly consider this structure and incorporate individual-level models of vaccine response.

To account for the mark in testing for vaccine efficacy, we develop tests for

$$H_0^0 \colon \lambda_1(t, v) = \lambda_2(t, v), \quad \text{for } (t, v) \in [0, \tau] \times [0, 1],$$

against the following alternative hypotheses:

$$H_1^0 \colon \lambda_1(t, v) \leqslant \lambda_2(t, v), \quad \text{for all } (t, v) \in [0, \tau] \times [0, 1],$$

$$H_2^0 \colon \lambda_1(t, v) \neq \lambda_2(t, v), \quad \text{for some } (t, v) \in [0, \tau] \times [0, 1],$$

with strict inequality for some $(t, v) \in [0, \tau] \times [0, 1]$ in H_1^0. Testing H_0^0 evaluates $\mathrm{VE}(t, v) = 0$ for all t and v, that is, whether there is any vaccine efficacy against any HIV strain. As we show in simulations, tests of H_0^0 can have much greater power than standard tests of vaccine efficacy that ignore the mark, that is, that evaluate the null hypothesis $\lambda_1(t) = \lambda_2(t)$ for all $t \in [0, \tau]$. A test ignoring the mark should be done in conjunction with a test of H_0^0, however, to assess the overall clinical/public health benefit of the vaccine. To illustrate the importance of carrying out both tests, if $\mathrm{VE}(t) = 0$ and $\mathrm{VE}(t, v)$ is positive (negative) for $v \leqslant (>)\ 0.5$, then the vaccine clearly should not be declared effective. Yet, the analysis accounting for the mark would lead to follow-up studies of the mechanism by which the vaccine impacted mark-specific infection risk.

If H_0^0 is rejected, then it is of interest to assess if vaccine efficacy varies with strain distance. Accordingly, we also develop tests for

$$H_0 \colon \lambda_1(t, v)/\lambda_2(t, v) \text{ does not depend on } v \text{ for } t \in [0, \tau]$$

against the following alternative hypotheses:

$$H_1\colon\ \lambda_1(t,v_1)/\lambda_2(t,v_1) \leqslant \lambda_1(t,v_2)/\lambda_2(t,v_2), \quad \text{for all } v_1 \leqslant v_2,\ \ t \in [0,\tau],$$

$$H_2\colon\ \lambda_1(t,v_1)/\lambda_2(t,v_1) \neq \lambda_1(t,v_2)/\lambda_2(t,v_2), \quad \text{for some } v_1 \leqslant v_2,\ \ t \in [0,\tau],$$

with strict inequality for some t, v_1, v_2 in H_1. Because H_0 and H_1 can be reexpressed as H_0: $\mathrm{VE}(t,v) = \mathrm{VE}(t)$ for all t, v and H_1: $\mathrm{VE}(t,v_1) \leqslant \mathrm{VE}(t,v_2)$ for all $t, v_1 \geqslant v_2$ (with $<$ for some $v_1 > v_2$), testing H_0 versus H_1 assesses whether vaccine efficacy decreases with HIV sequence divergence. Scientifically, this is of particular interest to assess.

To develop test statistics for evaluating H_0, we will exploit the observation that H_0 holds if and only if the mark-specific relative risk coincides with the ordinary relative risk, that is, $\lambda_1(t,v)/\lambda_2(t,v) = \lambda_1(t)/\lambda_2(t)$ for all t, v, where $\lambda_k(t) = \int_0^1 \lambda_k(t,v)\mathrm{d}v$ is the group-k hazard irrespective of the mark. As for discrete competing risks, the density of the mark makes no contribution to the total hazard other than through the mark-specific hazard, which represents the hazard rate accounting for the density of a specific strain. In Section 2, we introduce the proposed procedures for testing H_0^0 and H_0. Large-sample results and a simulation technique needed to implement the test procedures are summarized in Section 3. In Section 4, we discuss nonparametric estimation of the mark-specific vaccine efficacy. We summarize the results of a simulation experiment in Section 5, and an application to data from the VaxGen trial is provided in Section 6. Section 7 contains concluding remarks. The supplementary material available at *Biostatistics* online (http://www.biostatistics.oxfordjournals.org) contains expanded details on the interpretation of $\mathrm{VE}(t,v)$ (Section A), details of the large-sample results and Gaussian multipliers simulation technique for estimating critical values (Section B), expanded simulation results (Section C), and proofs of the large-sample results (Section D).

2. TEST PROCEDURE

We base our approach on estimates of the doubly cumulative mark-specific hazard functions $\Lambda_k(t,v) = \int_0^v\int_0^t \lambda_k(s,u)\mathrm{d}s\,\mathrm{d}u$, $k = 1, 2$. Given the observation of i.i.d. replicates $(X_{ki}, \delta_{ki}, \delta_{ki}V_{ki})$, $i = 1, \ldots, n_k$, of $(X_k, \delta_k, \delta_k V_k)$, $k = 1, 2$, the nonparametric maximum likelihood estimator (MLE) of $\Lambda_k(t,v)$ is provided by the Nelson–Aalen-type estimator

$$\hat{\Lambda}_k(t,v) = \int_0^t \frac{N_k(\mathrm{d}s, v)}{Y_k(s)}, \quad t \geqslant 0,\ \ v \in [0,1], \tag{2.1}$$

where $Y_k(t) = \sum_{i=1}^{n_k} I(X_{ki} \geqslant t)$ is the size of the risk set for group k at time t, and

$$N_k(t,v) = \sum_{i=1}^{n_k} I(X_{ki} \leqslant t, \delta_{ki} = 1, V_{ki} \leqslant v)$$

is the marked counting process with jumps at the uncensored failure times X_{ki} and associated marks V_{ki}, see Huang and Louis (1998, (3.2)).

Our tests of H_0^0 are based on comparing $\hat{\Lambda}_1(t,v)$ and $\hat{\Lambda}_2(t,v)$ and of H_0 are based on comparing the nonparametric MLE of $\Lambda_1(t,v) - \Lambda_2(t,v)$ with an estimate under H_0. Since H_0 is equivalent to $\Lambda_1(t,v) = \int_0^t[\lambda_1(s)/\lambda_2(s)]\Lambda_2(\mathrm{d}s, v)$ for all t, v, under H_0 we may estimate the difference $\Lambda_1(t,v) - \Lambda_2(t,v)$ by $\int_0^t[(\hat{\lambda}_1(s)/\hat{\lambda}_2(s)) - 1]\hat{\Lambda}_2(\mathrm{d}s, v)$, where $\hat{\lambda}_k(t)$ is a nonparametric estimator of $\lambda_k(t)$, as discussed below. Alternatively, under a proportional marginal hazards assumption, $\lambda_1(t)/\lambda_2(t) = \exp(\beta)$, this difference may be estimated by $\int_0^t[\exp(\hat{\beta}) - 1]\hat{\Lambda}_2(\mathrm{d}s, v)$, where $\hat{\beta}$ is the maximum partial

likelihood estimator of β, which leads to a semiparametric test for H_0. The nonparametric approach makes minimal assumptions but requires smoothing over time, whereas the semiparametric approach avoids smoothing and in principle may provide greater power when the proportional hazards assumption holds.

For the nonparametric approach, we estimate each hazard function $\lambda_k(t)$ by kernel smoothing:

$$\hat{\lambda}_k(t) = \frac{1}{b_k}\int_0^{\tau+\delta} K\left(\frac{t-s}{b_k}\right)\mathrm{d}\hat{\Lambda}_k(s), \tag{2.2}$$

where $\hat{\Lambda}_k(s) = \int_0^t (1/Y_k(s))\mathrm{d}N_k(s)$ is the Nelson–Aalen estimator of $\Lambda_k(t) = \int_0^t \lambda_k(s)\mathrm{d}s$, with $N_k(t) = \sum_{i=1}^{n_k} I(X_{ki} \leqslant t, \delta_{ki} = 1)$. The kernel K is a bounded symmetric function with support $[-1, 1]$ and integral 1. The bandwidth b_k is a positive parameter that indicates the window $[t - b_k, t + b_k]$ over which $\hat{\Lambda}_k(t)$ is smoothed and converges to zero as $n_k \to \infty$.

2.1 *Test processes and test statistics*

Based on the above discussion, we introduce test processes of the form

$$L_n^r(t, \upsilon) = \sqrt{\frac{n_1 n_2}{n}}\int_a^t H_n(s)[\hat{\Lambda}_1(\mathrm{d}s, \upsilon) - \hat{r}(s)\hat{\Lambda}_2(\mathrm{d}s, \upsilon)] \tag{2.3}$$

for $t \geqslant 0, 0 \leqslant \upsilon \leqslant 1$, where $H_n(\cdot)$ is a suitable weight process converging to a nonrandom function $H(\cdot)$ and $a \geqslant 0$. The process $H_n(\cdot)$ may be used to upweight regions with less variability, improving power, or for other reasons considered in Sections 3.2 and B.3 of the supplementary material available at *Biostatistics* online.

The superscript r reflects the choice of process $\hat{r}(s)$ in the test process and indicates whether it is used to test H_0^0 (indicated by r as 1, corresponding to $\hat{r}(s) = 1$), to test H_0 nonparametrically (indicated by r as np; $\hat{r}(s) = \hat{\lambda}_1(s)/\hat{\lambda}_2(s)$) or to test H_0 semiparametrically (indicated by r as sp; $\hat{r}(s) = \exp(\hat{\beta})$). A simple calculation shows that for r as np, $[\cdot]$ in (2.3) compares $\hat{\Lambda}_1(\mathrm{d}s, \upsilon) - \hat{\Lambda}_2(\mathrm{d}s, \upsilon)$ to the nonparametric estimate of $\Lambda_1(\mathrm{d}s, \upsilon) - \Lambda_2(\mathrm{d}s, \upsilon)$ under H_0 described above. The parallel result holds for r as sp using the semiparametric estimate of $\Lambda_1(\mathrm{d}s, \upsilon) - \Lambda_2(\mathrm{d}s, \upsilon)$ under H_0.

A variety of test statistics can be formulated as functionals of $L_n^r(t, \upsilon)$. We develop integration-type and supremum-type statistics. With $w_V(\upsilon)$ a known nonnegative weight function, large values of the following statistics provide evidence against H_0^0 in the direction of H_0^1 (first 2 statistics) or H_0^2 (second 2 statistics):

$$\hat{U}_1^1 = L_n^1(\tau, 1), \quad \hat{U}_2^1 = \int_0^1 w_V(\upsilon)L_n^1(\tau, \upsilon)\mathrm{d}\upsilon, \tag{2.4}$$

$$\hat{U}_3^1 = |L_n^1(\tau, 1)|, \quad \hat{U}_4^1 = \int_0^1 w_V(\upsilon)(L_n^1(\tau, \upsilon))^2\,\mathrm{d}\upsilon. \tag{2.5}$$

For testing H_0, let $y_k(t) = P(X_k \geqslant t)$, let $\tilde{\tau} = \sup\{t\colon y_1(t) > 0 \text{ and } y_2(t) > 0\}$, and assume $\tau < \tilde{\tau}$. To simplify the proofs and the conditions on the rates of convergence concerning b_k, we take $a > 0$ and construct the test statistics from the process $L_n^r(t, \upsilon)$ over $a \leqslant t \leqslant \tau, 0 \leqslant \upsilon \leqslant 1$. In practice, however, there would be no harm in taking $a = 0$ in order to use as much of the data as possible (this is done in the simulations and application).

Set $\Delta_n^r(t, \upsilon_1, \upsilon_2) = L_n^r(t, \upsilon_1) + L_n^r(t, \upsilon_2) - 2L_n^r(t, (\upsilon_1 + \upsilon_2)/2)$. For r as np or sp, the following test statistics measure departures from H_0 in the direction of H_1 ($\hat{U}_1^r$) or H_2 ($\hat{U}_2^r$):

$$\hat{U}_1^r = \sup_{\upsilon_1 < \upsilon_2} \sup_{0 \leqslant t_1 < t_2 < \tau} \{\Delta_n^r(t_2, \upsilon_1, \upsilon_2) - \Delta_n^r(t_1, \upsilon_1, \upsilon_2)\}, \tag{2.6}$$

$$\hat{U}_2^r = \sup_{\upsilon_1 < \upsilon_2} \sup_{0 \leqslant t_1 < t_2 < \tau} |\Delta_n^r(t_2, \upsilon_1, \upsilon_2) - \Delta_n^r(t_1, \upsilon_1, \upsilon_2)|. \tag{2.7}$$

To motivate the statistics $\hat{U}_1^r$ and $\hat{U}_2^r$, we note from the proof of Theorem 2 that $(n/n_1n_2)^{1/2}[\Delta_n^r(t_2, \upsilon_1, \upsilon_2) - \Delta_n^r(t_1, \upsilon_1, \upsilon_2)]$ converges in probability to $\delta(t_1, t_2, \upsilon_1, \upsilon_2) = \int_{t_1}^{t_2} \int_{\frac{\upsilon_1+\upsilon_2}{2}}^{\upsilon_2} H(s)(\lambda_1(s, \upsilon) - r(s)\lambda_2(s, \upsilon))$ $\mathrm{d}\upsilon\, \mathrm{d}s - \int_{t_1}^{t_2} \int_{\upsilon_1}^{\frac{\upsilon_1+\upsilon_2}{2}} H(s)(\lambda_1(s, \upsilon) - r(s)\lambda_2(s, \upsilon))\mathrm{d}\upsilon\, \mathrm{d}s$, where $r(s) = \lambda_1(s)/\lambda_2(s)$ or $\exp(\beta)$. Under H_0, $\delta(t_1, t_2, \upsilon_1, \upsilon_2) = 0$ for all $t_1, t_2 \in [0, \tau]$ and $\upsilon_1, \upsilon_2 \in [0, 1]$. Under H_1 and some smoothness conditions, $\delta(t_1, t_2, \upsilon_1, \upsilon_2) > 0$ for some $t_1 < t_2 \in [0, \tau]$ and $\upsilon_1 < \upsilon_2 \in [0, 1]$. Therefore, large values of $\hat{U}_1^r$ ($\hat{U}_2^r$) provide evidence against H_0 in the direction of H_1 (H_2).

In Section 3, we provide results that all 3 processes $L_n^r(t, \upsilon)$ (indexed by r) converge weakly to a Gaussian process under the appropriate null hypothesis. We also state results on the consistency of the proposed tests against their alternatives and summarize a simulation procedure for determining the critical values of the $\hat{U}_j^r$.

3. LARGE-SAMPLE RESULTS

We summarize the asymptotic results with theorems and proofs relegated to the supplementary material available at *Biostatistics* online. Theorem 1 gives regularity conditions under which $L_n^r(t, \upsilon)$ defined in (2.3) converges weakly to a process $L^r(t, \upsilon)$ under H_0 as $n \to \infty$.

Let U_j^r be defined the same as $\hat{U}_j^r$ in (2.6) and (2.7), with $L_n^r(t, \upsilon)$ replaced with $L^r(t, \upsilon)$. By the continuous mapping theorem, $\hat{U}_j^{\mathrm{np}} \xrightarrow{\mathcal{D}} U_j^{\mathrm{np}}$ under H_0, so $P(\hat{U}_j^{\mathrm{np}} > c_{j\alpha}) \to \alpha$, where $c_{j\alpha}$ is the upper α-quantile of U_j^{np}. However, the $c_{j\alpha}$ are unknown and very difficult to estimate due to the complicated nature of the limit process $L^{\mathrm{np}}(t, \upsilon)$. In Section 3.1, we summarize a Monte Carlo procedure to obtain each $c_{j\alpha}$. Theorem 2 establishes that each $\hat{U}_j^{\mathrm{np}}$ is consistent against its alternative, that is, $P(\hat{U}_1^{\mathrm{np}} > c_{1\alpha}) \to 1$ as $n \to \infty$ under H_1 and $P(\hat{U}_2^{\mathrm{np}} > c_{2\alpha}) \to 1$ as $n \to \infty$ under H_2. The parallel results hold for $\hat{U}_j^{\mathrm{l}}$ and $\hat{U}_j^{\mathrm{sp}}$.

3.1 *Gaussian multipliers simulation procedure*

We now summarize a Gaussian multipliers technique for simulating each of the test processes $L_n^{\mathrm{np}}(t, \upsilon)$, $L_n^{\mathrm{sp}}(t, \upsilon)$, and $L_n^{\mathrm{l}}(t, \upsilon)$ under the null hypothesis, cf. Lin *and others* (1993). Section B of the supplementary material available at *Biostatistics* online describes a process

$$L_n^{\mathrm{np}*}(t, \upsilon) = \sqrt{\frac{n_2}{n}} n_1^{-1/2} \sum_{i=1}^{n_1} \hat{h}_{1i}(t, \upsilon) W_{1i} - \sqrt{\frac{n_1}{n}} n_2^{-1/2} \sum_{i=1}^{n_2} \hat{h}_{2i}(t, \upsilon) W_{2i}, \tag{3.1}$$

where $\hat{h}_{1i}(t, \upsilon)$ and $\hat{h}_{2i}(t, \upsilon)$ are functions of the data and $W_{ki}, i = 1, \ldots, n_k, k = 1, 2$, are i.i.d. standard normal random variables. Theorem 3 states that the conditional weak limit of the process $L_n^{\mathrm{np}*}(t, \upsilon)$ given the observed data is the same as the weak limit of $L_n^{\mathrm{np}}(t, \upsilon)$ under H_0. This result implies that asymptotically consistent estimates of the critical values $c_{j\alpha}$ can be obtained by comparing $\hat{U}_j^{\mathrm{np}}$ to a null reference

distribution formed by the statistics $\hat{U}_j^{\mathrm{np}*}$ computed from $L_n^{\mathrm{np}*}(t,\upsilon)$ using replicate sampling of W_{1i} and W_{2i}. The supplementary material available at *Biostatistics* online describes parallel processes $L_n^{\mathrm{sp}*}(t,\upsilon)$ and $L_n^{1*}(t,\upsilon)$ and shows that the same Gaussian multipliers procedure can be used to consistently estimate critical values for the semiparametric tests of H_0 and the tests of H_0^0.

3.2 *Choice of weight process*

The test process is more variable at larger failure times, so it is advisable to choose the weight process to downweight the upper tail of the integral, and we suggest

$$H_n(s) = \sqrt{Y_1(s)Y_2(s)/n_1n_2}. \tag{3.2}$$

The test can be made invariant to the order of the 2 groups by including $\hat{r}(s)^{-1/2}$ in $H_n(s)$. The weight $H_n(s)$ can also be chosen to increase power against specific alternatives (Sun, 2001).

4. Estimation of mark-specific vaccine efficacy

Precise estimation of $\mathrm{VE}(t,\upsilon)$ introduced in Section 1 requires huge sample sizes because smoothing is required in both υ and t, and generally efficacy trials do not provide sufficient samples (Gilbert *and others*, 2002). Accordingly, we consider an alternative notion of mark-specific vaccine efficacy defined in terms of cumulative incidences:

$$\mathrm{VE}^{\mathrm{c}}(t,\upsilon) = 1 - F_1(t,\upsilon)/F_2(t,\upsilon),$$

which we call cumulative vaccine efficacy. This represents a time-averaged — rather than instantaneous — measure of vaccine efficacy and is much easier to estimate than $\mathrm{VE}(t,\upsilon)$. We also consider the doubly cumulative vaccine efficacy

$$\mathrm{VE}^{\mathrm{dc}}(t,\upsilon) = 1 - P(T_1 \leqslant t, V_1 \leqslant \upsilon)/P(T_2 \leqslant t, V_2 \leqslant \upsilon),$$

which can be estimated without any smoothing and with greater precision than $\mathrm{VE}^{\mathrm{c}}(t,\upsilon)$.

A nonparametric estimator of $\mathrm{VE}^{\mathrm{c}}(t,\upsilon)$ is given by $\widehat{\mathrm{VE}}^{\mathrm{c}}(t,\upsilon) = 1 - \hat{F}_1(t,\upsilon)/\hat{F}_2(t,\upsilon)$, where

$$\hat{F}_k(t,\upsilon) = \frac{1}{b_k}\int_0^1\int_0^t \frac{\hat{S}_k(s-)}{Y_k(s)} K\left(\frac{\upsilon-u}{b_k}\right) N_k(\mathrm{d}s,\mathrm{d}u), \tag{4.1}$$

$\hat{S}_k(t)$ is the Kaplan–Meier estimate of $S_k(t)$, $K(\cdot)$ is a bounded symmetric kernel function with support $[-1,1]$ and integral 1, and $b_k > 0$ is a bandwidth. The estimator $\hat{F}_k(t,\upsilon)$ is the continuous analog of the estimator that has been used for a discrete mark (Prentice *and others*, 1978).

If $F_1(t,\upsilon) \neq 0$ and $F_2(t,\upsilon) \neq 0$, a $100(1-\alpha)\%$ pointwise confidence interval for $\mathrm{VE}^{\mathrm{c}}(t,\upsilon)$ can be computed by transforming symmetric confidence limits about $\log(F_1(t,\upsilon)/F_2(t,\upsilon))$:

$$1 - (1 - \widehat{\mathrm{VE}}^{\mathrm{c}}(t,\upsilon))\exp\left(\pm z_{\alpha/2}\sqrt{\frac{\widehat{\mathrm{Var}}\{\hat{F}_1(t,\upsilon)\}}{\hat{F}_1(t,\upsilon)^2} + \frac{\widehat{\mathrm{Var}}\{\hat{F}_2(t,\upsilon)\}}{\hat{F}_2(t,\upsilon)^2}}\right),$$

$$\widehat{\mathrm{Var}}\{\hat{F}_k(t,\upsilon)\} = \frac{1}{b_k^2}\int_0^1\int_0^t \left[\frac{\hat{S}_k(s-)}{Y_k(s)} K\left(\frac{\upsilon-u}{b_k}\right)\right]^2 N_k(\mathrm{d}s,\mathrm{d}u). \tag{4.2}$$

To estimate $\mathrm{VE}^{\mathrm{dc}}(t,\upsilon)$, each $P(T_k \leqslant t, V_k \leqslant \upsilon)$ is simply estimated by $\int_0^t \{\hat{S}_k(s-)/Y_k(s)\}\, N_k(\mathrm{d}s,\upsilon)$, the standard estimator for the discrete cumulative incidence function for cause of failure defined by $V \leqslant \upsilon$, and its variance is estimated by $\int_0^t \{\hat{S}_k(s-)/Y_k(s)\}^2 N_k(\mathrm{d}s,\upsilon)$.

5. SIMULATION EXPERIMENT

The simulations are based on the features of the VaxGen trial described in Section 1. We study performance of the test statistics $\hat{U}_j^1$, $j = 1, 2, 3, 4$; $\hat{U}_j^{\text{np}}$ and $\hat{U}_j^{\text{sp}}$, $j = 1, 2$; and of $\widehat{\text{VE}}^{\text{c}}(\tau, \upsilon)$, with $\tau = 3$ years. We focus on the case that the mark-specific hazard function factors as $\lambda_k(t, \upsilon) = \lambda_k(t)c_k(\upsilon)$. Limited simulations under more complicated models showed comparable performance of the procedures (see supplementary Table 1 available at *Biostatistics* online). Under the factorization, the cumulative incidence function for group k is $F_k(t, \upsilon) = P\{T_k \leqslant t\}c_k(\upsilon)$. In the first set of simulations, we specify T_1 and T_2 to be exponential with parameters $\theta\lambda_2$ and λ_2, respectively, so that the cumulative vaccine efficacy by time τ irrespective of the mark V is given by $\text{VE}^{\text{c}}(\tau) = 1 - (1 - \exp(-\lambda_2\theta\tau))/(1 - \exp(-\lambda_2\tau))$, where λ_2 is the constant infection hazard rate in the placebo group. Here, θ is the constant infection hazard ratio between groups 1 and 2. In the second set of simulations, we specify non-proportional hazards $\lambda_1(t)$ and $\lambda_2(t)$ to examine the effect of violating the assumption used by the semiparametric tests of H_0. In this case, $\lambda_2(t) = \lambda_2$ as above and T_1 is distributed as Weibull with $\lambda_1(t) = 2\lambda_1 t$.

We select λ_2 so that 50% of placebo recipients are expected to be infected by $\tau = 36$ months, and consider $\text{VE}^{\text{c}}(\tau) = 0.0$, 0.33, and 0.67. Next, we specify

$$c_k(\upsilon) = [\beta_k(1.5^{1/\beta_k} - 0.5^{1/\beta_k})]^{-1}(\upsilon + 0.5)^{(1/\beta_k)-1}, \quad \text{for } 0 \leqslant \upsilon \leqslant 1.$$

The cumulative vaccine efficacy is given by

$$\text{VE}^{\text{c}}(\tau, \upsilon) = 1 - (1 - \text{VE}^{\text{c}}(\tau))\frac{\beta_2}{\beta_1}\left[\frac{1.5^{1/\beta_2} - 0.5^{1/\beta_2}}{1.5^{1/\beta_1} - 0.5^{1/\beta_1}}\right](\upsilon + 0.5)^{(1/\beta_1)-(1/\beta_2)}.$$

Table 1. *Empirical power* (×100%) *for testing* H_1^0 *and* H_2^0 *for data simulated with* $\lambda_k(t, \upsilon) = \lambda_k c_k(\upsilon)$, $k = 1, 2$

n_k	Test	Alternative	VE(τ) = 0	VE(τ) = 0.33				VE(τ) = 0.67			
			β_1	β_1			2-sided	β_1			2-sided
			1	1	0.5	0.25		1	0.5	0.25	
100	Cox†		5.2	65.1	65.1	65.1	61.8	99.9	99.9	99.9	99.8
(48)‡	$\hat{U}_1^1$	H_1^0	7.9	68.1	72.3	78.8	58.7	99.8	100	100	96.8
	$\hat{U}_2^1$	H_1^0	7.7	58.5	81.0	97.8	56.5	97.8	100	100	97.7
	$\hat{U}_3^1$	H_2^0	5.9	55.4	60.2	69.7	47.3	98.9	99.5	100	94.8
	$\hat{U}_4^1$	H_2^0	6.7	47.6	71.8	94.8	43.1	96.8	99.3	100	94.6
200	Cox		5.0	90.6	90.6	90.6	100	100	100	100	100
(95)‡	$\hat{U}_1^1$	H_1^0	5.0	92.7	94.3	97.2	91.5	100	100	100	100
	$\hat{U}_2^1$	H_1^0	5.3	86.0	98.4	100	88.1	100	100	100	100
	$\hat{U}_3^1$	H_2^0	7.0	87.5	90.3	94.7	84.7	100	100	100	100
	$\hat{U}_4^1$	H_2^0	5.3	81.0	95.4	100	79.4	100	100	100	100
400	Cox		5.8	99.7	99.7	99.7	100	100	100	100	100
(190)‡	$\hat{U}_1^1$	H_1^0	6.6	99.9	99.9	100	99.5	100	100	100	100
	$\hat{U}_2^1$	H_1^0	6.0	99.0	100	100	98.8	100	100	100	100
	$\hat{U}_3^1$	H_2^0	5.3	99.6	99.9	100	99.0	100	100	100	100
	$\hat{U}_4^1$	H_2^0	5.2	97.9	100	100	97.6	100	100	100	100

†Test statistic is a Wald Z-statistic based on the standard Cox model that ignores the mark.
‡Average number of subjects infected in group 2 (placebo).

Note that $\mathrm{VE}(\tau, \upsilon) = \mathrm{VE}(\tau)$ and $\mathrm{VE}^c(\tau, \upsilon) = \mathrm{VE}^c(\tau)$ if and only if $\beta_1 = \beta_2$ so that setting $\beta_2/\beta_1 = 1.0$ represents H_0. Furthermore, $\beta_2/\beta_1 > 1$ implies $\mathrm{VE}(\tau, \upsilon)$ and $\mathrm{VE}^c(\tau, \upsilon)$ decrease with υ, and the extent of departure from H_0 increases with β_2/β_1. We also consider a 2-sided alternative with $c_2(\upsilon) = 1$ and $c_1(\upsilon) = \frac{16}{3}\upsilon I\left(\upsilon < \frac{1}{2}\right) + \left(\frac{8}{3} - \frac{8}{3}\upsilon\right) I\left(\upsilon \geqslant \frac{1}{2}\right)$.

The weight process $H_n(\cdot)$ of (3.2) is used for the test statistics and $\hat{w}_V(\cdot) = 1$ for $\hat{U}_2^1$ and $\hat{U}_4^1$. For kernel estimation of $\lambda_k(t), k = 1, 2$, the Epanechnikov kernel $K(x) = 0.75(1-x^2)I(|x| \leqslant 1)$ is used. For each simulation iteration, the optimal bandwidth b_k is chosen to minimize an asymptotic approximation to the mean integrated squared error of $\hat{\lambda}_k$ (Andersen *and others*, 1993, p 240) separately for $k = 1, 2$ and the method of Gasser and Müller (1979) is used to correct for bias in the tails.

The nominal level of the tests is set at 0.05, and critical values are calculated using 500 replicates of the Gaussian multipliers technique summarized in Section 3.2. We choose $n = 100$, 200, or 400, and in addition to the 50% administrative censoring for the failure times at 36 months, we use a 10% random censoring rate in each arm. The performance statistics are calculated based on 1000 simulated data sets.

The results in Table 1 indicate that the tests of H_0^0 have appropriate sizes and high powers. When $\mathrm{VE}(t, \upsilon)$ declines with υ, they have greater power than the Cox model Wald test of $\mathrm{VE}(t) = 0$. Therefore, accounting for the mark variable can substantially improve efficiency. This is especially the case for $\hat{U}_2^1$ although this test has less power than the Cox model test if $\mathrm{VE}(t, \upsilon)$ is constant in υ (i.e. $\beta_1 = \beta_2$). In contrast, the power of $\hat{U}_1^1$ is less sensitive to how strongly $\mathrm{VE}(t, \upsilon)$ varies in υ. The corresponding 2-sided tests $\hat{U}_3^1$ and $\hat{U}_4^1$ show a similar comparative pattern but with lower power for the 1-sided alternatives.

The results in Table 2 show that the tests of H_0 perform well at moderate sample sizes. Somewhat surprisingly, for small/moderate samples, the semiparametric tests did not provide greater power than the nonparametric tests in the case that the failure times had proportional hazards. To explain this, note that the nonparametric and semiparametric test processes involve contrasts $\hat{\Lambda}_1(\mathrm{d}t, \upsilon) - \hat{r}(t)\hat{\Lambda}_2(\mathrm{d}t, \upsilon)$, with $\hat{r}(t) = \hat{\lambda}_1(t)/\hat{\lambda}_2(t)$ and $\exp(\hat{\beta})$, respectively, and the alternative hypothesis involves changes of

Table 2. *Empirical power* ($\times 100\%$) *for testing* H_1 *and* H_2 *for data simulated with* $\lambda_k(t, \upsilon) = \lambda_k c_k(\upsilon)$, $k = 1, 2$

n_k	Test	Alternative	$\mathrm{VE}(\tau) = 0.33$				$\mathrm{VE}(\tau) = 0.67$			
			β_1			2-sided	β_1			2-sided
			1	0.5	0.25		1	0.5	0.25	
100	$\hat{U}_1^{\mathrm{np}}$	H_1	6.4	21.8	59.0	42.7	7.1	17.0	35.2	22.9
(48)†	$\hat{U}_2^{\mathrm{np}}$	H_2	6.2	15.9	47.7	43.3	6.7	12.2	26.1	20.4
	$\hat{U}_1^{\mathrm{sp}}$	H_1	6.2	18.3	52.9	35.8	5.7	12.8	30.2	17.8
	$\hat{U}_2^{\mathrm{sp}}$	H_2	4.4	11.1	41.4	38.8	3.5	7.3	18.7	15.3
200	$\hat{U}_1^{\mathrm{np}}$	H_1	6.3	32.4	87.0	78.3	6.7	21.0	62.7	48.8
(95)†	$\hat{U}_2^{\mathrm{np}}$	H_2	6.8	23.0	81.4	80.9	6.5	14.3	54.2	51.4
	$\hat{U}_1^{\mathrm{sp}}$	H_1	5.6	29.7	84.8	76.8	5.5	20.0	61.1	46.3
	$\hat{U}_2^{\mathrm{sp}}$	H_2	5.4	20.8	79.5	81.4	4.8	13.2	49.6	45.6
400	$\hat{U}_1^{\mathrm{np}}$	H_1	5.8	48.2	99.5	98.3	6.2	33.7	93.3	87.4
(190)†	$\hat{U}_2^{\mathrm{np}}$	H_2	5.2	35.8	98.6	98.7	5.8	25.4	89.2	90.4
	$\hat{U}_1^{\mathrm{sp}}$	H_1	5.4	46.7	99.0	98.3	5.5	32.7	92.9	86.1
	$\hat{U}_2^{\mathrm{sp}}$	H_2	4.8	35.3	98.5	98.7	5.1	23.8	87.9	89.4

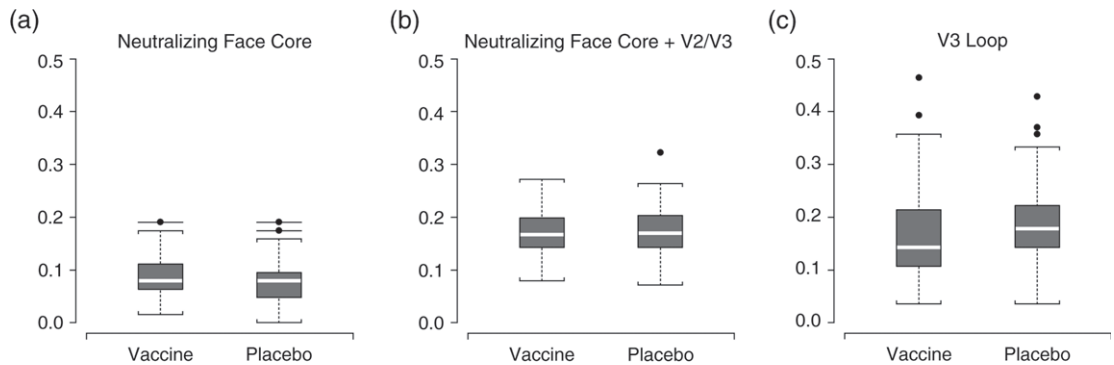
† Average number of subjects infected in group 2 (placebo).

$\lambda_1(t,v)/\lambda_2(t,v)$ in v— but not in t. Since $\hat{\Lambda}_k(\mathrm{d}t,v)$ and $\hat{\lambda}_k(t)$ approximately "track" each other in t, the nonparametric test process can reduce the noise from perturbations of $\hat{\lambda}_1(t)/\hat{\lambda}_2(t)$ in t, whereas the semiparametric test process cannot dampen this noise.

The small simulation study under non-proportional hazards, with H_0 true with $(\beta_1,\beta_2) = (1.0, 1.0)$, $(0.5, 0.5)$, or $(0.25, 0.25)$, demonstrates (as predicted from the theory) that the semiparametric tests are not valid when the marginal proportional hazards condition is not met. The empirical sizes of the tests frequently missed 0.05 by an amount more than 2 or 3 Monte Carlo standard deviations (see supplementary Table 2 available at *Biostatistics* online). Finally, the point and interval estimators of $\mathrm{VE}^{\mathrm{c}}(36,v)$ performed well, with details given in Section C of the supplementary material available at *Biostatistics* online.

6. APPLICATION

We apply the methods to the data from the VaxGen trial described in Section 1. The 32 infected subjects with a missing HIV sequence (and hence a missing mark) were excluded from the analysis. Figure 1 shows box plots of the 3 amino acid percent mismatch distances of the infecting HIV viruses to the nearest virus (MN or GNE8) represented in the tested vaccine. The testing procedures were implemented using the same weight functions $H_n(\cdot)$ and $w_V(\cdot)$, kernel $K(\cdot)$, and procedures for optimal bandwidth selection and tail correction that were used in the simulations. The p-values were approximated using 10 000 Monte Carlo simulations. The mean integrated squared error-optimized bandwidths b_k for the estimated hazards of infection $\hat{\lambda}_1(\cdot)$ and $\hat{\lambda}_2(\cdot)$ were $b_1 = 1.83$ months and $b_2 = 2.10$ months. For the



(d) P−values of Tests

Cox Wald:	0.58
$\hat{U}_1^1$ for H_0^0:	0.15
$\hat{U}_2^1$ for H_0^0:	0.05
$\hat{U}_3^1$ for H_0^0:	0.32
$\hat{U}_4^1$ for H_0^0:	0.14
$\hat{U}_1^{np}$ for H_0:	0.24
$\hat{U}_2^{np}$ for H_0:	0.38
$\hat{U}_1^{sp}$ for H_0:	0.11
$\hat{U}_2^{sp}$ for H_0:	0.2

(e) P−values of Tests

Cox Wald:	0.58
$\hat{U}_1^1$ for H_0^0:	0.18
$\hat{U}_2^1$ for H_0^0:	0.26
$\hat{U}_3^1$ for H_0^0:	0.36
$\hat{U}_4^1$ for H_0^0:	0.59
$\hat{U}_1^{np}$ for H_0:	0.84
$\hat{U}_2^{np}$ for H_0:	0.97
$\hat{U}_1^{sp}$ for H_0:	0.5
$\hat{U}_2^{sp}$ for H_0:	0.74

(f) P−values of Tests

Cox Wald:	0.58
$\hat{U}_1^1$ for H_0^0:	0.15
$\hat{U}_2^1$ for H_0^0:	0.61
$\hat{U}_3^1$ for H_0^0:	0.3
$\hat{U}_4^1$ for H_0^0:	0.72
$\hat{U}_1^{np}$ for H_0:	0.94
$\hat{U}_2^{np}$ for H_0:	0.33
$\hat{U}_1^{sp}$ for H_0:	0.73
$\hat{U}_2^{sp}$ for H_0:	0.11

Fig. 1. The top panel shows box plots of amino acid distances in HIV gp120 between the infecting viruses and the nearest vaccine strain MN or GNE8, for the 3 studied HIV distances. The bottom panel shows p-values of the studied tests.

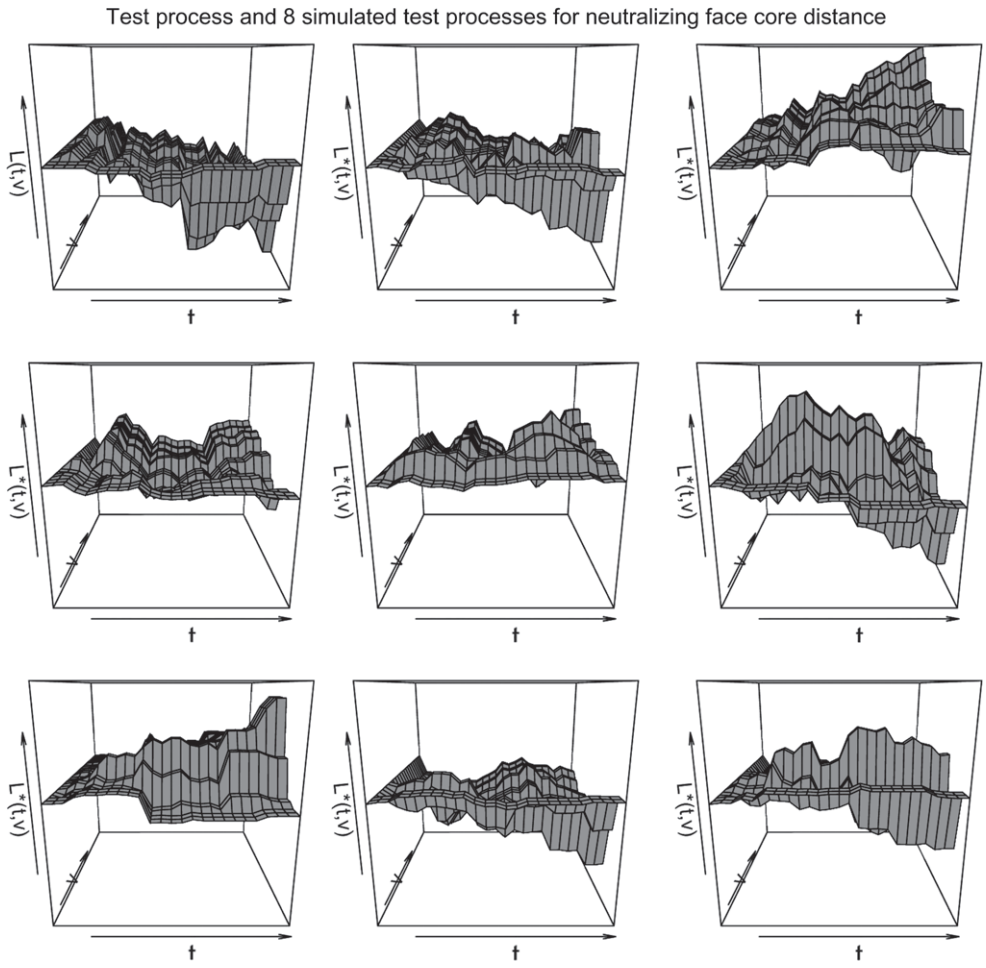


Fig. 2. For the neutralizing face core distances, the top-left panel shows the observed test process $L_n^{\mathrm{np}}(t, \upsilon)$ and the other panels show 8 randomly selected realizations of the simulated null test process $L_n^{\mathrm{np}*}(t, \upsilon)$. The value 0 of $L_n^{\mathrm{np}*}(t, \upsilon)$ is found as the value present at $t = 0$.

neutralizing face core distances, the 4 tests of H_0^0: $\mathrm{VE}(t, \upsilon) = 0$ gave p-values spanning 0.05 to 0.32 (Figure 1(d)), with $\hat{U}_2^1$ rejecting H_0^0 at the level 0.05. Based on this evidence (albeit weak) that $\mathrm{VE}(t, \upsilon) \neq 0$, we go on to test H_0: $\mathrm{VE}(t, \upsilon) = \mathrm{VE}(t)$. Neither nonparametric test rejected H_0 (Figure 1(d)). The proportional hazards assumption seemed reasonable based on a goodness-of-fit test ($p = 0.35$), justifying the semiparametric tests of H_0, which gave nonsignificant results (Figure 1(d)). To illustrate the graphical procedure, Figure 2 shows the test process $L_n^{\mathrm{np}}(t, \upsilon)$ together with 8 randomly selected realizations of the null test process $L_n^{\mathrm{np}*}(t, \upsilon)$, using a unit weight process $H_n(\cdot) = 1$. The maximum absolute deviation of $L_n^{\mathrm{np}}(t, \upsilon)$ in t is larger than that for all but one of the null test processes. Figure 1(e) and (f) shows p-values of the tests for the other 2 distances, which all exceeded 0.05.

With bandwidths $b_{\upsilon 1}$ and $b_{\upsilon 2}$ separately optimized using 2-fold cross-validation, we next estimated $\mathrm{VE}^{\mathrm{c}}(36, \upsilon)$ and $\mathrm{VE}^{\mathrm{dc}}(36, \upsilon)$ (Figure 3). The $\mathrm{VE}^{\mathrm{c}}(36, \upsilon)$ curves are estimated with reasonable precision at mark values υ not in the tail regions, and $\mathrm{VE}^{\mathrm{dc}}(36, \upsilon)$ is estimated with reasonable precision for υ not in the left tail, with precision increasing with υ. For neutralizing face core distances, the estimates of

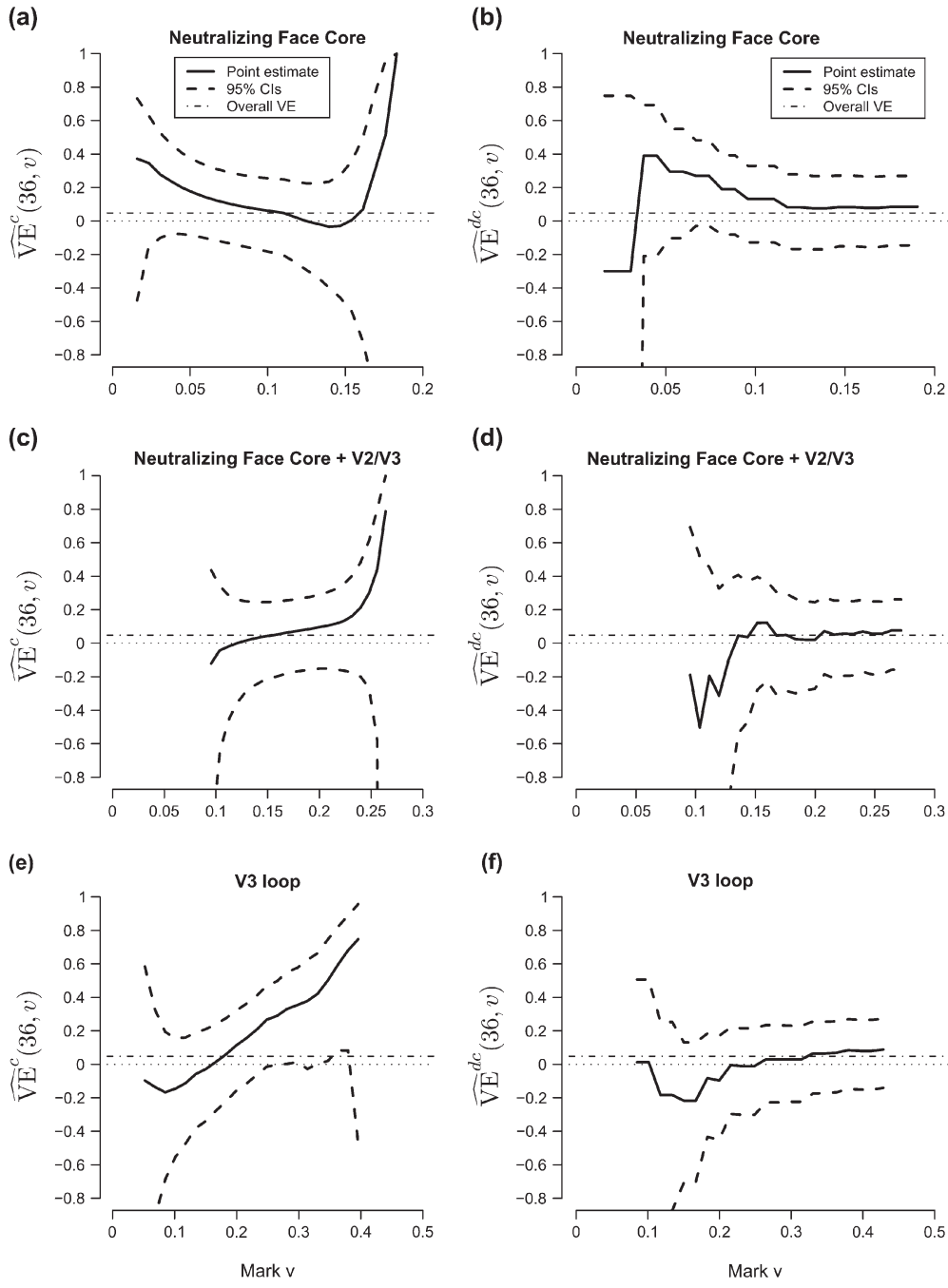


Fig. 3. The left panels show point and 95% confidence interval estimates of $\mathrm{VE}^{\mathrm{c}}(36, v) = 1 - F_1(36, v)/F_2(36, v)$ versus the HIV gp120 amino acid distance between the infecting viruses and the nearest vaccine antigen MN or GNE8, for the 3 studied HIV distances, with bandwidths $(b_{v1}, b_{v2}) = (0.096, 0.092)$, $(0.10, 0.10)$, $(0.10, 0.10)$ for (a), (b), (c). The right panels show corresponding point and interval estimates of $\mathrm{VE}^{\mathrm{dc}}(36, v) = 1 - P(T_1 \leqslant 36, V_1 \leqslant v)/P(T_2 \leqslant 36, V_2 \leqslant v)$. The dashed horizontal line is the overall vaccine efficacy estimate $\widehat{\mathrm{VE}}^{\mathrm{c}}(36) = 0.048$.

$VE^c(36, \upsilon)$ and $VE^{dc}(36, \upsilon)$ in the regions of precision diminished with viral distance, suggesting that the closeness of match of amino acids in the exposing strain versus vaccine strain in the core amino acids may have impacted the ability of the vaccine to stimulate protective antibodies that neutralized the exposing strain. This nonsignificant trend is intriguing because this distance has the soundest biological rationale—3-dimensional structural analysis has demonstrated that the amino acid positions used for this distance constitute conserved neutralizing antibody epitopes (Wyatt *and others*, 1998).

7. Concluding remarks

Nonparametric and semiparametric methods have been developed for testing and estimation of relative risks taking into account a continuous mark variable observed only at uncensored failure times and for evaluating the relationship between the relative risk and the mark. We showed that if the mark-specific relative risk varies with the mark, then a standard Cox model test of a unit hazard ratio (ignoring the mark) is less powerful (and sometimes much less) than the newly developed nonparametric procedures that test the null hypothesis $H_0^0 : \lambda_1(t, \upsilon)/\lambda_2(t, \upsilon) = 1$ of a unit mark-specific hazard ratio. This finding raises the novel idea to consider accounting for the mark variable in secondary hypothesis tests in clinical trials for which there are strong reasons to suspect that the mark-specific relative risk varies in the mark. Among the statistics developed for testing H_0^0, we recommend $\hat{U}_2^1$ if the mark-specific relative risk is thought to vary strongly with the mark, and $\hat{U}_1^1$ otherwise.

For testing dependency of the mark-specific relative risk on the mark, $H_0 : \lambda_1(t, \upsilon)/\lambda_2(t, \upsilon) = \lambda_1(t)/\lambda_2(t)$, the simulations suggest that the nonparametric procedures perform better than their semiparametric counterparts that assume proportional marginal hazards. The test based on $\hat{U}_1^{np}$ is recommended. The results also suggest that at least 100–200 failure events in the control group are needed to achieve high power to detect moderate departures from H_0. Furthermore, to achieve high power it is necessary that the trial population is highly exposed to HIV and the exposing HIVs have wide variation in mark values. As such a consideration for trial design is selecting sites with broad pathogen sequence diversity, which can increase generalizability of the trial results as well as conferring greater power for testing H_0^0 and H_0.

Although the methods were motivated by a particular scientific problem (the question in HIV vaccine efficacy trials of if and how efficacy of the tested vaccine varies with the genetic distance of the infecting HIV strain), we emphasize that they provide a general solution to the 2-sample survival analysis problem with a continuous mark variable, which may have many applications. An appeal of the procedures developed here is that they are based on a mark-specific version of the widely applied and well-understood Nelson–Aalen-type nonparametric MLE and naturally extend the scope of methods that have been developed for competing risks data with discrete (cause-of-failure) marks. Code for implementing the procedures is available upon request.

Acknowledgments

The authors gratefully acknowledge David Jobes and VaxGen Inc. for providing the HIV sequence data and thank Per Andersen and the anonymous reviewers for their helpful suggestions. Part of the research for this paper was done while Ian W. McKeague was visiting the Institute for Mathematical Sciences, National University of Singapore in 2005/2006. The visit was supported by the institute. *Conflict of Interest:* None declared.

Funding

National Institutes of Health (2 R01 AI54165-04 to P.B.G.); National Science Foundation (DMS-0505201 to I.W.M., DMS-0304922 and DMS-0604576 to Y.S.).

REFERENCES

ANDERSEN, P. K., BORGAN, O., GILL, R. D. AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

FLYNN, N. M., FORTHAL, D. N., HARRO, C. D., MAYER, K. H., PARA, M. F. AND THE RGP120 HIV VACCINE STUDY GROUP. (2005). Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *The Journal of Infectious Diseases* **191**, 654–665.

GASSER, T. AND MÜLLER, H.-G. (1979). Kernel estimation of regression functions. In: Gasser, T. and Rosenblatt, M. (editors), *Smoothing Techniques for Curve Estimation*. Lecture Notes in Mathematics, Volume 757. Berlin: Springer, pp. 23–68.

GILBERT, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *Annals of Statistics* **28**, 151–194.

GILBERT, P. B., LELE, S. AND VARDI Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika* **86**, 27–43.

GILBERT, P. B., MCKEAGUE, I. W. AND SUN, Y. (2004). Tests for comparing mark-specific hazards and cumulative incidence functions. *Lifetime Data Analysis* **10**, 5–28.

GILBERT, P. B., WEI, L. J., KOSOROK, M. R. AND CLEMENS, J. D. (2002). Simultaneous inference on the contrast of two hazard functions with censored observations. *Biometrics* **58**, 773–780.

HALLORAN, M. E., HABER, M. J. AND LONGINI, I. M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity. *American Journal of Epidemiology* **136**, 328–343.

HALLORAN, M. E., STRUCHINER, C. J. AND LONGINI, I. M. (1997). Study designs for different efficacy and effectiveness aspects of vaccination. *American Journal of Epidemiology* **146**, 789–803.

HUANG, Y. AND LOUIS, T. A. (1998). Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika* **85**, 785–798.

LIN, D. Y., WEI, L. J. AND YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V., FLOURNEY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**, 541–554.

SUN, Y. (2001). Generalized nonparametric test procedures for comparing multiple cause-specific hazard rates. *Journal of Nonparametric Statistics* **13**, 171–207.

WYATT, R., KWONG, P. D., DESJARDINS, E., SWEET, R. W., ROBINSON, J., HENDRICKSEN, W. A. AND SODROSKI, J. G. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**, 705–711.

[*Received December 1, 2006; revised May 23, 2007; accepted for publication July 9, 2007*]