Biostatistics (2005), **6**, 3, *pp.* 374–394
doi:10.1093/biostatistics/kxi014
Advance Access publication on April 14, 2005

Failure time analysis of HIV vaccine effects on viral load and antiretroviral therapy initiation

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

YANQING SUN

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

SUMMARY

The world's first efficacy trial of a preventive HIV vaccine was completed in 2003. Study participants who became HIV infected were followed for 2 years and monitored for HIV viral load and initiation of antiretroviral therapy (ART). In order to determine if vaccination may have altered HIV progression in persons who acquired HIV, a pre-specified objective was to compare the time until a composite endpoint between the vaccine and placebo arms, where the composite endpoint is the first event of ART initiation or viral failure (HIV viral load exceeds a threshold x_{vl} copies/ml). Specifically, with vaccine efficacy, $\mathrm{VE}(\tau, x_{\mathrm{vl}})$, defined as one minus the ratio (vaccine/placebo) of the cumulative probability of the composite endpoint (with failure threshold x_{vl}) occurring by τ months, the aim was to estimate the four parameters $\{\mathrm{VE}(\tau, x_{\mathrm{vl}})$: $x_{\mathrm{vl}} \in \{1500, 10\,000, 20\,000, 55\,000\}$ copies/ml$\}$ with simultaneous 95% confidence bands. A Gaussian multipliers simulation method is devised for constructing confidence bands for $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ with x_{vl} spanning multiple discrete values or a continuous range. The new method is evaluated in simulations and is applied to the vaccine trial data set.

Keywords: Gaussian multipliers technique; HIV vaccine efficacy trial; Kaplan–Meier estimator; Simultaneous confidence bands.

1. INTRODUCTION

Development of a preventive HIV vaccine (administered to HIV uninfected persons) is a global public health priority. A preventive vaccine may reduce morbidity and mortality due to HIV infection in at least three ways: (1) lower susceptibility to acquiring HIV infection, (2) decrease secondary transmission of HIV from vaccine recipients who become infected and (3) ameliorate HIV disease progression in vaccine recipients who become infected. Classically designed Phase III vaccine efficacy trials allow evaluation of vaccine effect (1), but do not allow direct evaluation of (2) and (3). This is the case because secondary transmission events are not observed, and the numbers of AIDS and death endpoints are low due to the several-year disease progression period of HIV and the ethical mandate to provide antiretroviral therapy

*To whom correspondence should be addressed.

(ART) to trial participants who acquire HIV (UNAIDS, 2001). Nonetheless, it is important to attempt to evaluate vaccine effects (2) and (3) within classically designed trials, because most licensed vaccines protect through these mechanisms (Murphy and Chanock, 1996; Clemens *et al*., 1997; Halloran *et al*., 1997; Clements-Mann, 1998), and a series of candidate vaccines are under development that are designed specifically to ameliorate transmission and disease post-acquisition of HIV (Nabel, 2001; Shiver *et al*., 2002; HVTN, 2004; IAVI, 2004).

In this article, we consider an indirect approach to assessing (2) and (3) within a classically designed efficacy trial. This objective is important because (i) classical designs are simpler and cheaper than augmented partners (Longini *et al.*, 1996) and cluster-randomized (Halloran *et al*., 1997; Hayes, 1998) designs that would permit direct assessments of (2), (ii) no design is available for assessing (3) directly within a 2–4 year time frame and (iii) the first two completed HIV vaccine efficacy trials (rgp120 HIV Vaccine Study Group, 2004) and the ongoing efficacy trial in Thailand use a classical design.

In the indirect approach, we consider methods for evaluating vaccine effects on a biomarker variable measured post-HIV infection that is putatively a surrogate endpoint for secondary transmission and/or progression to clinical disease. The level of plasma HIV RNA (viral load) is an important putative surrogate endpoint, since it has been found to be highly prognostic for both of these endpoints in observational studies (cf. Mellors *et al.*, 1997; HIV Surrogate Marker Collaborative Group, 2000; Quinn *et al.*, 2000; Gray *et al.*, 2001), and has been used as a primary endpoint in many ART trials (Gilbert *et al.*, 2001). The completed and ongoing efficacy trials use viral load measurements as the basis for assessing vaccine effects on transmission and disease.

In addition to the complication that a vaccine effect to reduce biomarker levels may not predict a vaccine effect to reduce the rate of clinical endpoints (cf. Fleming, 1992; Fleming and DeMets, 1996; Albert *et al*., 1998), the assessment of viral load is complicated by the fact that some trial participants will likely receive ART following the diagnosis of HIV infection (DHHS Guidelines, 2002). The therapy will suppress viral replication to undetectable levels in many treated persons (DHHS Guidelines, 2002). Consequently, the comparison of viral load between vaccine and placebo recipients is confounded by the effect of ART. This complication can be avoided by basing the analysis only on viral load measurements made on blood samples drawn soon after the diagnosis of HIV infection and before the initiation of ART. Though useful, this analysis provides no direct information about the durability of the vaccine effect. Initial suppression of virus by vaccine may wane over time due to emergent HIV vaccine resistance mutations; this phenomenon has been observed in monkey challenge studies that evaluated leading HIV vaccine approaches (Barouch *et al.*, 2002, 2003), and is a major potential problem for HIV vaccines in humans (Lukashov *et al.*, 2002). Therefore, it is important to analyze study endpoints that capture longer-term vaccine effects on viral load.

In the first HIV vaccine efficacy trial, VAX004, the Statistical Analysis Plan (SAP) specified the main post-infection study endpoint as a composite endpoint, defined as either virologic failure (a rise in HIV viral load above a pre-specified failure threshold x_{vl} copies/ml) or initiation of ART, whichever occurs first. This endpoint has recently been proposed for use as a co-primary endpoint (together with HIV infection) for efficacy trials of HIV vaccines designed to ameliorate viremia (Gilbert *et al.*, 2003). The composite endpoint is directly tied to clinical events, because virologic failure places a subject at increased risk for AIDS and HIV transmission to others, and initiating ART exposes a patient to drug toxicities, drug resistance and the loss of future ART options (Hirsch *et al.*, 2000; DHHS Guidelines, 2002). The composite endpoint measures the magnitude of viremic control through the choice of failure threshold x_{vl}, with a vaccine effect on the endpoint with a lower threshold indicating greater suppression. In addition, the endpoint measures the durability of the vaccine effect by counting events during a sufficiently long period following the diagnosis of HIV. Virologic failure has been used as a primary endpoint in many clinical trials of ARTs for HIV infected persons (Gilbert *et al.*, 2000, 2001).

An analytic advantage of the composite endpoint is that it can be assessed validly using standard survival analysis techniques such as Kaplan–Meier curves and log-rank tests. Such methods would yield biased inferences if applied to assess the time to virologic failure with censoring of subjects who initiate ART, because ART initiation is almost certainly associated with the risk of virologic failure, since physicians use information on viral load in decisions to prescribe ART (DHHS Guidelines, 2002).

The SAP for VAX004 specified analyzing a vaccine efficacy parameter, $\mathrm{VE}(\tau, x_{\mathrm{vl}})$, defined as one minus the ratio (vaccine/placebo) of cumulative probabilities of the composite endpoint occuring by $\tau = 12$ months post-infection diagnosis. $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ is interpreted as the percent reduction (vaccine versus placebo) in the cumulative risk of the composite endpoint by τ months. A parameter based on cumulative rather than instantaneous incidence rates was used in order to capture durability of the vaccine effect to 12 months. $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ can be estimated using Kaplan–Meier estimates of the composite endpoint survival curves for the vaccine and placebo groups. The SAP specified making inferences on $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ at the four thresholds $x_{\mathrm{vl}} = 1500, 10\,000, 20\,000, 55\,000$. These thresholds were selected based on an HIV-discordant heterosexual partners study in Uganda, which showed that persons with viral load $<$1500 copies/ml rarely transmit (Quinn *et al.*, 2000; Gray *et al.*, 2001), and on the Multicenter AIDS Cohort Study (MACS), which demonstrated that the viral thresholds 1500, 10 000, 20 000 and 55 000 discriminated the risk of progressing to AIDS within 3 years after infection (DHHS Guidelines, 2002). Furthermore, the MACS population of men who have sex with men (MSM) is similar to the VAX004 study population, which was 94.3% MSM, and the MACS data provided the basis for the recent U.S. recommendations for when to initiate ART (DHHS Guidelines, 2002).

To control the Type I error rate, the SAP specified calculation of simultaneous confidence intervals for $\mathrm{VE}(\tau, x_{\mathrm{vl}})$, $x_{\mathrm{vl}} = 1500$, 10 000, 20 000 and 55 000, with 95% joint coverage probability. To our knowledge no solution to this problem exists in the literature, and we develop a solution here. Given that these four particular thresholds are not validated as important thresholds for measuring HIV vaccine effects, and that typically scant information is available a priori to predict how low the tested vaccine may be capable of suppressing viral load, it is also important to compute simultaneous confidence bands for $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ with x_{vl} varying over a continuous range. Such bands convey a full picture of the magnitude of vaccine efficacy, for example allowing identification of the threshold (if any) at which the lower simultaneous confidence limit crosses zero. Making inferences over a pre-specified interval of thresholds also avoids the need to guess at the discrete set of most important thresholds, and prevents post hoc cheating, i.e. selective reporting of $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ estimates at the thresholds that yield the largest estimates. We develop a general procedure for constructing confidence bands that applies to both cases of x_{vl} spanning discrete levels and a continuous range. Work related to the problem addressed here includes methodology for constructing confidence bands for a functional of two survival curves (Parzen *et al.*, 1997) or of two cause-specific cumulative incidence functions (McKeague *et al.*, 2001). These procedures approximated the distribution of interest using the Gaussian multipliers technique introduced by Lin *et al.* (1993); we also apply this technique.

How to interpret the estimated $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ curve over a range of x_{vl} values? First, note that the lower the threshold x_{vl} at which there is efficacy, the more potent (and efficacious) the vaccine, as greater viral suppression predicts greater reductions in both disease progression and HIV transmissibility to others. Therefore, the lowest threshold at which the lower simultaneous confidence limit for $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ exceeds 0 indicates the greatest potency of viral suppression that the vaccine provides with high confidence. Second, inference on $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ at the threshold x_{vl} at which starting ART is recommended (and offered/provided to trial participants) has important policy implications, because the efficacy parameter at this threshold has interpretation as the percent vaccine reduction in the fraction of persons who need ART by time τ. Third, albeit with interpretation complicated by ART initiation, the shape of the estimated curve $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ reflects the mechanism by which vaccination impacts viral load. In the clearest case that trial participants adhere to the ART guidelines used in the trial, if the vaccine operates by lowering viral loads at all levels

by a constant amount (i.e. a location-shift effect), then $\mathrm{VE}(t, x_{\mathrm{vl}})$ is positive for all x_{vl}. Under other mechanisms of vaccine effects, the efficacy can vanish to zero above a certain threshold x_{vl}; for example this may occur if vaccination only impacts viral loads below a certain level.

This article is organized as follows. The procedure for generating simultaneous confidence bands is developed in Section 2, and is studied in simulations in Section 3. Section 4 applies the methods to the VAX004 data. Section 5 discusses alternative and complementary approaches to studying the composite endpoint. Section 6 provides discussion on how to apply the new method in future vaccine trials, and an Appendix contains theoretical details of the method.

2. Method for constructing simultaneous confidence bands

2.1 *Preliminaries and the estimand*

With τ a fixed time point and x_{vl} a fixed virologic failure threshold, define

$$\mathrm{VE}(\tau, x_{\mathrm{vl}}) = 1 - F_1(\tau, x_{\mathrm{vl}})/F_2(\tau, x_{\mathrm{vl}}),$$

where $F_1(\tau, x_{\mathrm{vl}})(F_2(\tau, x_{\mathrm{vl}}))$ is the cumulative probability that a vaccinated (placebo) subject fails virologically or starts treatment by τ months post-infection diagnosis. Let $T_{k1}, \ldots, T_{kn_k}$ be the times between infection diagnosis and treatment initiation and $Y_{k1}(t), \ldots, Y_{kn_k}(t)$ be the viral loads at time t for the n_k infected subjects in group k ($k = 1$, vaccine; $k = 2$, placebo). Assume that $\{Y_{ki}(t), T_{ki}\}$, $i = 1, \ldots, n_k$, are independent, identically distributed (iid) within each group, and the two samples are independent of one another. We also assume that $F_k(t, x_{\mathrm{vl}})$ is continuous on $[0, \tau] \times [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$ with $F_k(\tau, x_{\mathrm{vl}}^{\mathrm{L}}) < 1$ for $k = 1, 2$. The total number of infected subjects is $n = n_1 + n_2$. Let $\rho_k = \lim_{n\to\infty} n_k/n$ and $0 < \rho_k < 1$. The goal is to construct simultaneous confidence bands for $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ for x_{vl} spanning a pre-specified range $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$, where $x_{\mathrm{vl}}^{\mathrm{L}} < x_{\mathrm{vl}}^{\mathrm{U}}$ and $F_2(\tau, x_{\mathrm{vl}}^{\mathrm{U}}) > 0$. The widest possible range of thresholds is specified by $x_{\mathrm{vl}}^{\mathrm{L}}$ and $x_{\mathrm{vl}}^{\mathrm{U}}$ equal to the lower- and upper-quantification limits of the viral load assay, respectively.

The time for subject i in group k to fail virologically given the virologic failure threshold x_{vl} or starting treatment, whichever comes first, is $\widetilde{T}_{ki}(x_{\mathrm{vl}}) = \min\{\inf\{t : \sup_{0\leqslant s\leqslant t} Y_{ki}(s) \geqslant x_{\mathrm{vl}}\}, T_{ki}\}$. Let C_{ki} be the censoring time for subject i in group k, $\widetilde{X}_{ki}(x_{\mathrm{vl}}) = \min\{\widetilde{T}_{ki}(x_{\mathrm{vl}}), C_{ki}\}$, and $\delta_{ki}(x_{\mathrm{vl}}) = I(\widetilde{T}_{ki}(x_{\mathrm{vl}}) \leqslant C_{ki})$. We assume $\widetilde{T}_{ki}(x_{\mathrm{vl}})$ and C_{ki} are independent for each k.

Throughout this article we define the time of virologic failure as the time of the first study visit at which the viral load is observed to equal or exceed x_{vl}. Alternatively, this event time could be taken to be the true time at which viral load first exceeds x_{vl}. This event time is interval censored, and the estimation of $\mathrm{VE}(t, x_{\mathrm{vl}})$ could be biased if interval censoring is ignored. We restrict attention to the observable viral failure detection time because (i) it is clinically relevant to define failure at the clinic visit of failure detection, because this is the event observed by physicians that influences treatment decisions; (ii) the time of ART initiation is defined by the clinic visit at which ART is prescribed, so that using the clinic visit time for viral failure creates a cohesive definition of the composite endpoint event time and (iii) there is greatest interest in assessing $\mathrm{VE}(t, x_{\mathrm{vl}})$ at the latest time point $t = \tau$, and inferences on $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ are minimally susceptible to bias from interval censoring, since interval censoring up to the last visit time prior to τ does not impact estimates of the proportion failing by τ.

For fixed x_{vl}, the cumulative probability that an infected subject in group k fails virologically or starts treatment by time τ is equal to

$$F_k(\tau, x_{\mathrm{vl}}) = P\{\widetilde{T}_{ki}(x_{\mathrm{vl}}) \leqslant \tau\} = 1 - P\left\{\sup_{0\leqslant s\leqslant\tau} Y_k(s) < x_{\mathrm{vl}}, T_k > \tau\right\}.$$

2.2 *Estimation*

Let $S_k(\tau, x_{\mathrm{vl}}) = 1 - F_k(\tau, x_{\mathrm{vl}})$ be the survival function of $\widetilde{T}_{ki}(x_{\mathrm{vl}})$ at time τ and let $\widehat{S}_k(\tau, x_{\mathrm{vl}})$ be the Kaplan–Meier estimator of $S_k(\tau, x_{\mathrm{vl}})$ based on $\{\widetilde{X}_{ki}(x_{\mathrm{vl}}), \delta_{ki}(x_{\mathrm{vl}})\}$ for $i = 1, \ldots, n_k$. Then $\widehat{F}_k(\tau, x_{\mathrm{vl}}) = 1 - \widehat{S}_k(\tau, x_{\mathrm{vl}})$. Let $\widehat{\Lambda}_k(\tau, x_{\mathrm{vl}})$ be the Nelson–Aalen estimator for the cumulative hazard function $\Lambda_k(\tau, x_{\mathrm{vl}}) = -\log S_k(\tau, x_{\mathrm{vl}})$. For explicit forms of these estimators, we introduce the following notations. Let $N_{ki}(t, x_{\mathrm{vl}}) = I(\widetilde{X}_{ki}(x_{\mathrm{vl}}) \leqslant t, \delta_{ki}(x_{\mathrm{vl}}) = 1)$, $R_{ki}(t, x_{\mathrm{vl}}) = I(\widetilde{X}_{ki}(x_{\mathrm{vl}}) \geqslant t)$, $M_{ki}(t, x_{\mathrm{vl}}) = N_{ki}(t, x_{\mathrm{vl}}) - \int_0^t R_{ki}(s, x_{\mathrm{vl}})\, \mathrm{d}\Lambda_k(s, x_{\mathrm{vl}})$ and $R_k(t, x_{\mathrm{vl}}) = \sum_{i=1}^{n_k} R_{ki}(t, x_{\mathrm{vl}})$. Let $r_k(t, x_{\mathrm{vl}}) = P\{\widetilde{X}_{ki}(x_{\mathrm{vl}}) \geqslant t\}$. The Nelson–Aalen estimator for the given x_{vl} is then

$$\widehat{\Lambda}_k(t, x_{\mathrm{vl}}) = \sum_{i=1}^{n_k} \int_0^t \frac{\mathrm{d}N_{ki}(s, x_{\mathrm{vl}})}{R_k(s, x_{\mathrm{vl}})}.$$

It is well known that for the given value of x_{vl}, we have the following martingale representation for the Kaplan–Meier estimator (Fleming and Harrington, 1991):

$$\sqrt{n_k}(\widehat{F}_k(\tau, x_{\mathrm{vl}}) - F_k(\tau, x_{\mathrm{vl}})) = S_k(\tau, x_{\mathrm{vl}}) \int_0^\tau \frac{\sqrt{n_k} \sum_{i=1}^{n_k} \mathrm{d}M_{ki}(s, x_{\mathrm{vl}})}{r_k(s, x_{\mathrm{vl}})} + \mathrm{o}_p(1). \tag{2.1}$$

It is shown in the Appendix that (2.1) holds uniformly for $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$ and that (2.1) converges in distribution to a mean-zero normal random variable with variance equal to $\sigma_k^2(\tau, x_{\mathrm{vl}}) = S_k^2(\tau, x_{\mathrm{vl}}) \int_0^\tau \mathrm{d}\Lambda_k(s, x_{\mathrm{vl}})/r_k(s, x_{\mathrm{vl}})$. In the absence of censoring, $\sigma_k^2(\tau, x_{\mathrm{vl}})$ reduces to $F_k(\tau, x_{\mathrm{vl}})\, S_k(\tau, x_{\mathrm{vl}})$. The asymptotic variance $\sigma_k^2(\tau, x_{\mathrm{vl}})$ can be consistently estimated by $\widehat{\sigma}_k^2(\tau, x_{\mathrm{vl}}) = n_k \widehat{S}_k^2(\tau, x_{\mathrm{vl}}) \int_0^\tau \mathrm{d}\widehat{\Lambda}_k(s, x_{\mathrm{vl}})/R_k(s, x_{\mathrm{vl}})$.

For ease of notation, in what follows, we drop the first component τ in the functions. Then

$$\begin{aligned} U(x_{\mathrm{vl}}) &= \sqrt{n}\left(\frac{\widehat{F}_1(x_{\mathrm{vl}})}{\widehat{F}_2(x_{\mathrm{vl}})} - \frac{F_1(x_{\mathrm{vl}})}{F_2(x_{\mathrm{vl}})}\right) \\ &= \sqrt{n}\left(\frac{1}{F_2(x_{\mathrm{vl}})}(\widehat{F}_1(x_{\mathrm{vl}}) - F_1(x_{\mathrm{vl}})) - \frac{F_1(x_{\mathrm{vl}})}{(F_2(x_{\mathrm{vl}}))^2}(\widehat{F}_2(x_{\mathrm{vl}}) - F_2(x_{\mathrm{vl}}))\right) + \mathrm{o}_p(1), \end{aligned} \tag{2.2}$$

uniformly in $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$.

2.3 *Pointwise confidence bands for* $\mathrm{VE}(x_{\mathrm{vl}})$

It follows from the central limit theorem that for each fixed x_{vl}, $U(x_{\mathrm{vl}})$ converges in distribution to a mean-zero normal random variable with variance

$$\sigma^2(x_{\mathrm{vl}}) = \rho_1^{-1}(F_2(x_{\mathrm{vl}}))^{-2}\sigma_1^2(x_{\mathrm{vl}}) + \rho_2^{-1}(F_1(x_{\mathrm{vl}}))^2(F_2(x_{\mathrm{vl}}))^{-4}\sigma_2^2(x_{\mathrm{vl}}),$$

which can be estimated by $\widehat{\sigma}^2(x_{\mathrm{vl}})$ obtained by replacing ρ_k with n_k/n, $F_k(x_{\mathrm{vl}})$ with $\widehat{F}_k(x_{\mathrm{vl}})$ and $\sigma_k^2(x_{\mathrm{vl}})$ with $\widehat{\sigma}_k^2(x_{\mathrm{vl}})$. Let $\widehat{\mathrm{VE}}(x_{\mathrm{vl}}) = 1 - \widehat{F}_1(x_{\mathrm{vl}})/\widehat{F}_2(x_{\mathrm{vl}})$. Large sample $100(1-\alpha)\%$ pointwise confidence bands for $\mathrm{VE}(x_{\mathrm{vl}})$ at x_{vl} are given by

$$\widehat{\mathrm{VE}}(x_{\mathrm{vl}}) \pm n^{-1/2} z_{\alpha/2} \widehat{\sigma}(x_{\mathrm{vl}}), \tag{2.3}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution.

2.4 *Simultaneous confidence bands for* $\mathrm{VE}(x_{\mathrm{vl}})$

From (2.1) and (2.2), we have

$$U(x_{\mathrm{vl}}) = (n/n_1)^{1/2}(F_2(x_{\mathrm{vl}}))^{-1}S_1(x_{\mathrm{vl}})\int_0^\tau \frac{n_1^{-1/2}\sum_{i=1}^{n_1} \mathrm{d}M_{1i}(s, x_{\mathrm{vl}})}{r_1(s, x_{\mathrm{vl}})}$$
$$- (n/n_2)^{1/2}\frac{F_1(x_{\mathrm{vl}})}{(F_2(x_{\mathrm{vl}}))^2}S_2(x_{\mathrm{vl}})\int_0^\tau \frac{n_2^{-1/2}\sum_{i=1}^{n_2} \mathrm{d}M_{2i}(s, x_{\mathrm{vl}})}{r_2(s, x_{\mathrm{vl}})} + \mathrm{o}_p(1). \tag{2.4}$$

Let $Z_{1i}, Z_{2j}, i = 1, \ldots, n_1, j = 1, \ldots, n_2$, be iid standard normal random variables. Let

$$U^*(x_{\mathrm{vl}}) = (n/n_1)^{1/2}(\widehat{F}_2(x_{\mathrm{vl}}))^{-1}\widehat{S}_1(x_{\mathrm{vl}})\int_0^\tau \frac{n_1^{-1/2}\sum_{i=1}^{n_1} Z_{1i}\,\mathrm{d}\widehat{M}_{1i}(s, x_{\mathrm{vl}})}{R_1(s, x_{\mathrm{vl}})}$$
$$- (n/n_2)^{1/2}\frac{\widehat{F}_1(x_{\mathrm{vl}})}{(\widehat{F}_2(x_{\mathrm{vl}}))^2}\widehat{S}_2(x_{\mathrm{vl}})\int_0^\tau \frac{n_2^{-1/2}\sum_{i=1}^{n_2} Z_{2i}\,\mathrm{d}\widehat{M}_{2i}(s, x_{\mathrm{vl}})}{R_2(s, x_{\mathrm{vl}})}, \tag{2.5}$$

where $\widehat{M}_{ki}(t, x_{\mathrm{vl}}) = N_{ki}(t, x_{\mathrm{vl}}) - \int_0^t R_{ki}(s, x_{\mathrm{vl}})\,\mathrm{d}\widehat{\Lambda}_k(s, x_{\mathrm{vl}})$. It is shown in the Appendix that $U(x_{\mathrm{vl}})$ converges weakly to a mean-zero Gaussian process for $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$ and that conditional on the observed data, the process $U^*(x_{\mathrm{vl}})$ converges weakly to the same limiting Gaussian process as $U(x_{\mathrm{vl}})$. Also, by the uniform almost sure convergence of $\widehat{\sigma}(x_{\mathrm{vl}})$ to $\sigma(x_{\mathrm{vl}})$ over $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$, it follows that

$$\lim_{n\to\infty} P^*\left\{\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} |U^*(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})| \leqslant x\right\} \overset{\text{a.s.}}{=} \lim_{n\to\infty} P\left\{\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} |U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})| \leqslant x\right\}, \tag{2.6}$$

where $P^*\{A\}$ is the conditional probability of A given the observed data sequence. Let $c_{\alpha/2}$ be the asymptotic $1-\alpha$ quantile of $\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} |U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|$. Let $U_b^*(x_{\mathrm{vl}}), b = 1, \ldots, B$, be B independent copies of $U^*(x_{\mathrm{vl}})$, obtained by repeatedly generating independent sets of iid standard normal random variables $\{Z_{1i}, Z_{2j}, i = 1, \ldots, n_1, j = 1, \ldots, n_2\}$ while holding the observed data fixed. The quantile $c_{\alpha/2}$ can be estimated consistently by the $1-\alpha$ quantile of the set $\{\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} |U_b^*(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|, b = 1, \ldots, B\}$. Large sample $100(1-\alpha)\%$ uniform confidence bands for $\mathrm{VE}(x_{\mathrm{vl}})$ over $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$ are then given by

$$\widehat{\mathrm{VE}}(x_{\mathrm{vl}}) \pm n^{-1/2}c_{\alpha/2}\widehat{\sigma}(x_{\mathrm{vl}}). \tag{2.7}$$

3. Simulations

A complicated question is how to simulate viral loads and the times to treatment initiation in the most realistic way. The time to treatment initiation depends heavily on the current science on when to start ART and on the policy that is used to provide treatment for infected trial participants; these factors vary over time and with the geographic region of the trial. Current science suggests that individuals with high viral load and/or low CD4 cell counts should start treatment. In particular, U.S. guidelines recommend starting treatment when viral load $>$ 55 000 copies/ml or when CD4 $<$ 350 copies/ml (DHHS Guidelines, 2002). For trials in developed countries, considerable heterogeneity in treatment initiation among infected individuals is expected; some will follow the guidelines and others will start treatment apart from the guidelines. In contrast, trials in developing countries are expected to operate under strict standardized guidelines that are adhered to by most or all infected participants.

3.1 *Simulation model setup*

We develop a simulation model based on the viral load and treatment initiation data from the VAX004 trial:

1. $n = 347$ infected subjects, $n_1 = 225$ in group 1 (vaccine) and $n_2 = 122$ in group 2 (placebo).
2. Subjects are followed for 24 months after the diagnosis of HIV infection.
3. 20% random dropout prior to the composite endpoint by 24 months for each group.
4. Viral loads are measured from samples drawn at times near nine scheduled visits at Months 0.5, 1, 2, 4, 8, 12, 16, 20 and 24 post-infection diagnosis, denoted by t_j, $1 \leqslant j \leqslant 9$. The actual visit times in months for each individual are normally distributed with means at the scheduled times. Specifically, for the ith individual in group k, the jth visit time t_{kij} is $N(t_j, \sigma_j^2)$, where $\sigma_1 = 0.05$, $\sigma_2 = 0.06$, $\sigma_3 = 0.10$ and $\sigma_j = 0.12$ for $j = 4, \ldots, 9$.
5. The viral loads ($\log_{10}$ transformed) from a subject in the placebo group satisfy a standard linear mixed effects (lme) model,

$$Y_{2i}(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 t^4 + r_{0i} + r_{1i} t + \epsilon_i(t), \tag{3.1}$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (4.3884, -0.2808, 0.0363, -0.0019, 0.000035)^\top$ are fixed effects parameters. The random effects $(r_{0i}, r_{1i})^\top$ have a bivariate normal distribution with mean 0 and covariance matrix given by $\mathrm{Var}(r_{0i}) = 0.4745$, $\mathrm{Var}(r_{1i}) = 0.00233$ and $\mathrm{Cov}(r_{0i}, r_{1i}) = -0.0138$. The measurement errors $\epsilon_i(t_{kij})$ are iid with mean 0 and variance 0.4977.

The viral load processes for the vaccine group are simulated in three ways:

(a) null model (denoted by NULL) where the viral load processes follow (3.1);

(b) constant mean shift model (denoted by CONS) with a mean shift of s_{vl} at all 9 time points, lower in vaccine than placebo. We take $s_{\mathrm{vl}} = 0.33$ and 0.5 on the $\log_{10}$ scale;

(c) non-constant mean shift model (denoted by NCONS) with a mean shift of s_{vl} lower at Months 0.5, 1 and 2, mean $0.5 s_{\mathrm{vl}}$ lower at Month 4 and 0 lower at Months 8, 12, 16, 20, 24. For this scenario the vaccine initially lowers viral load, but then vaccine resistance develops, which ruins the suppression.

Once the simulation process for viral load is set, the time to treatment initiation is generated in one of two ways: (i) (INDEP) independent of viral load and CD4 cell count and (ii) (DEP) dependent on viral load and CD4 cell count.

(i) *INDEP of biomarkers.* The times to treatment initiation are simulated from exponential distributions in each group with approximate probability of starting treatment by 24 months, 0.5 in the placebo group and 0.5 (null case) or 0.25 (alternative cases) in the vaccine group.

(ii) *DEP on biomarkers.* Based on the U.S. treatment guidelines (DHHS Guidelines, 2002), subjects whose CD4 counts decline to low levels (<350 cells/mm^3) have a high chance of starting ART, subjects whose CD4 counts decline to moderate levels (<500 cells/mm^3) have a moderate chance of starting ART, subjects whose viral load becomes high (>55 000 copies/ml) have a moderate chance of starting treatment and subjects whose CD4 stays above 500 cells/mm^3 and viral load stays below 55 000 copies/ml have a low chance of starting treatment. These ideas can be formalized by first simulating a CD4 process for each subject. Fitting a simple lme model to the real CD4 count data from VAX004 yields the following setup. There are two fixed effects parameters, the intercept $\beta_0 = 627.9$ and slope $\beta_1 = -0.203$. There are two random effects that represent subject-specific intercepts (b_0) and slopes (b_1), which have a bivariate normal distribution with mean 0 and $\mathrm{Var}(b_0) = 41375.0$, $\mathrm{Var}(b_1) = 102.9$ and $\mathrm{Cov}(b_0, b_1) = -635.6$. The Gaussian error ϵ has mean 0 and variance 15724.9.

For simulations with a vaccine effect to lower viral load by mean 0.33 (0.5), we assume a vaccine effect to increase the mean CD4 count by 100 (150) cells/mm^3. Simulation configurations with no vaccine effect on viral load also have no vaccine effect on CD4 cell count.

At each visit time, the probability of starting ART within the next month (for visits at Months 0.5, 1, 2) and within the next 2 months (for visits at Months 4, 8, 12, 16, 20) is set as a function of the current CD4 count and viral load. Specifically, the probabilities of ART initiation at a visit during the next 1 or 2 month interval are fixed as follows:

CD4 count	Viral load	Probability of ART initiation
CD4 $\leqslant$ 350	VL $>$ 55 000	0.7
CD4 $\leqslant$ 350	VL $\leqslant$ 55 000	0.3
350 $<$ CD4 $\leqslant$ 500	VL $>$ 55 000	0.1
350 $<$ CD4 $\leqslant$ 500	VL $\leqslant$ 55 000	0.05
CD4 $>$ 500	VL $>$ 55 000	0.02
CD4 $>$ 500	VL $\leqslant$ 55 000	0.01

Under this scenario, about 40% in the placebo group and 25–40% in the vaccine group start treatment by 24 months.

For a single data set randomly generated under each scenario defined by the INDEP and DEP models of ART initiation crossed with the NULL, CONS(2) and NCONS(2) models of viral load ((2) denotes a mean shift of $s_{\mathrm{vl}} = 0.5$), Figure 1 illustrates 95% confidence bands for $\mathrm{VE}(14, x_{\mathrm{vl}})$ for $x_{\mathrm{vl}} \in [1500, 55\,000]$.

3.2 *Coverage probability and empirical power*

To evaluate the coverage probability of the confidence bands and the ability of the bands to identify nonzero vaccine efficacy, we consider testing the following hypotheses:

$$H_{0i}\colon \mathrm{VE}(14, x_{\mathrm{vl}}) = 0 \ \text{ for all } x_{\mathrm{vl}} \in R_i \ \text{ versus } \ H_{ai}\colon \mathrm{VE}(14, x_{\mathrm{vl}}) \neq 0 \quad \text{for some } x_{\mathrm{vl}} \in R_i,$$

where R_i, $i = 1, \ldots, 8$, represent the following ranges of x_{vl}: $R_1, x_{\mathrm{vl}} \in [1500, 55\,000]$; $R_2, x_{\mathrm{vl}} \in [10\,000, 55\,000]$; $R_3, x_{\mathrm{vl}} \in \{1500, 10\,000, 20\,000, 55\,000\}$; $R_4, x_{\mathrm{vl}} \in \{10\,000, 55\,000\}$; $R_5, x_{\mathrm{vl}} = 1500$; $R_6, x_{\mathrm{vl}} = 10\,000$; $R_7, x_{\mathrm{vl}} = 20\,000$; $R_8, x_{\mathrm{vl}} = 55\,000$. Since the null hypothesis H_{0i} is rejected if and only if the confidence bands for $x_{\mathrm{vl}} \in R_i$ exclude zero at one or more thresholds x_{vl}, assessing these eight scenarios informs on the coverage probability of the confidence bands. In addition, evaluating these scenarios informs the power/precision trade-offs for various ways of conducting the analysis. When designing the real trial we struggled with the question of what was the best range of thresholds to study.

We propose two types of test statistics for testing H_{0i} versus H_{ai}. Specifically,

$$S_i = \sup_{x_{\mathrm{vl}} \in R_i} |U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|$$

and

$$Q_i = \sum_{x_{\mathrm{vl}} \in R_i} |U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|^2 \quad \text{or} \quad Q_i = \int_{x_{\mathrm{vl}} \in R_i} |U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|^2 \, \mathrm{d}x_{\mathrm{vl}},$$

depending on whether R_i is a finite set or a continuous interval. The null hypothesis is rejected for large values of the test statistics. The supremum tests S_i are known to be omnibus but may have lower

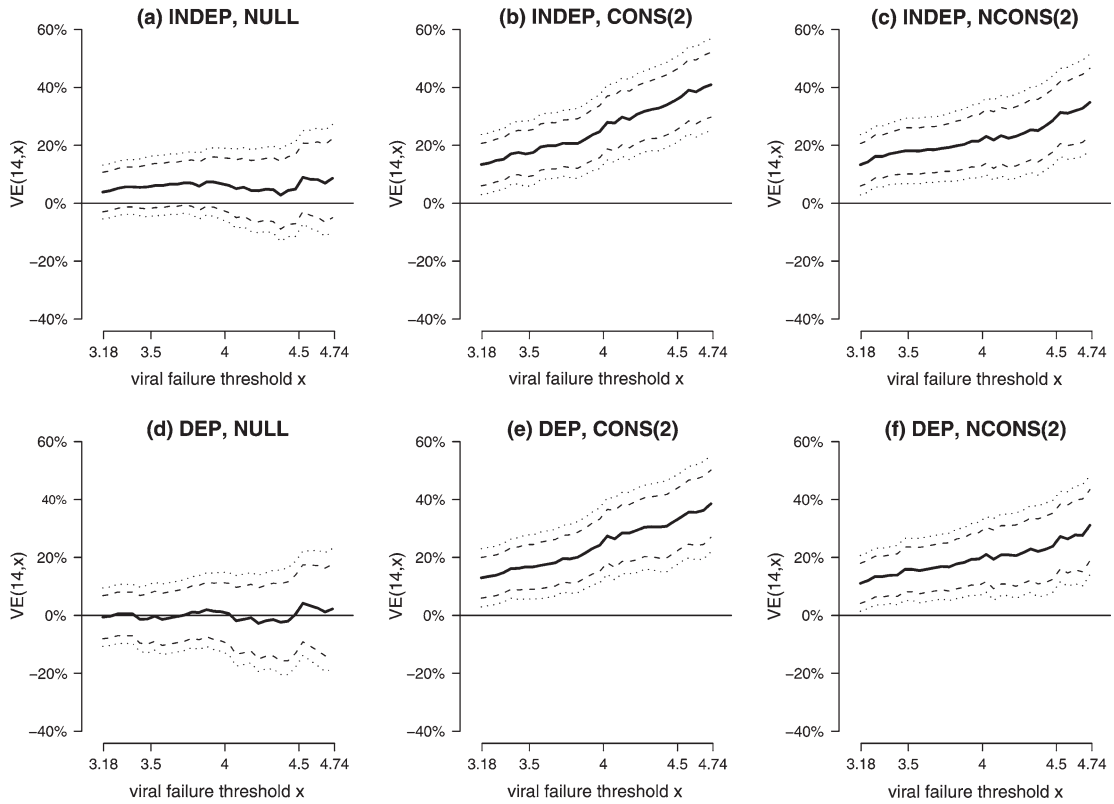

Fig. 1. For a single data set randomly generated under the scenarios (a) INDEP, NULL; (b) INDEP, CONS(2); (c) INDEP, NCONS(2); (d) DEP, NULL; (e) DEP, CONS(2) and (f) DEP, NCONS(2) described in Section 3.1, the plots show the estimate of $\mathrm{VE}(14, x_{\mathrm{vl}})$ (solid lines) with 95% pointwise (dotted lines) and simultaneous (dashed lines) confidence bands, for $x_{\mathrm{vl}} \in [1500, 55\,000]$ on the $\log_{10}$ scale.

power because of lack of specificity to specific alternatives. The sum/integrated square tests Q_i combine information across thresholds and are more powerful against monotone alternatives where the vaccine always improves over the placebo.

For the pointwise tests corresponding to $i = 5, 6, 7, 8$, the tests S_i and Q_i at significance level α are equivalent to a normal test with the test statistic $Z = U(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})$ and rejection region $|Z| > z_{\alpha/2}$. For simultaneous tests corresponding to $i = 1, 2, 3, 4$, the critical values $c_{i\,\alpha/2}$ for S_i are estimated by the $1 - \alpha$ quantile of the data set $\{\sup_{x_{\mathrm{vl}} \in R_i} |U_b^*(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|, b = 1, \ldots, B\}$. The critical values for Q_i are estimated by the $1 - \alpha$ quantile of the data set $\{\sum_{x_{\mathrm{vl}} \in R_i} |U_b^*(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|^2, b = 1, \ldots, B\}$ or $\{\int_{x_{\mathrm{vl}} \in R_i} |U_b^*(x_{\mathrm{vl}})/\widehat{\sigma}(x_{\mathrm{vl}})|^2 \, \mathrm{d}x_{\mathrm{vl}}, b = 1, \ldots, B\}$, depending on whether R_i is a finite set or continuous interval. When R_i is discrete with m thresholds, a computationally simple alternative to the above Gaussian multiplier approach is to use m normal statistics Z (one for each x_{vl} in R_i) and to apply the Bonferroni correction to determine significance. Such a procedure is likely to be conservative, resulting in wider intervals and reduced power, especially when R_i contains a large number of threshold values. Table 1 describes the empirical sizes and powers of the supremum tests and the sum/integrated square tests using Gaussian multiplier critical values, and Table 2 shows the results for the normal tests, with Bonferroni correction when $m > 1$. Each entry in Tables 1 and 2 are calculated based on 1000 repetitions and $B = 1000$.

Table 1. *Empirical sizes and powers × 100 of the supremum tests S_i and sum/integrated square tests Q_i: the models under which the data were simulated [NULL, CONS(1), CONS(2), NCONS(1) and NCONS(2)] are defined in Section 3.1. $s_{vl} = 0.33$ in models CONS(1) and NCONS(1) and $s_{vl} = 0.5$ in models CONS(2) and NCONS(2)*

Test		Supremum S_i				Sum/integrated square Q_i			
Model		R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
INDEP	NULL	7.3	6.2	6.6	7.0	8.6	7.5	6.7	7.0
	CONS(1)	91.2	91.5	93.3	92.7	91.0	93.4	94.0	94.1
	CONS(2)	99.4	99.4	99.9	99.5	99.3	99.6	99.9	99.6
	NCONS(1)	82.9	81.7	80.2	80.4	84.1	84.1	82.0	80.7
	NCONS(2)	95.3	94.7	94.0	93.6	97.0	96.9	95.8	94.7
DEP	NULL	5.5	4.4	5.6	4.8	5.4	6.0	4.5	5.1
	CONS(1)	84.2	85.4	87.4	87.1	85.0	87.4	89.2	88.0
	CONS(2)	99.3	99.4	99.6	99.6	99.2	99.6	99.9	99.8
	NCONS(1)	72.7	68.8	68.4	66.4	75.9	72.8	70.8	68.2
	NCONS(2)	95.0	95.0	94.1	92.3	96.3	96.4	94.6	93.7

Table 2. *Empirical sizes and powers × 100 of the normal tests, with Bonferroni correction for R_3 and R_4: the models under which the data were simulated [NULL, CONS(1), CONS(2), NCONS(1) and NCONS(2)] are defined in Section 3.1. $s_{vl} = 0.33$ in models CONS(1) and NCONS(1) and $s_{vl} = 0.5$ in models CONS(2) and NCONS(2)*

Test		Normal test Z				Normal test Z with Bonferroni correction	
Model		R_5	R_6	R_7	R_8	R_3	R_4
INDEP	NULL	0.8	5.0	5.7	5.9	4.3	6.1
	CONS(1)	38.5	77.8	86.4	91.7	90.6	91.9
	CONS(2)	69.9	97.0	98.6	99.3	99.8	99.5
	NCONS(1)	25.3	59.2	67.7	77.4	75.8	78.1
	NCONS(2)	47.2	82.0	87.2	92.0	91.6	92.7
DEP	NULL	1.5	4.5	5.5	4.1	3.2	4.1
	CONS(1)	37.1	71.3	79.3	84.9	82.7	85.0
	CONS(2)	71.7	97.4	98.7	99.3	99.3	99.4
	NCONS(1)	21.6	49.8	58.3	63.6	63.6	63.9
	NCONS(2)	46.3	81.1	88.0	89.8	91.9	91.8

Based on the NULL simulations, the confidence band procedures consistently have sizes near the nominal 0.05 level. An exception is for the null hypothesis R_5 in Table 2, for which the empirical size is 0.008 and 0.015 for the INDEP and DEP cases, respectively. The low size occurs because when $x_{vl} = 1500$, almost every subject fails by $\tau = 14$ months, so that the risk sets $R_1(t, x_{vl})$ and $R_2(t, x_{vl})$ in formulas (2.4) and (2.5) are very small near τ. The tiny risk sets cause the asymptotic approximation to be unreliable.

Based on the non-null simulations, the following observations were made regarding the comparative power for evaluating $\mathrm{VE}(14, x_{vl})$ in the ranges $R_1, \ldots, R_8$. First, the sum/integrated square test has slightly higher power than the supremum test for thresholds in R_1, R_2, R_3 or R_4. For R_3 and R_4, both tests

show greater power than the normal tests with Bonferroni correction. Second, power is comparable for R_1 through R_4 under each test; therefore, in practice fixed thresholds can be added without appreciably compromising power. Third, for hypothesis tests at single threshold values R_5 through R_8, the power increases with the magnitude of the threshold. Along the lines described above, this result occurs because almost all subjects fail by time τ when the threshold x_{vl} is relatively small. Power was consistently lower for the DEP versus INDEP simulations, which occurs because the alternative hypothesis is closer to the null hypothesis for the DEP simulations. Finally, as expected, power was consistently higher for the simulations with true viral load mean shift of $s_{\text{vl}} = 0.5$ compared to $s_{\text{vl}} = 0.33$.

4. EXAMPLE

The world's first HIV vaccine efficacy trial (VAX004) was conducted in North America and the Netherlands from 1998 to 2003 (rgp120 HIV Vaccine Study Group, 2004). Participants were randomized in a 2:1 ratio to receive the subunit protein vaccine AIDSVAX (3598 subjects) or a blinded placebo (1805 subjects). Participants were immunized at Months 0, 1, 6, 12, 18, 24 and 30 post-randomization, and were tested for HIV infection at Months 6, 12, 18, 24, 30 and 36. Subjects diagnosed with HIV infection were re-consented and followed on a Month 0.5, 1, 2, 4, 8, 16, 20 and 24 post-infection diagnosis visit schedule. At each of these visits, HIV viral load and status of ART initiation were recorded. The comprehensive results of the analyses of the data in VAX004 will be presented in clinical journals (including rgp120 HIV Vaccine Study Group, 2004); here we present a subset of the results needed to demonstrate and apply the statistical methodology developed here.

The primary objective of the trial was to assess whether vaccination reduced the rate of HIV infection. Unfortunately it did not, as 7% of participants were infected in each study arm (vaccine: 241/3598 infected; placebo: 127/1805 infected). The secondary objective, of interest for this article, was to assess whether vaccination altered the course of HIV progression. Of the 368 infected subjects, 347 enrolled into the post-infection cohort and are analyzable for post-infection endpoints, 225 and 122 in the vaccine and placebo groups, respectively. The composite endpoint was analyzed for the entire randomized cohort as well as for the cohort of HIV infected subjects. Analyses of the infected subcohort are important because vaccine effects on HIV pathogenesis are most clearly measured in infected subjects, and it is feasible to monitor this subcohort intensively for several years. However, this analysis is not intent-to-treat (ITT) and is susceptible to post-randomization selection bias (Hudgens *et al.*, 2003; Gilbert *et al.*, 2003), and therefore, it is important to also conduct unbiased ITT analyses of the composite endpoint in all randomized subjects. The ITT analyses evaluate the time between randomization and the composite endpoint, and approximate a classical assessment of vaccine efficacy to prevent clinically significant disease (Clements-Mann, 1998). A drawback of the ITT approach is that the follow-up period for capturing endpoints is restricted to the interval during which the entire cohort is followed.

Viral load tends to be highly variable in the first few weeks following HIV infection (Schacker *et al.*, 1998). A small fraction of infected trial participants may have a Month 0.5 viral load value that was measured in this acute phase. For such subjects vaccination may be efficacious to control viral load, but suppression is not yet achieved. To eliminate the influence of possibly unstable Month 0.5 values, measurements at this visit were not used for determining composite endpoints. Therefore, composite endpoints were registered at the earliest date of ART initiation or virologic failure based on a viral load measurement at the Month 1 visit or later. For analyses of the infected subcohort, subjects who did not experience the composite endpoint by 14 months post-infection diagnosis were censored at 14 months, and for randomized cohort analyses, subjects who did not experience the composite endpoint within 36 months of randomization were censored at 36 months. In both analyses subjects lost to follow-up were censored at the date of last contact.

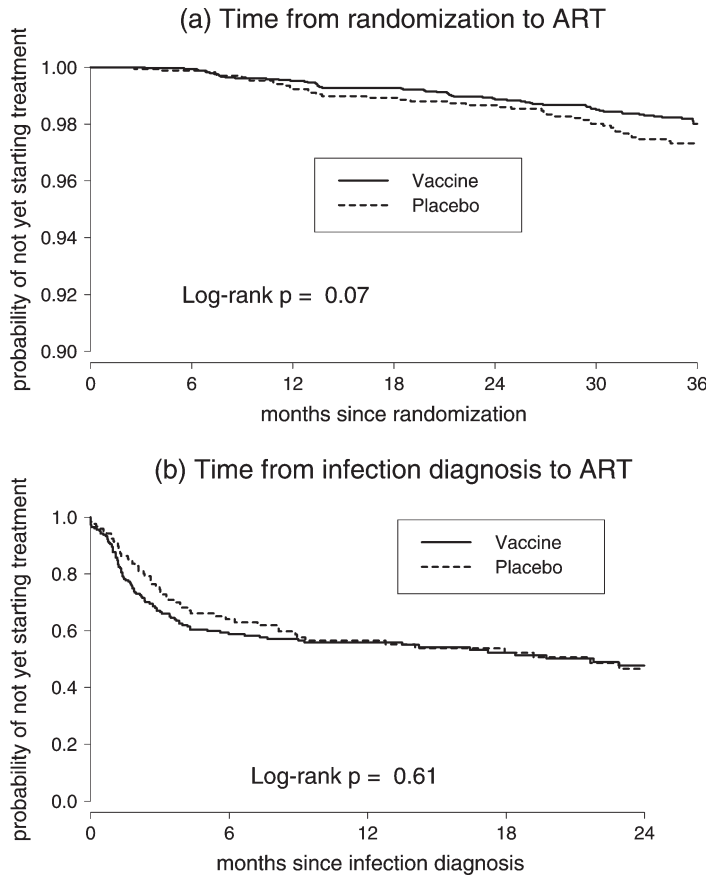

Fig. 2. For the VAX004 trial data, the figure shows Kaplan–Meier curves of (a) the time between randomization and ART initiation and (b) the time between HIV infection diagnosis and ART initiation.

For each cohort and by study arm, Figure 2 shows Kaplan–Meier curves of the time to ART initiation, and Figure 3 shows the pre-ART measurements of viral load. A Cox model analysis verified strongly dependent censoring of pre-ART viral profiles by ART initiation, with estimated hazard ratio 1.88 (95% CI 1.51–2.34, $p < 0.0001$) for each $\log_{10}$ higher value of most recent pre-ART viral load. This result implies that a Kaplan–Meier analysis of the time-to-viral failure with censoring by ART would be severely biased, and motivates analysis of the composite endpoint. For the four pre-specified virologic failure thresholds $x_{\mathrm{vl}} = 1500, 10\,000, 20\,000, 55\,000$ copies/ml, Figure 4 shows Kaplan–Meier curves of the time to the composite endpoint. In the ITT analysis, 290 randomized subjects reached the composite endpoint with $x_{\mathrm{vl}} = 1500$, 227 (78.3%) of whom failed virologically, and 211 subjects reached the composite endpoint with $x_{\mathrm{vl}} = 55\,000$, 117 (55.5%) of whom failed virologically. For the infected cohort, 320 subjects reached the composite endpoint with $x_{\mathrm{vl}} = 1500$, 261 (81.6%) of whom failed virologically, and 237 subjects reached the composite endpoint with $x_{\mathrm{vl}} = 55\,000$, 144 (60.8%) of whom failed virologically. Figure 4 suggests comparable distributions of time-to-composite endpoints in the vaccine and placebo arms.

For formal inferences, the parameter $\mathrm{VE}(14, x_{\mathrm{vl}})$ was assessed for x_{vl} ranging between 1500 and 55 000 copies/ml, where a 14-month time frame post-infection diagnosis was chosen so as to capture all events occurring by the Month 12 visit. Since most subjects failed by the Month 12 visit, an analysis

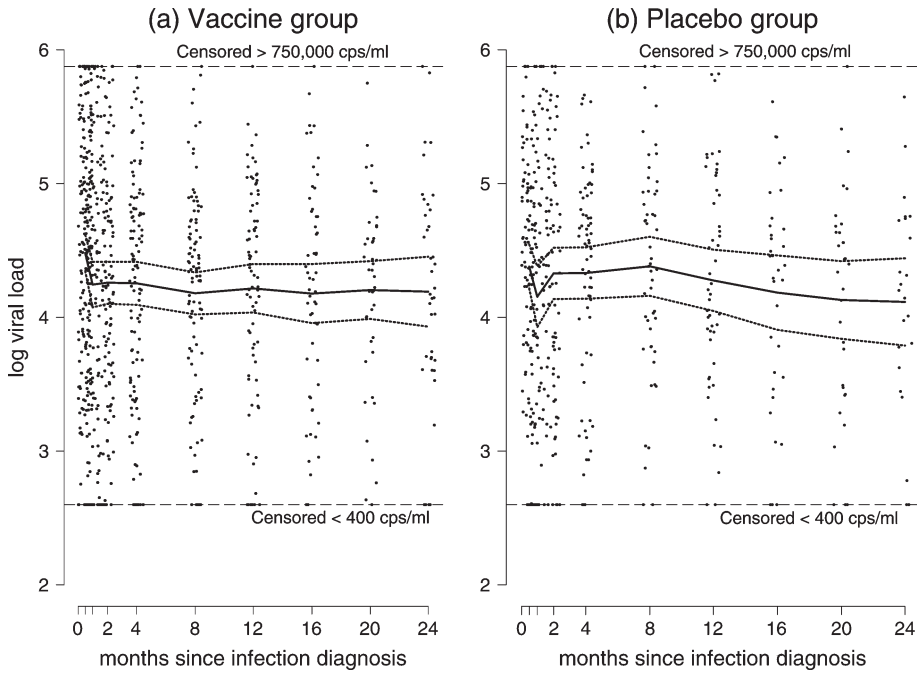

Fig. 3. For the VAX004 trial data, the figure shows the pre-ART measurements of viral load for the (a) vaccine group and (b) placebo group, as a function of the time of sampling post-infection diagnosis. For pre-ART viral loads sampled at the Month 0.5, 1, 2, 4, 8, 12, 16, 20 and 24 visits, the solid lines are mean estimates and the dotted lines are pointwise 95% confidence intervals.

that would use a longer follow-up duration would provide little additional information over the 12-month analysis. For ITT inferences, the parameter $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ was assessed for x_{vl} spanning the same values as for the infected subcohort analysis, where $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ is one minus the ratio (vaccine/placebo) of the cumulative incidence of the composite endpoint occurring between randomization and 36 months. Inference on $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ evaluates the combined effects of vaccination to reduce the infection rate and composite endpoint rate. Figure 5 shows estimates of $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ and $\mathrm{VE}(14, x_{\mathrm{vl}})$, with pointwise and simultaneous 95% confidence interval estimates. Bold vertical segments indicate the simultaneous 95% confidence intervals for the four fixed values of x_{vl}. The confidence coefficient $c_{\alpha/2}$ for each of the bands was obtained by generating $B = 1000$ copies of $U^*(x_{\mathrm{vl}})$.

The point estimates of $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ varied between 0.03 (at $x_{\mathrm{vl}} = 7286$; 3.88 $\log_{10}$) and 0.27 (at $x_{\mathrm{vl}} = 43\,652$; 4.64 $\log_{10}$) over the range of thresholds x_{vl}. The 95% simultaneous bands included zero at all thresholds x_{vl}, indicating no significant differences in the risk of composite endpoints among the groups. The fact that the point estimates of $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ were consistently above zero is explained by the trend toward a longer time until ART initiation in the vaccine group ($p = 0.07$, Figure 2(a)).

The point estimates of $\mathrm{VE}(14, x_{\mathrm{vl}})$ varied between -0.05 and 0.05 and steadily increased with x_{vl}. The simultaneous confidence bands included zero at all threshold values, and were most narrow for $x_{\mathrm{vl}} = 1500$, -0.12 to 0.05, and steadily widened with x_{vl}, with span -0.24 to 0.30 at $x_{\mathrm{vl}} = 55\,000$. This pattern occurred because the number of events decreased with x_{vl}, from 320 events for $x_{\mathrm{vl}} = 1500$ to 237 events for $x_{\mathrm{vl}} = 55\,000$. In comparing the analyses of the randomized and infected cohorts, the simultaneous confidence bands were substantially narrower for the latter analysis, with average half-width 0.39 and 0.16, respectively. This result occurred in part because there were fewer endpoints for the ITT analysis

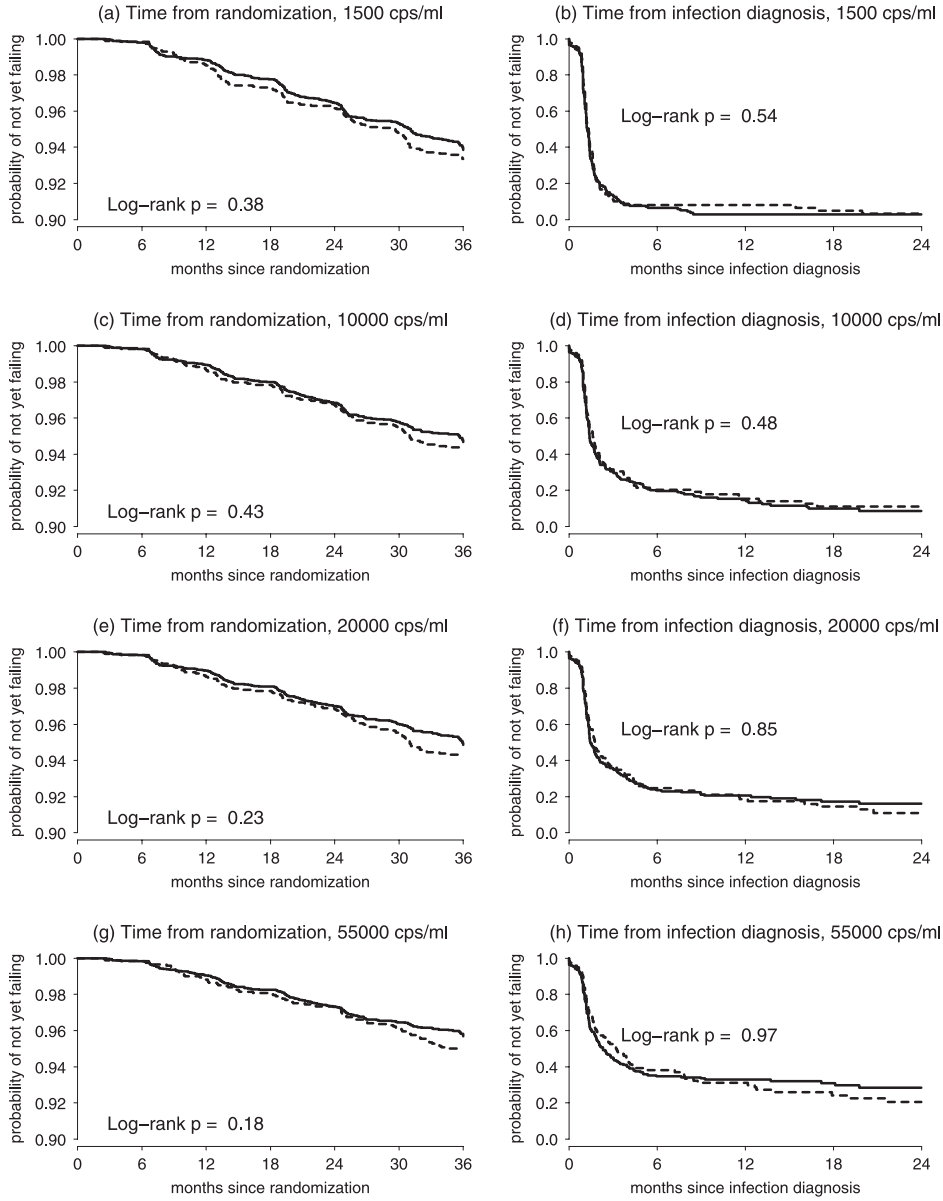

Fig. 4. For the four pre-specified virologic failure thresholds $x_{\text{vl}} = \log_{10}$ 1500, 10 000, 20 000 and 55 000 copies/ml (i.e. levels 3.18, 4.00, 4.30 and 4.74) in the VAX004 trial, the figure shows Kaplan–Meier curves of (left panel) the time between randomization and the composite endpoint, and of (right panel) the time between infection diagnosis and the composite endpoint. The solid (dotted) line denotes the vaccine (placebo) group.

(since composite endpoints occurring beyond 36 months post-randomization were excluded in the ITT inferences; Figure 4).

Notice that for both the ITT and infected cohort analyses, the simultaneous confidence intervals at the four fixed thresholds are substantially narrower than the simultaneous bands computed over the continuous

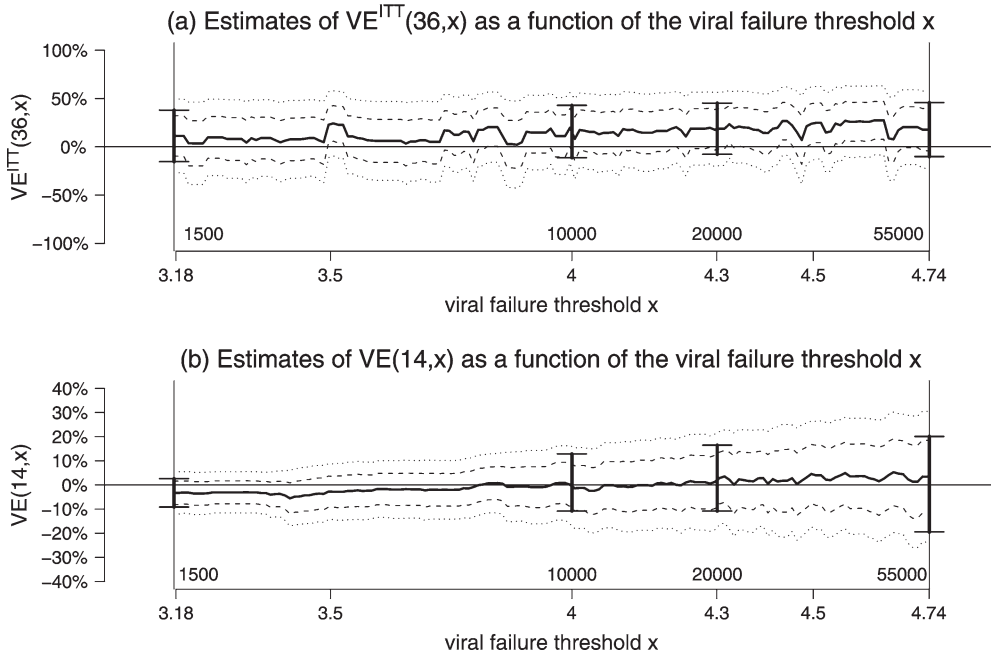

Fig. 5. For the VAX004 trial data, (a) shows 95% pointwise (dashed lines) and simultaneous (dotted lines) confidence intervals about $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ for x_{vl} ranging between 1500 and 55 000 copies/ml on the $\log_{10}$ scale. Solid lines denote estimates of $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$. Bold vertical segments are 95% simultaneous confidence intervals for $\mathrm{VE}^{\mathrm{ITT}}(36, x_{\mathrm{vl}})$ for x_{vl} set at $\log_{10}$ 1500, 10 000, 20 000 and 55 000 and (b) shows the comparable analysis of $\mathrm{VE}(14, x_{\mathrm{vl}})$ for the failure time measured from infection diagnosis.

range of thresholds. This result suggests that one reasonable strategy for future vaccine trials is to apply the procedure using a fixed set of several discrete thresholds that have clinical relevance, if available.

5. COMPLEMENTARY ASSESSMENTS OF POST-INFECTION VACCINE EFFECTS

Alternative approaches to studying vaccine effects on viral load and ART initiation include assessments based on marginal distributions, cause-specific hazard functions or cumulative incidence functions. We consider the value of these approaches. First, the assessment of the vaccine effect on the marginal distribution of the time to ART inititation provides important interpretable information, since ART initiation itself, regardless of reason, is a clinically significant endpoint. This marginal analysis should be done in addition to the composite endpoint analysis. Second, the assessment of the vaccine effect on the marginal distribution of the time-to-viral failure is of little value unless the post-infection follow-up period is very long, because very few viral failure events will occur after ART initiation within a 1–2 year time period. Third, given the arguments made above for focusing inferences on vaccine efficacy parameters that are cumulative rather than instantaneous in time, assessment of cumulative incidence functions is more pertinent than assessment of cause-specific hazard functions. It is informative to study the cumulative incidence functions for viral failure, $F_k^{\mathrm{vl}}(t, x_{\mathrm{vl}}) = P\{\widetilde{T}_k(x_{\mathrm{vl}}) \leqslant t, \delta_k^{\mathrm{vl}} = 1\}, k = 1, 2$, where $\delta_k^{\mathrm{vl}} = 1$ if failure is due to viral load $> x_{\mathrm{vl}}$ and 0 if failure is due to ART initiation. The methods developed here can be adapted to provide simultaneous confidence intervals for $\mathrm{VE}^{\mathrm{vl}}(t, x_{\mathrm{vl}}) = 1 - F_1^{\mathrm{vl}}(t, x_{\mathrm{vl}})/F_2^{\mathrm{vl}}(t, x_{\mathrm{vl}})$ in x_{vl}. Plotting estimates of both $\mathrm{VE}(\tau, x_{\mathrm{vl}})$ and $\mathrm{VE}^{\mathrm{vl}}(\tau, x_{\mathrm{vl}})$ provides information on the degree to which

vaccine efficacy to prevent the composite endpoint is due to prevention of viral failure. In addition, the parameter $\text{PVE}^{\text{vl}}(t, x_{\text{vl}}) = [F_1^{\text{vl}}(t, x_{\text{vl}})/F_1(t, x_{\text{vl}})]/[F_2^{\text{vl}}(t, x_{\text{vl}})/F_2(t, x_{\text{vl}})]$ can be shown to equal the relative probability (vaccine versus placebo) that a composite endpoint failure event by time t was due to viral failure: $\text{PVE}^{\text{vl}}(t, x_{\text{vl}}) = \{\delta_1^{\text{vl}} = 1 | \widetilde{T}_1(x_{\text{vl}}) \leqslant t\}/P\{\delta_2^{\text{vl}} = 1 | \widetilde{T}_2(x_{\text{vl}}) \leqslant t\}$. This ratio can be interpreted as the proportion of the efficacy to prevent the composite endpoint attributable to prevention of viral failure, and estimates of it can also be plotted alongside $\widehat{\text{VE}}(\tau, x_{\text{vl}})$and $\widehat{\text{VE}}^{\text{vl}}(\tau, x_{\text{vl}})$ to provide complementary information.

6. Discussion

Future HIV vaccine efficacy trials are planned to operate under standardized ART initiation guidelines based on viral load and/or CD4 cell count criteria. The guidelines used, and adherence to these guidelines, influence the interpretation of the composite endpoint analysis and the choice of virologic failure thresholds x_{vl}. Based on the current U.S. guidelines (DHHS Guidelines, 2002), it is sensible to assess the composite endpoint for thresholds ranging up to $x_{\text{vl}} = 55\,000$ copies/ml. Because pre-ART virologic failure above x_{vl} for $x_{\text{vl}} \leqslant 55\,000$ usually precedes pre-ART CD4 decline <350 cells/mm^3 (in fact the contrary event never occurred in VAX004), if this guideline is followed, then estimates of $\text{VE}(\tau, x_{\text{vl}})$ for $x_{\text{vl}} \leqslant 55\,000$ have clear interpretations as vaccine effects on the virologic failure rate with little or no confounding by treatment. Although an upper threshold $x_{\text{vl}} = 55\,000$ copies/ml was selected for VAX004 based on the current U.S. guidelines, it should be noted that this choice is somewhat arbitrary, because standardized guidelines were not used for this trial, and prevailing opinions about when to start treatment evolved during the 5 year period of the trial.

For future planned trials that will use standardized ART initiation guidelines, achieving high rates of adherence to the guidelines will make the composite endpoint analysis easier to interpret. In the world's second HIV vaccine efficacy trial, conducted by VaxGen in intravenous drug users in Thailand from 1998 to 2003, the Thai government freely provided ART to infected participants whose CD4 declined below a threshold. Adherence to this national guideline was perfect in that no participant initiated ART prior to meeting the threshold. If there is substantial non-adherence to treatment initiation criteria in an efficacy trial, then the value of the composite endpoint analysis erodes with the degree of non-adherence. In the case that a large fraction of infected subjects start ART prior to meeting treatment criteria, the composite endpoint analysis would contribute little independent information beyond the marginal analysis of ART initiation.

The method developed here applies for analyzing a general composite endpoint defined as the first event of ART initiation or any biomarker-defined endpoint. The method has been applied to assess the first event of CD4 count failure (CD4 count $< x_{\text{CD4}} \in [200, 500]$ cells/mm^3) or ART initiation in the VaxGen Thai trial (unpublished data). Like viral load, CD4 cell count is highly prognostic for progression to AIDS and death (cf. Mellors *et al.*, 1997; HIV Surrogate Marker Collaborative Group, 2000), and based on some studies may be a better predictor of AIDS than viral load near the time of AIDS (Lyles *et al.*, 2000; HIV Surrogate Marker Collaborative Group, 2000). In many developing countries including Botswana, South Africa, Thailand and Uganda, criteria for providing ART through national programs are based on CD4 cell count thresholds but do not consider viral load information. In trials where such treatment policies are operative, analysis of the CD4 cell count/ART initiation composite endpoint may be easier to interpret and have a more direct link to progression to AIDS/death than the analysis of the viral load/ART initiation composite endpoint. A drawback of the CD4-based composite endpoint is that in some trial populations the rate of CD4 cell count decline is quite low (this result was observed in VAX004, with 26% of infected subjects reaching CD4 < 350 cells/mm^3 by 24 months), which restricts the power of the composite endpoint analysis. However, in some populations (e.g. in developing countries) CD4 cells may decline quickly enough to give the analysis reasonably high power; in the VaxGen Thai trial 55% of

infected subjects reached CD4 $<$ 350 cells/mm^3 by 24 months. In any case, any efficacy trial is expected to collect data on both viral load and CD4 cell counts, and analyses of composite endpoints based on both biomarkers will likely be useful for inferring HIV vaccine effects on HIV progression and transmission. The ongoing HIV vaccine efficacy trial in Thailand is using a composite endpoint that includes all three events, ART initiation, viral failure and CD4 failure.

This article has focused on studying $\mathrm{VE}(t, x_{\mathrm{vl}})$ at the latest time point of follow-up after infection diagnosis $t = \tau$. This analysis has greatest importance, because efficacy at later time points predicts greater clinical benefit, and implies greater robustness of the vaccine's efficacy to the possible development of vaccine resistance. It is also of interest to study $\mathrm{VE}(t, x_{\mathrm{vl}})$ over time t, to assess if and how efficacy wanes over time. For a fixed threshold x_{vl}, the procedure of Parzen *et al.* (1997) can be applied to obtain simultaneous confidence intervals for $\mathrm{VE}(t, x_{\mathrm{vl}})$ for t in an interval $[t_1, t_2]$. Since Parzen *et al.*'s (1997) method is based on the same technique used in this article (a martingale approximation and Gaussian multipliers), and our convergence result (A.4) in the Appendix is uniform in t and x_{vl}, it should be possible to combine the two methods into a procedure for constructing a confidence region for $\mathrm{VE}(t, x_{\mathrm{vl}})$ simultaneously in both $t \in [t_1, t_2]$ and $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$. This is left as an open problem.

Finally, note that the proposed procedure can be used to construct simultaneous confidence bands for $F_k(\tau, x_{\mathrm{vl}})$ or for any continuous functional of $F_1(\tau, x_{\mathrm{vl}})$ and $F_2(\tau, x_{\mathrm{vl}})$; for example in some applications it may be of interest to study $F_1(\tau, x_{\mathrm{vl}}) - F_2(\tau, x_{\mathrm{vl}})$.

ACKNOWLEDGMENTS

This work was supported by NIH grant 1RO1 AI054165-01 (Gilbert) and NSF grant DMS-0304922 (Sun).

APPENDIX

We prove (2.1) and the weak convergence of $U(x_{\mathrm{vl}})$ and $U^*(x_{\mathrm{vl}})$. Note that $\widetilde{T}_{ki}(x_{\mathrm{vl}})$ increases as x_{vl} increases. Thus, $V_k(x_{\mathrm{vl}}) = \max_{1 \leqslant i \leqslant n_k} \widetilde{X}_{ki}(x_{\mathrm{vl}})$ increases as x_{vl} increases. We have $\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} I(V_k(x_{\mathrm{vl}}) < \tau) \leqslant I(V_k(x_{\mathrm{vl}}^{\mathrm{L}}) < \tau) \to^p 0$ as $n_k \to \infty$. By Corollary 3.2.1 (Fleming and Harrington 1991, p. 98), it follows that

$$\sqrt{n_k}(\widehat{F}(\tau, x_{\mathrm{vl}}) - F(\tau, x_{\mathrm{vl}})) = S_k(\tau, x_{\mathrm{vl}})\sqrt{n_k} \int_0^\tau \frac{\widehat{S}_k(s-, x_{\mathrm{vl}})}{S_k(s, x_{\mathrm{vl}})} \frac{I(R_k(s, x_{\mathrm{vl}}) > 0)}{R_k(s, x_{\mathrm{vl}})} \, \mathrm{d}M_k(s, x_{\mathrm{vl}}) + \mathrm{o}_p(1), \tag{A.1}$$

where $M_k(t, x_{\mathrm{vl}}) = N_k(t, x_{\mathrm{vl}}) - \int_0^\tau R_k(s, x_{\mathrm{vl}}) \, \mathrm{d}\Lambda(s, x_{\mathrm{vl}})$.

Note that both $S_k(t, x_{\mathrm{vl}})$ and $\widehat{S}_k(t, x_{\mathrm{vl}})$ increase as x_{vl} increases and that $S_k(t, x_{\mathrm{vl}})$ is continuous on $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$. Further, it is known (Fleming and Harrington, 1991) that $\sup_{0 \leqslant t \leqslant \tau} |\widehat{S}_k(t, x_{\mathrm{vl}}) - S_k(t, x_{\mathrm{vl}})| \to^p 0$ pointwise for $x_{\mathrm{vl}} \in [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$. By some elementary analysis, we have $\sup_{x_{\mathrm{vl}}^{\mathrm{L}} \leqslant x_{\mathrm{vl}} \leqslant x_{\mathrm{vl}}^{\mathrm{U}}} \sup_{0 \leqslant t \leqslant \tau} |\widehat{S}_k(t, x_{\mathrm{vl}}) - S_k(t, x_{\mathrm{vl}})| \to^p 0$ as $n_k \to \infty$. Similar arguments lead to the convergence of $R_k(t, x_{\mathrm{vl}})/n_k$ to $r_k(t, x_{\mathrm{vl}})$ in probability, uniformly in $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$.

Next, applying the modern empirical process theory (van der Vaart, 1998), we show that $n_k^{-1/2} M_k(t, x_{\mathrm{vl}})$ converges weakly to a mean-zero Gaussian process with continuous paths. Let

$$M_{ki}^*(u_1, u_2, v_1, v_2, v_3) = N_{ki}(u_1, v_1) - \int_0^{u_2} R_{ki}(s, v_2) \, \mathrm{d}\Lambda_k(s, v_3).$$

We have $M_{ki}^*(t, t, x, x, x) = M_{ki}(t, x)$. Let $\mathcal{F}$ be the class of coordinate projections such that $f_{t,x}(M_{ki}^*) = M_{ki}^*(t, t, x, x, x)$, for $(t, x) \in [0, \tau] \times [x_{\mathrm{vl}}^{\mathrm{L}}, x_{\mathrm{vl}}^{\mathrm{U}}]$. Let $0 = t_0 < t_1 < t_2 < \cdots < t_R = \tau$ and $x_{\mathrm{vl}}^{\mathrm{L}} = x_0 < x_1 < x_2 < \cdots < x_M = x_{\mathrm{vl}}^{\mathrm{U}}$. By the monotone properties of $N_{ki}(t, x_{\mathrm{vl}})$, $R_{ki}(t, x_{\mathrm{vl}})$ and $\Lambda_k(t, x_{\mathrm{vl}})$ on each

coordinate, we have, for $(t, x) \in [t_{r-1}, t_r] \times [x_{m-1}, x_m]$, $M_{ki}(t, x) \leqslant M^*_{ki}(t_r, t_{r-1}, x_{m-1}, x_{m-1}, x_m) \equiv f_u(M^*_{ki})$ and $M_{ki}(t, x) \geqslant M^*_{ki}(t_{r-1}, t_r, x_m, x_m, x_{m-1}) \equiv f_l(M^*_{ki})$. For any $\epsilon > 0$, we can take the number of grids, R and M, in t and in x to be at the order of $1/\epsilon$ such that $E(f_u(M^*_{ki}) - f_l(M^*_{ki}))^2 \leqslant \epsilon$ under the continuity assumptions on the distributions. Hence the bracketing number $N_{[\,]}(\sqrt{\epsilon}, \mathcal{F}, L_2(\mathcal{P}))$ is of the polynomial order $(1/\epsilon)^5$, following the arguments in the proof of Theorem 19.5 and Example 19.6 (van der Vaart, 1998). Therefore, the bracketing integral $J_{[\,]}(1, \mathcal{F}, L_2(\mathcal{P})) < \infty$. By the Glivenko–Cantelli Theorem and Donsker's Theorem (Theorems 19.4 and 19.5 of van der Vaart, 1998), $n_k^{-1/2} M_k(t, x_{\mathrm{vl}}) = n_k^{-1/2} \sum_{i=1}^{n_k} M^*_{ki}(t, t, x_{\mathrm{vl}}, x_{\mathrm{vl}}, x_{\mathrm{vl}})$ converges weakly to a mean-zero Gaussian process with continuous paths, say $G_k(t, x_{\mathrm{vl}})$ on $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x^{\mathrm{L}}_{\mathrm{vl}}, x^{\mathrm{U}}_{\mathrm{vl}}]$.

Applying the Cramér–Wold device and Slutsky's Theorem, we have

$$\left(n_k^{-1/2} M_k(t, x_{\mathrm{vl}}), \widehat{S}_k(t-, x_{\mathrm{vl}}), R_k(t, x_{\mathrm{vl}})/n_k\right) \xrightarrow{D} (G_k(t, x_{\mathrm{vl}}), S_k(t, x_{\mathrm{vl}}), r_k(t, x_{\mathrm{vl}}))$$

on $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x^{\mathrm{L}}_{\mathrm{vl}}, x^{\mathrm{U}}_{\mathrm{vl}}]$. By the strong embedding theorem (Shorack and Wellner, 1986, pp. 47–48), we obtain in a new probability space almost sure convergence of an equivalent process

$$\left(n_k^{-1/2} M^*_k(t, x_{\mathrm{vl}}), \widehat{S}^*_k(t-, x_{\mathrm{vl}}), R^*_k(t, x_{\mathrm{vl}})/n_k\right) \longrightarrow (G^*_k(t, x_{\mathrm{vl}}), S_k(t, x_{\mathrm{vl}}), r_k(t, x_{\mathrm{vl}}))$$

uniformly in $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x^{\mathrm{L}}_{\mathrm{vl}}, x^{\mathrm{U}}_{\mathrm{vl}}]$, where $(n_k^{-1/2} M^*_k(t, x_{\mathrm{vl}}), \widehat{S}^*_k(t-, x_{\mathrm{vl}}), R^*_k(t, x_{\mathrm{vl}})/n_k)$ and $G^*_k(t, x_{\mathrm{vl}})$ are equal in law to $(n_k^{-1/2} M_k(t, x_{\mathrm{vl}}), \widehat{S}_k(t-, x_{\mathrm{vl}}), R_k(t, x_{\mathrm{vl}})/n_k)$ and $G_k(t, x_{\mathrm{vl}})$, respectively. Further, $(n_k^{-1/2} M^*_k(t, x_{\mathrm{vl}}), \widehat{S}^*_k(t-, x_{\mathrm{vl}}), R^*_k(t, x_{\mathrm{vl}})/n_k)$ and $G^*_k(t, x_{\mathrm{vl}})$ can be chosen to have the same sample paths as the original processes. Now, applying the Lemma of Bilias *et al.* (1997) and integration by parts, we have

$$\sqrt{n_k} \int_0^t \left(\frac{\widehat{S}^*_k(s-, x_{\mathrm{vl}})}{S_k(s, x_{\mathrm{vl}})} - 1\right) \frac{I(R^*_k(s, x_{\mathrm{vl}}) > 0)}{R^*_k(s, x_{\mathrm{vl}})} \, \mathrm{d}M^*_k(s, x_{\mathrm{vl}}) \xrightarrow{a.s.} 0 \tag{A.2}$$

and

$$n_k^{-1/2} \int_0^t \left(\frac{n_k I(R^*_k(s, x_{\mathrm{vl}}) > 0)}{R^*_k(s, x_{\mathrm{vl}})} - \frac{1}{r_k(s, x_{\mathrm{vl}})}\right) \mathrm{d}M^*_k(s, x_{\mathrm{vl}}) \xrightarrow{a.s.} 0, \tag{A.3}$$

uniformly in $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x^{\mathrm{L}}_{\mathrm{vl}}, x^{\mathrm{U}}_{\mathrm{vl}}]$. By (A.1),

$$\begin{aligned}
&\sqrt{n_k}(\widehat{F}(t, x_{\mathrm{vl}}) - F(t, x_{\mathrm{vl}})) \\
&\quad = S_k(t, x_{\mathrm{vl}}) \sqrt{n_k} \int_0^t \left(\frac{\widehat{S}_k(s-, x_{\mathrm{vl}})}{S_k(s, x_{\mathrm{vl}})} - 1\right) \frac{I(R_k(s, x_{\mathrm{vl}}) > 0)}{R_k(s, x_{\mathrm{vl}})} \, \mathrm{d}M_k(s, x_{\mathrm{vl}}) \\
&\qquad + S_k(t, x_{\mathrm{vl}}) n_k{}^{-1/2} \int_0^t \left(\frac{I(R_k(s, x_{\mathrm{vl}}) > 0)}{R_k(s, x_{\mathrm{vl}})/n_k} - \frac{1}{r_k(s, x_{\mathrm{vl}})}\right) \mathrm{d}M_k(s, x_{\mathrm{vl}}) \\
&\qquad + S_k(t, x_{\mathrm{vl}}) n_k{}^{-1/2} \int_0^t \frac{1}{r_k(s, x_{\mathrm{vl}})} \, \mathrm{d}M_k(s, x_{\mathrm{vl}}) + \mathrm{o}_p(1).
\end{aligned} \tag{A.4}$$

The first two terms, as processes in (t, x_{vl}), are equal in law to the left side of (A.2) and (A.3), respectively, therefore converge to zero in probability uniformly in $(t, x_{\mathrm{vl}}) \in [0, \tau] \times [x^{\mathrm{L}}_{\mathrm{vl}}, x^{\mathrm{U}}_{\mathrm{vl}}]$ by (A.2) and (A.3). This completes the proof of (2.1).

The weak convergence of $U(x_{\mathrm{vl}})$ and $U^*(x_{\mathrm{vl}})$ can be proved similarly to that of $n_k^{-1/2} M_k(t, x_{\mathrm{vl}})$, $k = 1, 2$, using Lemma 1 of Sun and Wu (2003). We omit the details here.

REFERENCES

ALBERT, J. M., IOANNIDIS, J. P. A., REICHELDERFER, P., CONWAY, B., COOMBS, R. W., CRANE, L., DEMASI, R., DIXON, D. O., FLANDRE, P., HUGHES, M. D., *et al*. (1998). Statistical issues for HIV surrogate endpoints: point/counterpoint. *Statistics in Medicine* **17**, 2435–2462.

BAROUCH, D. H., KUNSTMAN, J., GLOWCZWSKIE, J., KUNSTMAN, K. J., EGAN, M. A., PEYERL, F. W., SANTRA, S., KURODA, M. J., SCHMITZ, J. E., BEAUDRY, K., *et al*. (2003). Viral escape from dominant simian immunodefficiency virus epitope-specific cytotoxic T lymphocytes in DNA-vaccinated rhesus monkeys. *Journal of Virology* **77**, 7367–7375.

BAROUCH, D. H., KUNSTMAN, J., KURODA, M. J., SCHMITZ, J. E., SANTRA, S., PEYERL, F. W., KRIVULKA, G. R., BEAUDRY, K., LIFTON, M. A., GORGONE, D. A., *et al*. (2002). Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. *Nature* **415**, 335–339.

BILIAS, Y., GU, M. AND YING, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Annals of Statistics* **25**, 662–682.

CLEMENS, J. D., NAFICY, A. AND RAO, M. R. (1997). Long-term evaluation of vaccine protection: methodological issues for phase 3 trials and phase 4 studies. In Levine, M. M., Woodrow, G. C., Kaper, J. B. and Cobon, G. S. (eds), *New Generation Vaccines*. New York: Marcel Dekker, pp. 47–67.

CLEMENTS-MANN, M. L. (1998). Lessons for AIDS vaccine development from non-AIDS vaccines. *AIDS Research and Human Retroviruses* **14 (Suppl 3)**, S197–S203.

DHHS GUIDELINES. (2002). *Panel on Clinical Practices for Treatment of HIV Infection. Department of Health and Human Services. Guidelines for the Use of Antiretroviral Agents in HIV-Infected Adults and Adolescents*. February 4, 2002. Available at http://www.aidsinfo.nih.gov/guidelines/.

FLEMING, T. R. (1992). Evaluating therapeutic interventions (with Discussion and Rejoinder). *Statistical Science* **7**, 428–456.

FLEMING, T. R. AND DEMETS, D. L. (1996). Surrogate endpoints in clinical trials: are we being misled? *Annals of Internal Medicine* **125**, 605–613.

FLEMING, T. R. AND HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons.

GILBERT, P., BOSCH, R. AND HUDGENS, M. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59**, 531–541.

GILBERT, P., DEGRUTTOLA, V., HAMMER, S. AND KURITZKES, D. (2001). Virological and regimen termination surrogate endpoints in AIDS clinical trials. *Journal of the American Medical Association* **285**, 775–782.

GILBERT, P., RIBAUDO, H., GREENBERG, L., YU, G., BOSCH, R., TIERNEY, C. AND KURITZKES, D. (2000). Considerations in choosing a primary endpoint that measures durability of virological suppression in an antiretroviral trial. *AIDS* **14**, 1961–1972.

GRAY, R. H., WAWER, M. J., BROOKMEYER, R., SEWANKAMBO, N. K., SERWADDA, P., WABWIRE-MANGEN, F., LUTALO, T., LI, X., VANCOTT, T., QUINN, T. C.; RAKAI PROJECT TEAM. (2001). Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1 discordant couples in Rakai, Uganda. *Lancet* **357**, 1149–1153.

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

HAYES, R. (1998). Design of human immunodeficiency virus intervention trials in developing countries. *Journal of the Royal Statistical Society Series A* **161 (Part 2)**, 251–263.

HIRSCH, M. S., BRUN-VEZINET, F., D'AQUILA, R. T., HAMMER, S. M., JOHNSON, V. A., KURITZKES, D. R., LOVEDAY, C., MELLORS, J. W., CLOTET, B., CONWAY, B., *et al*. (2000). Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society–USA Panel. *Journal of the American Medical Association* **283**, 2417–2426.

HIV SURROGATE MARKER COLLABORATIVE GROUP. (2000). Human immunodeficiency virus type 1 RNA level and CD4 count as prognostic markers and surrogate endpoints: a meta-analysis. *AIDS Research and Human Retroviruses* **16**, 1123–1133.

HUDGENS, M. G., HOERING, A. AND SELF, S. G. (2003). On the analysis of viral load endpoints in HIV vaccine trials. *Statistics in Medicine* **22**, 2281–2298.

HVTN (HIV VACCINE TRIALS NETWORK). (2004). *The Pipeline Project*. Available at http://www.hvtn.org/.

IAVI (INTERNATIONAL AIDS VACCINE INITIATIVE). (2004). *State of Current AIDS Vaccine Research*. Available at http://www.iavi.org/.

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.

LONGINI, I. M., SUSMITA, D. AND HALLORAN, M. E. (1996). Measuring vaccine efficacy for both susceptibility to infection and reduction in infectiousness for prophylactic HIV-1 vaccines. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **13**, 440–447.

LUKASHOV, V. V., GOUDSMIT, J. AND PAXTON, W. A. (2002). The genetic diversity of HIV-1 and its implications for vaccine development. In Wong-Staal, F. and Gallo, R. C. (eds), *AIDS Vaccine Research*. New York: Marcel Dekker, pp. 93–120.

LYLES, R. H., MUNOZ, A., YAMASHITA, T. E., BAZMI, H., DETELS, R., RINALDO, C. R., MARGOLICK, J. B., PHAIR, J. B. AND MELLORS, J. W. (2000). Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS cohort study. *Journal of Infectious Diseases* **181**, 872–880.

MCKEAGUE, I. W., GILBERT, P. B. AND KANKI, P. J. (2001). Omnibus tests for comparison of competing risks with adjustment for covariate effects. *Biometrics* **57**, 818–828.

MELLORS, J. W., MUNOZ, A., GIORGI, J. V., MARGOLICK, J. B., TASSONI, C. J., GUPTA, P., KINGSLEY, L. A., TODD, J. A., SAAH, A. J., DETELS, R., *et al*. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine* **126**, 946–954.

MURPHY, B. R. AND CHANOCK, R. M. (1996). Immunization against virus disease. In Fields, B. N., Knipe, D. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizman, B. and Straus, S. E. (eds), *Fields Virology*. Philadelphia, PA: Lippincott-Raven, pp. 467–497.

NABEL, G. J. (2001). Challenges and opportunities for development of an AIDS vaccine. *Nature* **410**, 1002–1007.

PARZEN, M. I., WEI, L. J. AND YING, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics* **24**, 309–314.

QUINN, T. C., WAWER, M. J., SEWANKAMBO, N., SERWADDA, D., LI, C., WABWIRE-MANGEN, F., MEEHAN, M. O., LUTALO, T. AND GRAY, R. H. (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. *New England Journal of Medicine* **342**, 921–929.

RGP120 HIV VACCINE STUDY GROUP. (2004). Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases*, in press.

SCHACKER, T., COLLIER, A. C., HUGHES, J., SHEA, T. AND COREY, L. (1998). Biological and virologic characteristics of primary HIV infection. *Annals of Internal Medicine* **128**, 613–620.

SHIVER, J. W., FU, T.-M., CHEN, L., CASMIRO, D. R., DAVIES, M. E., EVANS, R. K., ZHANG, Z. Q., SIMON, A. J., TRIGENA, W. L., DUBEY, S. A., *et al*. (2002). Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency virus immunity. *Nature* **415**, 331–335.

SHORACK, G. R. AND WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: John Wiley and Sons.

SUN, Y. AND WU, H. (2003). Time-varying coefficients regression model for longitudinal data. *Technical Report*. University of North Carolina at Charlotte.

UNAIDS (JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS). (2001). Symposium: ethical considerations in HIV preventive vaccine research: examining the 18 UNAIDS guidance points. *Developing World Bioethics* **1**, 121–134.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

[*Received April 12, 2004; revised December 14, 2004; accepted for publication December 15, 2004*]