Biometrics 62, 893–900
September 2006

DOI: 10.1111/j.1541-0420.2005.00516.x

A Comparison of Eight Methods for the Dual-Endpoint Evaluation of Efficacy in a Proof-of-Concept HIV Vaccine Trial

Devan V. Mehrotra,[1,*] Xiaoming Li,[1] and Peter B. Gilbert[2]

[1]Merck Research Laboratories, UN-A102, 785 Jolly Road, Blue Bell, Pennsylvania 19422, U.S.A.
[2]Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington 98109, U.S.A.
[*]*email:* devan_mehrotra@merck.com

Summary. To support the design of the world's first proof-of-concept (POC) efficacy trial of a cell-mediated immunity-based HIV vaccine, we evaluate eight methods for testing the composite null hypothesis of no-vaccine effect on either the incidence of HIV infection or the viral load set point among those infected, relative to placebo. The first two methods use a single test applied to the actual values or ranks of a burden-of-illness (BOI) outcome that combines the infection and viral load endpoints. The other six methods combine separate tests for the two endpoints using unweighted or weighted versions of the two-part z, Simes', and Fisher's methods. Based on extensive simulations that were used to design the landmark POC trial, the BOI methods are shown to have generally low power for rejecting the composite null hypothesis (and hence advancing the vaccine to a subsequent large-scale efficacy trial). The unweighted Simes' and Fisher's combination methods perform best overall. Importantly, this conclusion holds even after the test for the viral load component is adjusted for bias that can be introduced by conditioning on a postrandomization event (HIV infection). The adjustment is derived using a selection bias model based on the principal stratification framework of causal inference.

Key words: Burden of illness; Causal inference; Cell-mediated immunity; Fisher's test; HIV vaccine; Multiple endpoints; Principal stratification; Selection bias; Simes' test.

1. Introduction

More than 20 million people worldwide have died of AIDS since the first cases were identified in 1981, including 3 million deaths in 2004 alone. An estimated 40 million people are currently living with HIV/AIDS, and approximately 15,000 new HIV infections are being added each day (UNAIDS, 2004). An efficacious prophylactic HIV vaccine (administered to HIV uninfected persons) is urgently needed.

The first-generation candidate HIV vaccines, developed in the 1980s and early 1990s, were designed to prevent HIV acquisition by stimulating anti-HIV antibodies. However, antibody-based vaccines failed to lower the rate of HIV infection compared to placebo in the first two large-scale HIV vaccine efficacy trials (The rgp120 HIV Vaccine Study Group, 2005). The absence of protection has been explained, in part, by the inability of the tested vaccines to elicit antibodies that neutralize HIV particles freshly sampled from populations (Burton et al., 2004). Due to HIV's expansive genetic diversity and its many mechanisms of evading neutralization, development of an effective antibody-based HIV vaccine has proven to be an extremely difficult task.

Second-generation HIV vaccine candidates have been designed not to elicit humoral immune responses (antibodies), but rather to elicit cell-mediated immune (CMI) responses (Graham, 2002). These candidates are motivated by increasing evidence that CMI responses, mediated primarily by CD8+ cytotoxic T lymphocytes, play a key role in the control of acute and chronic HIV infection (Borrow et al., 1994; Shiver et al., 2002).

To establish the efficacy of an antibody-based HIV vaccine in a randomized, placebo-controlled clinical trial, it would suffice to demonstrate a statistical difference in the HIV infection rates between vaccine and placebo recipients. But how does one establish the efficacy of a CMI-based HIV vaccine? Vaccine-induced CMI responses (unlike antibody responses) are not expected to impact the initial entry of host cells by HIV. However, they could abort an infection before it becomes fully established (implying a negative HIV diagnostic test), or contain the viral load at a low "set point" in people who become infected despite vaccination. As noted in Gilbert et al. (2003b), the latter outcome would likely provide substantial clinical benefit by preventing or delaying the onset of AIDS, and would decrease the rate of secondary transmission of HIV. These considerations support the use of HIV infection and viral load set point as co-primary endpoints in an efficacy trial of a CMI-based HIV vaccine.

The first proof-of-concept (POC) efficacy trial of a CMI-based HIV vaccine began enrolling volunteers in December 2004. This groundbreaking trial is being conducted by Merck Research Laboratories, in collaboration with the HIV Vaccine Trials Network and the Division of AIDS in the U.S. National Institutes of Health. The candidate vaccine, developed

© 2006, The International Biometric Society

by Merck, consists of a mixture of three identical nonreplicating adenovirus serotype-5 vectors, each encoding the HIV gag, pol, or nef genes as vaccine antigens.

In this article, we use simulations to evaluate eight methods for testing the composite null hypothesis of no-vaccine effect on either efficacy endpoint (infection or viral load set point). The first two methods use a single unconditional test based on the actual values or ranks of a burden-of-illness (BOI) outcome that combines the two endpoints. In contrast, the remaining six methods generate a test statistic (or p-value) by combining two separate tests: an unconditional test for the infection endpoint and a conditional (on HIV infection) test for the viral load endpoint. The approaches used to combine the tests include methods for linearly combining two Z-statistics, methods based on the maximum and minimum of the p-values from the two tests, and methods based on a geometric mean of the two p-values. These methods can incorporate prespecified weights that allow prior data and beliefs on the mechanism of vaccine efficacy to be accounted for to optimize power. While our focus is on HIV vaccine trials, the methods studied can be used more generally to test a composite null hypothesis with multiple efficacy endpoints.

The rest of this article is organized as follows. In Section 2 we define the composite null hypothesis and the data collected for testing it. In Section 3 we describe the eight testing procedures, and in Section 4 we compare their powers in a comprehensive simulation study that was used to design the POC trial. In Section 5 we provide more power comparisons after modifying the combination test methods to build in robustness against potential postrandomization selection bias using the principal stratification framework of causal inference. We conclude with summary remarks in Section 6.

2. Composite Null Hypothesis and Data

In the POC trial, approximately 1500 HIV uninfected adults whose lifestyles put them at relatively high risk of acquiring HIV infection will be randomized in a 1:1 ratio to receive either the HIV vaccine or placebo. All subjects will be tested periodically for acquisition of HIV infection until a total of 50 cases of HIV infection ("events") have accrued; justification for 50 events is provided later. Subjects who are diagnosed as becoming HIV positive will be followed longitudinally for viral load and CD4 cell count evaluations. The viral load set point is defined for this trial as the average of the $\log_{10}$ HIV RNA plasma levels at 2 and 3 months after diagnosis of HIV infection.

Corresponding to the two primary endpoints are two vaccine efficacy parameters of interest: VE_S ("vaccine efficacy for susceptibility") is one minus the true relative risk of HIV infection, and δ_{VL} ("vaccine efficacy for viral load") is the true between-group difference (placebo minus vaccine) in the means of the viral load set points of subjects who become HIV infected. The composite null hypothesis for the POC trial is

$$\mathrm{H}_0 : VE_S = 0 \quad \text{and} \quad \delta_{VL} = 0. \tag{1}$$

Interest lies in testing H_0 versus the one-tailed alternative H_1 : $VE_S > 0$ and/or $\delta_{VL} > 0$. POC is established if H_0 is rejected in favor of H_1.

Let $N_v(N_p)$ be the number of subjects randomized to receive vaccine (placebo), and $n_v(n_p)$ be the number who become HIV infected during the trial, with $\hat{p}_v = n_v/N_v$ and $\hat{p}_p = n_p/N_p$ the proportions infected, and $\bar{p} = (n_v + n_p)/(N_v + N_p)$ the pooled proportion infected. For subjects infected in the vaccine (placebo) group, let $x_1, \ldots, x_{n_v}(y_1, \ldots, y_{n_p})$ be the viral load set points. Finally, let $r = N_p/N_v$ and $D = \hat{p}_v - \hat{p}_p$.

3. Methods for Testing the Composite Null Hypothesis

3.1 *Using a Single Test Based on a Composite Burden-of-Illness Outcome*

To test a composite efficacy hypothesis like (1), Chang, Guess, and Heyse (1994) proposed a method in which first a BOI outcome is observed for each randomized subject. In the context of the POC trial, the outcome is zero if the subject remains HIV uninfected, and is the viral load set point if the subject becomes HIV infected. The BOI per randomized subject is then compared between the placebo and vaccine groups. The numerator of the test statistic is

$$T = \frac{\sum_{i=1}^{n_v} x_i}{N_v} - \frac{\sum_{i=1}^{n_p} y_i}{N_p}.$$

Note that $\sum_{i=1}^{n_v} x_i/N_v = \hat{p}_v(\sum_{i=1}^{n_v} x_i/n_v)$ and $\sum_{i=1}^{n_p} y_i/N_p = \hat{p}_p(\sum_{i=1}^{n_p} y_i/n_p)$, so that the BOI method compares between groups the product of the HIV infection rate and the mean viral load set point among infected subjects. A standardized test statistic based on T is

$$Z_{\mathrm{BOI}} = \frac{T - E(T \mid n_v + n_p, \mathrm{H}_0)}{\sqrt{\hat{V}(T \mid n_v + n_p, \mathrm{H}_0)}}, \tag{2}$$

where $E(T \mid n_v + n_p, \mathrm{H}_0) = 0$, and the variance estimate was derived by Chang et al. (1994) as

$$\hat{V}(T \mid n_v + n_p = n, \mathrm{H}_0) = n\left(\frac{a^2}{N_v N_p} + \frac{s_x^2/N_v + s_y^2/N_p}{N_v + N_p}\right),$$

where $a = n^{-1}(\sum_{i=1}^{n_v} x_i + \sum_{i=1}^{n_p} y_i)$ and $s_x^2(s_y^2)$ is the sample variance of the x's (y's). H_0 is rejected in favor of H_1 at one-tailed level α if $Z_{\mathrm{BOI}} < -Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the 100 $(1-\alpha)$ percentile of N(0, 1).

An alternative to the original BOI method is to use the Wilcoxon rank sum test applied to the BOI outcomes; we refer to this as the rank-based BOI approach. In this approach, the $N_v + N_p$ BOI outcomes for the two randomized groups are pooled and ranked in the usual manner. All subjects who remain HIV uninfected are assigned a tied "best rank" of $0.5 \times (N_v + N_p + 1 - n_v - n_p)$, and among the HIV-infected subjects, those with larger BOIs (= viral load set points) get higher ranks. Let Z_{rankBOI} denote the resulting standardized Wilcoxon rank sum statistic; H_0 is rejected if $Z_{\mathrm{rankBOI}} < -Z_{1-\alpha}$. Of note, Mehrotra, Li, and Gilbert (2005) showed that this approach can inflate the probability of a type I error for an event-driven trial. This follows upon noting that Z_{rankBOI} is a weighted sum of Z_1^* and Z_2^*, where $Z_1^* = (\hat{p}_v - \hat{p}_p)/(\bar{p}(1-\bar{p})(n_v^{-1} + n_p^{-1}))^{1/2}$ is the score statistic for comparing two independent binomial proportions, and Z_2^* is the standardized Wilcoxon rank sum statistic for comparing viral load set point distributions between infected vaccine and infected placebo recipients. If the number of events $(n_v + n_p)$

is fixed, then $\hat{p}_v$ and $\hat{p}_p$ are negatively correlated, implying that the denominator of Z_1^* is smaller than it should be. This explains the inflated rate of rejecting H_0 based on $Z_{\text{rankBOI}} < -Z_{1-\alpha}$. However, because the simplicity of the rank-based BOI approach makes it appealing to clinical investigators, we have included it to explicitly draw attention to its pitfalls, as shown in Section 4.

3.2 *Combining Separate Tests for the Infection and Viral Load Endpoints*

The composite null hypothesis in (1) is an intersection hypothesis: $H_0 = H_0^{INF} \cap H_0^{VL}$, where $H_0^{INF}: VE_S = 0$ and $H_0^{VL}: \delta_{VL} = 0$. Let Z_1 and Z_2 be any valid statistics for one-tailed tests of H_0^{INF} versus $H_1^{INF}: VE_S > 0$, and H_0^{VL} versus $H_1^{VL}: \delta_{VL} > 0$, respectively. For example, if event times are used, the Cox model could be used. However, because of the anticipated low event rate in the POC trial, a test that incorporates event times will not provide appreciably more power than one based on binomial proportions (Cuzick, 1982). Accordingly, we use a test based on the binary HIV infection endpoint. Specifically, note that given $n_v + n_p = n$, n_v is approximately distributed as $Binomial(n, (1+r)^{-1})$ under H_0^{INF}. Hence, a one-tailed p-value p_1 can be obtained as $p_1 = \sum_{x=0}^{n_v} \binom{n}{x} (\frac{1}{1+r})^x (\frac{r}{1+r})^{n-x}$, or as the tail area under the $N(0,1)$ p.d.f. to the left of

$$Z_1 = \frac{D - E(D \mid n_v + n_p, H_0)}{\sqrt{\hat{V}(D \mid n_v + n_p, H_0)}} = \frac{n_v(n_v+n_p)^{-1} - (1+r)^{-1}}{\sqrt{r(1+r)^{-2}(n_v+n_p)^{-1}}}. \tag{3}$$

For Z_2, we use the standardized Wilcoxon rank sum test statistic applied to the viral load set points of the infected subjects; let p_2 denote the corresponding one-tailed p-value.

Let w_1 be a known constant between 0 and 1, and $w_2 = 1 - w_1$. Six procedures for testing the composite null hypothesis in (1) at one-tailed level α based on a combination of Z_1 and Z_2 or p_1 and p_2 are defined below. Note that p_1 and p_2 derived from Z_1 and Z_2 are stochastically independent under H_0. This result, proved by Shih and Quan (1997) in an unrelated context, establishes the validity of the combination tests.

(a) Two-part z-test (O'Brien, 1984): Reject H_0 if $Z = \frac{(Z_1+Z_2)}{\sqrt{2}} < -Z_{1-\alpha}$. (Lachenbruch, 2001 proposed a related two-tailed test based on $Z_1^2 + Z_2^2$, but our problem is one-tailed.)

(b) Weighted two-part z-test (Pocock, Geller, and Tsiatis, 1987; Follmann, 1995): Reject H_0 if $Z_w = \frac{w_1 Z_1 + w_2 Z_2}{\sqrt{w_1^2 + w_2^2}} < -Z_{1-\alpha}$.

(c) Simes' test (Simes, 1986): Reject H_0 if $\max(p_1, p_2) < \alpha$ or $\min(p_1, p_2) < \alpha/2$.

(d) Weighted Simes' test (Hochberg and Liberman, 1994): Reject H_0 if

$$\max\left(\frac{p_1}{2w_1}, \frac{p_2}{2w_2}\right) < \alpha \quad \text{or} \quad \min\left(\frac{p_1}{2w_1}, \frac{p_2}{2w_2}\right) < \alpha/2.$$

(e) Fisher's test (Fisher, 1932): Reject H_0 if $p < \alpha$, where $p = P(\chi_4^2 > -4 \log_e \sqrt{p_1 p_2})$.

(f) Weighted Fisher's test (Good, 1955): Reject H_0 if $p_w < \alpha$, where

$$p_w = \frac{w_1 \tilde{q}^{1/w_1}}{w_1 - w_2} + \frac{w_2 \tilde{q}^{1/w_2}}{w_2 - w_1}, \quad \text{for } w_1 \neq w_2, \quad \text{with } \tilde{q} = p_1^{w_1} \times p_2^{w_2}.$$

Tests 1, 3, and 5 implicitly assign equal weight to the two endpoints ($w_1 = w_2 = 0.5$), while the corresponding tests 2, 4, and 6 allow placing different weights.

If the viral load set points for infected subjects in the vaccine (X) and placebo (Y) groups have normal distributions with means μ_v and μ_p, and variances σ_v^2 and σ_p^2, respectively, then the optimal weight for the viral load endpoint in test 2 (and the presumed nearly optimal weight for tests 4 and 6) can be approximated as

$$w_{2,\text{optimal}} \approx 1 - \frac{VE_S}{VE_S - \sqrt{12(1-VE_S)}\left[\Phi\left(-\delta_{VL}/\sqrt{\sigma_v^2+\sigma_p^2}\right) - 0.5\right]}, \tag{4}$$

where $\delta_{VL} = \mu_p - \mu_v$; see details in the Appendix at `http://www.tibs.org/biometrics`. It is possible, however, that heterogeneity in host genetic characteristics (e.g., human leukocyte antigen alleles) may impact the response to vaccination, resulting in a mixed pool of "weak," "moderate," and "strong" responders to vaccination. Accordingly, we assume that the distribution of X will be similar to that of a mixture of three normal distributions, with mixing proportions $\pi_i (\sum_{i=1}^3 \pi_i = 1)$, means $\mu_{v,i}$, and a common variance $\sigma_{v,\text{all}}^2$. Hence, $w_{2,\text{optimal}}$ is obtained by replacing $\Phi(-\delta_{VL}/\sqrt{\sigma_v^2+\sigma_p^2})$ in (4) with $\sum_{i=1}^3 \pi_i \Phi(-\delta_{VL,i}/\sqrt{\sigma_{v,\text{all}}^2+\sigma_p^2})$, where $\delta_{VL,i} = \mu_p - \mu_{v,i}$.

Table 1 displays values of $w_{2,\text{optimal}}$ for various combinations of VE_S and δ_{VL}, assuming $\delta_{VL,1} = \delta_{VL} - 0.957$, $\delta_{VL,2} = \delta_{VL} - 0.457$, $\delta_{VL,3} = \delta_{VL} + 0.543$, $\sigma_p = 0.75$, $\sigma_{v,\text{all}} = 0.65$, $\pi_1 = 0.2$, $\pi_2 = 0.243$, and $\pi_3 = 0.557$. Based on existing preclinical data (Shiver et al., 2002) and discussions with experts, an educated guess for the POC trial is that the point estimate of VE_S will lie between 0% and 30%, and the point estimate of δ_{VL} will lie between 0.75 and 1.25 $\log_{10}$ copies/ml. The optimal weights for the "middle" value (VE_S, δ_{VL}) = (15%, 1.0) are approximately $w_1 = 0.14$ and $w_2 = 0.86$.

Table 1

Optimal weight $w_{2,\text{optimal}}$ for the viral load component in the weighted two-part z-test for different levels of VE_S and δ_{VL}

	δ_{VL} ($\log_{10}$ copies/ml)				
VE_S(%)	0.50	0.75	1.00	1.25	1.50
0%	∼1	∼1	∼1	∼1	∼1
15%	0.78	0.83	0.86	0.88	0.89
30%	0.62	0.70	0.74	0.77	0.79
45%	0.49	0.57	0.63	0.67	0.69
60%	0.38	0.46	0.52	0.56	0.59
75%	0.28	0.35	0.41	0.45	0.48
90%	0.17	0.22	0.27	0.30	0.32

4. Power Comparisons of the Eight Tests Using Simulations

The type I error rates and powers of the eight testing procedures were evaluated in a comprehensive simulation study to help identify an optimal method for the POC trial. Details of how the data were simulated are provided in the Appendix posted at `http://www.tibs.org/biometrics`. As expected, the observed type I error rate was inflated for the rank-based BOI method (up to 7.2%), but was always less than two standard errors of Monte Carlo variation above the nominal 5% level (<5.6%) for the other methods. Figure 1 shows estimated powers of the testing procedures, that is, the proportion of times that H_0 in (1) was rejected for each method in 5000 simulations at each fixed n. When all of the vaccine effect is on the viral load endpoint ($VE_S = 0\%$), the three weighted combination tests have high and equivalent power. The unweighted combination methods also do well but have slightly lower power, demonstrating that up-weighting the viral load endpoint had the intended effect. In contrast, the BOI methods have very low power when $VE_S = 0\%$, demonstrating that this approach cannot be recommended if there is high pretest probability that the vaccine is unable to lower susceptibility to HIV infection. When $VE_S = 30\%$, the combination test methods all perform well and comparably; the contribution of vaccine efficacy to prevent infection makes the weighted and unweighted methods perform similarly. Again the BOI methods have much lower power than the combination test methods. When $VE_S = 60\%$, the BOI method is competitive,

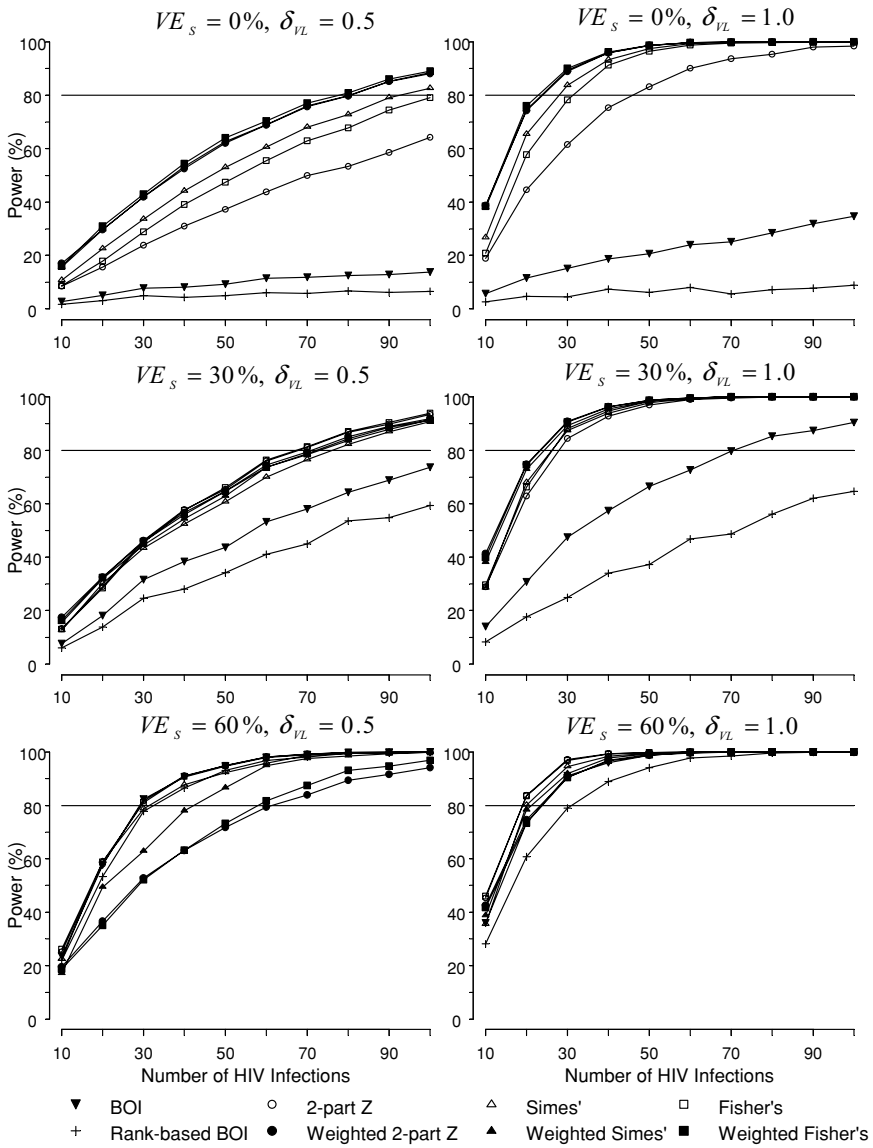

Figure 1. Estimated powers of the eight testing procedures for rejecting the composite null hypothesis $H_0 : VE_S = 0$ and $\delta_{VL} = 0$ as a function of the total number of HIV infections (events), for different combinations of VE_S (%) and δ_{VL} ($\log_{10}$ copies/ml).

Table 2

Total number of events (HIV infections) required to have 80% *power to reject the composite null hypothesis* H_0: $VE_S = 0$ *and* $\delta_{VL} = 0$ *for different levels of* VE_S *and* δ_{VL}, *for the Simes' (S), Fisher's (F), weighted Simes' (WS), and weighted Fisher's (WF) combination tests*

	δ_{VL} ($\log_{10}$ copies/ml)															
	0.0				0.5				0.75				1.0			
VE_S(%)	S	F	WS	WF	S	F	WS	WF	S	F	WS	WF	S	F	WS	WF
0%	>100	>100	>100	>100	93	>100	81	75	45	49	37	37	28	31	24	23
10%	>100	>100	>100	>100	92	92	79	75	44	47	38	37	28	29	24	23
20%	>100	>100	>100	>100	87	78	77	72	44	41	38	36	27	27	23	23
30%	>100	>100	>100	>100	77	64	74	71	42	36	39	36	27	25	24	23
40%	>100	>100	>100	>100	63	50	65	67	40	32	37	35	25	23	24	23
50%	77	78	>100	>100	49	37	53	64	33	27	33	35	23	20	23	23
60%	47	47	63	>100	35	29	44	57	28	22	31	34	23	17	23	23
70%	28	30	39	>100	25	21	31	52	23	17	26	33	17	15	20	23

with power equal to that of the best-performing combination test methods to detect $\delta_{VL} = 0.5$ and only slightly lower to detect $\delta_{VL} = 1.0$. The weighted combination methods have substantially lower power than their unweighted counterparts for $VE_S = 60\%$, $\delta_{VL} = 0.5$, demonstrating that up-weighting the endpoint on which there is a smaller effect size results in a power loss.

Based on the above power analysis, all of the evaluated combination test methods have acceptable performance for a POC trial of a CMI-based HIV vaccine. Simes' and Fisher's methods may be preferable to the two-part z-methods because their power is less affected by the choice of weight function (Figure 1, top two panels).

The total number of HIV infections required to have 80% power to establish POC is summarized in Table 2 for the four leading combination test methods. For example, 93 events will provide 80% power to reject H_0 using the unweighted Simes' method (S) when $VE_S = 0\%$ and $\delta_{VL} = 0.5$ $\log_{10}$ copies/ml. Note that for low values of $VE_S(\leq 30\%)$, the methods that up-weight the viral load endpoint require slightly fewer infections to detect the same viral load effect as the equal-weighted methods, but for moderate to high values of $VE_S(\geq 50\%)$ the former require notably more infections. In general, the number of infections required varies less over the range of efficacy parameter values for the equal-weighted procedures. These results suggest that assigning equal weight to the two endpoints provides the greatest robustness to uncertainties in the true nature of vaccine efficacy. Accordingly, the unweighted Simes' and Fisher's methods emerge as optimal choices for evaluating the efficacy of the CMI-based HIV vaccine. The former was selected for the POC trial, in part because rejection of the composite null hypothesis in (1) using Simes' test "automatically" provides conclusions about statistical significance separately for the two endpoints without any further multiplicity adjustment; this follows because with only two endpoints, Simes' procedure is identical to Hochberg's (1988).

5. Power Comparisons after Adjusting for Postrandomization Selection Bias

To this point, the combination test methods have used a test for the viral load endpoint that compares the mean viral load set points of HIV-infected subjects in the vaccine and placebo groups. Because the test is restricted to subjects who are selected based on a postrandomization event (HIV infection), it does not assess a causal effect of vaccine (Robins and Greenland, 1992). Rather, it assesses viral load differences due to a mixture of two effects: the causal vaccine effect and the effect of variables correlated with viral load that are (potentially) unevenly distributed among the infected subgroups (Frangakis and Rubin, 2002).

In Section 4, we discarded the BOI method and chose the unweighted Simes' and Fisher's combination test methods based on the latter having better power for the POC trial. However, the BOI method usefully provides unbiased inferences on a causal effect of the vaccine, because it is based on all randomized subjects. In contrast, the combination test methods provide unbiased causal inferences only under the untestable assumption of no selection bias for the viral load component. We now proceed to show that the BOI method is generally less powerful than the two leading combination test methods even after the latter are adjusted for plausible levels of selection bias in a manner that makes it harder for them to reject H_0. To do so, we use the potential outcomes framework for causal inference (Rubin, 1974).

Each subject i has two potential HIV infection outcomes: one under assignment to vaccine ($S_i(v)$) and one under assignment to placebo ($S_i(p)$). In addition, each subject if infected under assignment to vaccine has a potential viral load set point $VLS_i(v)$, and if infected under assignment to placebo has a potential viral load set point $VLS_i(p)$. Following Hudgens, Hoering, and Self (HHS) (2003) and Gilbert, Bosch, and Hudgens (GBH) (2003a), a causal vaccine effect on viral load can be defined for the "always-infected" principal stratum of subjects who would become HIV infected regardless of randomization to vaccine or placebo (i.e., those with $S_i(v) = S_i(p) = 1$). Any functional that measures a contrast of the distributions

$$F^{\text{alw. inf}}_{(v)}(y) \equiv P(VLS_i(v) \leq y \,|\, S_i(v) = S_i(p) = 1) \quad \text{and}$$
$$F^{\text{alw. inf}}_{(p)}(y) \equiv P(VLS_i(p) \leq y \,|\, S_i(v) = S_i(p) = 1) \tag{5}$$

is a causal estimand. Unfortunately, because neither distribution in (5) is readily identifiable (because $S_i(v)$ and $S_i(p)$ are

not both observed), it is possible to assess a causal vaccine effect on viral load only after making some assumptions.

Following Rubin (1974), HHS and GBH outlined three assumptions: (i) the potential outcomes for each subject are independent of the treatment assignments of other subjects, (ii) the treatment assignment for each subject is independent of his/her potential outcomes, and (iii) $S_i(v) \leq S_i(p)$ for all subjects i, that is, the vaccine does not increase the risk of acquiring HIV infection. Under these assumptions, $F_{(v)}^{\text{alw. inf}} = F_v$, that is, the distribution of potential viral loads under assignment to vaccine equals the identifiable distribution of viral loads in infected vaccine recipients. Moreover, as in GBH, the above three assumptions plus the selection model

$$F_{(p)}^{\text{alw. inf}}(y, \beta) = (1 - VE_S)^{-1} \int_0^y w(z, \beta)\, dF_p(z) \qquad (6)$$

identify $F_{(p)}^{\text{alw. inf}}$, where F_p is the c.d.f. of the viral load set point in infected placebo recipients, $w(y, \beta) = \exp(\tau + \beta y)/[1 + \exp(\tau + \beta y)]$ is a selection weight function, $\beta \in [-\infty, \infty]$ is a parameter fixed by the investigator that quantifies the degree of selection bias, and τ is determined by the equation $F_p(\infty \,|\, \beta) = 1$. For finite β, $e^{-\beta}$ is the odds ratio of HIV infection under assignment to vaccine given infection under assignment to placebo with viral load set point y versus with viral load set point $y + 1$. $\beta = 0$ specifies no selection bias, and $\beta > 0$ ($\beta < 0$) specifies bias toward the infected vaccinees having selectively higher (lower) viral load set points. To protect against selection bias that could artificially favor the vaccine, we focus on $\beta < 0$.

The causal null hypothesis of interest for the viral load endpoint is $\text{H}_{0,\text{causal}}^{VL} : \delta_{VL}^{ACE} = 0$, where $\delta_{VL}^{ACE} = \int y\, dF_{(p)}^{\text{alw. inf}}(y) - \int y\, dF_{(v)}^{\text{alw. inf}}(y)$ is the average causal effect. Note that given a fourth assumption: (iv) selection bias operates only through a vaccine effect on the infection endpoint, there is no opportunity for selection bias when $VE_S = 0$ (regardless of β), and $w(y, \beta) = 1$. Hence, $\delta_{VL} = \delta_{VL}^{ACE}$ when $VE_S = 0$, implying that under assumptions (i)–(iv) the composite null hypothesis in (1) can be rewritten as

$$\text{H}_0 : VE_S = 0 \quad \text{and} \quad \delta_{VL}^{ACE} = 0. \qquad (7)$$

To account for potential selection bias resulting from $VE_S > 0$, we replace Z_2 in the combination test methods with a statistic T_β that tests for a vaccine effect on viral load that is above and beyond a plausible level of selection bias indexed by β. Among several options for T_β, we propose using a rank statistic that is consistent with the analysis discussed earlier. Specifically, let $\bar{y} = n_p^{-1} \sum_{i=1}^{n_p} y_i$ denote the observed mean viral load set point for the placebo group, and let

$$y_{i,\beta}^* = y_i - \left(\bar{y} - \frac{\sum_{i=1}^{n_p} w(y_i \,|\, \hat{\tau}, \beta) y_i}{\sum_{i=1}^{n_p} w(y_i \,|\, \hat{\tau}, \beta)} \right) \qquad (8)$$

denote the "adjusted" (reduced) viral load set point for infected subject i in the placebo group, where $\hat{\tau}$ is obtained as described in GBH. Our proposed T_β is the Wilcoxon rank sum test statistic calculated using the adjusted and observed viral load set points in the placebo and vaccine groups, respectively. Because the null distribution of T_β is intractable, the p-value based on T_β, denoted by $p_{2,\beta}$, is obtained using an adaptation of the bootstrap procedure proposed by GBH. If the estimated VE_S is ≤ 0, then $w(y_i \,|\, \hat{\tau}, \beta) = 1\, \forall i, \beta$, in which case T_β will equal Z_2, the unadjusted Wilcoxon statistic.

Table 3 shows the estimated number of events required to have 80% power to establish POC for the selection-bias-adjusted Simes' and Fisher's combination test methods for $\beta = -1$ and -2 (selection odds ratios of $e^{-\beta} = 2.7$ and 7.4, respectively), and for $\beta = -\infty$. Corresponding results without a selection bias adjustment ($\beta = 0$) and results for the BOI method are included for comparison. Note that the number of events required for the adjusted Simes' and Fisher's tests increases as β becomes more negative, reflecting a higher hurdle for establishing POC as a larger amount of selection bias is assumed. When $VE_S = 15\%$, the adjusted combination test methods require notably fewer events than the BOI method under any degree of selection bias. This is also generally true when $VE_S = 30\%$, though the advantage over BOI is smaller. When $VE_S = 60\%$, the combination test and BOI methods require comparable numbers of events assuming no selection

Table 3

Total number of events (HIV infections) required to have 80% power to reject the composite null hypothesis $\text{H}_0 : VE_S = 0$ *and* $\delta_{VL} = 0$ *for different levels of* VE_S *and* δ_{VL}*, for the Simes' (S) and Fisher's (F) combination tests with* ($\beta < 0$) *or without* ($\beta = 0$) *an adjustment for potential selection bias, and for the BOI method*

δ_{VL} ($\log_{10}$ copies/ml)	BOI	$\beta = 0^*$		$\beta = -1$		$\beta = -2$		$\beta = -\infty$	
		S	F	S	F	S	F	S	F
$VE_S = 15\%$									
0.75	>100	44	45	54	52	61	59	78	72
1.00	>100	27	28	32	32	35	35	40	40
$VE_S = 30\%$									
0.75	91	42	36	59	51	77	63	>100	85
1.00	74	27	25	34	32	41	37	53	46
$VE_S = 60\%$									
0.75	24	28	22	37	30	42	36	86	71
1.00	20	23	17	30	23	35	28	56	47

*No adjustment for selection bias.

bias ($\beta = 0$), but after adjusting the former for selection bias the BOI method is more powerful (i.e., it needs relatively few events), with greater advantage when more selection bias is controlled for.

These simulations demonstrate that if the true vaccine effect on the infection endpoint is somewhere between absent to moderate (up to about 50%), then the adjusted Simes' and Fisher's methods are more powerful than the BOI method, even after building in robustness to selection bias. It is only when VE_S is fairly high (>50%) that the BOI method is the more powerful procedure. But in that case, the viral load comparison becomes less important because POC will likely be established based on a causal vaccine effect on the infection endpoint. These results suggest that for a POC efficacy trial of a CMI-based HIV vaccine that is more likely to work by lowering postinfection viral load set points rather than preventing HIV infection, the selection-bias-adjusted Simes' and Fisher's combination test methods are preferred over the BOI method.

6. Discussion

We have demonstrated that for a POC efficacy trial of a CMI-based HIV vaccine, the unconditional BOI method is generally less powerful than methods that combine an unconditional test for the infection endpoint with a conditional test for the viral load endpoint. In particular, we have staked a case for choosing either the unweighted Simes' or Fisher's combination test for establishing POC. Both methods are generally more powerful than the BOI method even after the test for the viral load component is adjusted for selection bias that might artificially favor the vaccine; the power advantages are substantial when the vaccine has at most a modest effect on the infection endpoint, which is more likely for a CMI-based vaccine.

We conclude by noting that arguments can be made both for and against adjusting for potential selection bias in a POC trial. Consider first the arguments against adjustment. We have shown that under assumptions (i)–(iv), rejection of H_0 implies that $VE_S > 0$ and/or $\delta_{VL}^{ACE} > 0$. This finding of some benefit, albeit with uncertainty about the component effects, may form sufficient justification for advancing a vaccine candidate to a subsequent large-scale trial. Additional support for using the combination tests without accounting for selection bias derives from noting that the most likely operative selection mechanisms would create selectively higher viral loads in infected vaccinees (e.g., due to a greater propensity for vaccine failure in those with relatively weak immune systems). It can therefore be argued that the unadjusted combination test methods (that assume $\beta = 0$) already have built-in robustness against the selection bias effects of interest.

To make the argument in favor of a selection-bias adjustment, we note that the prevailing majority opinion in the field is that immune responses induced by a CMI-based HIV vaccine are more likely to control postinfection HIV replication rather than reduce the risk of HIV infection. An observed moderate effect for the infection endpoint (e.g., $VE_S^{\text{obs}} = 25\%$) in tandem with a moderate effect for the viral load endpoint (e.g., $\delta_{VL}^{\text{obs}} = 0.75$ copies/ml) may be difficult to interpret, because there will be a high degree of uncertainty about the extent to which the observed vaccine effect on viral load is a causal effect. From this point of view, initiation of a subsequent large-scale trial may not be warranted unless there is "robust" evidence for a causal vaccine effect on viral load. Of note, even if an adjustment for plausible levels of selection bias is deemed necessary, approximately 50 events will provide at least 80% power to establish POC if $VE_S \geq 60\%$ or $\delta_{VL} \geq 1.0$ $\log_{10}$ copies/ml (Table 3, method S or F). Accordingly, the POC trial was designed to accrue 50 events, with a planned interim analysis of efficacy at 30 events (details omitted). A "positive" result at the interim analysis (based on prespecified statistical criteria) could advance the vaccine to a large-scale efficacy trial approximately 9–15 months sooner than the analysis at 50 events.

Acknowledgements

The authors thank the editor, associate editor, and two referees for helpful comments.

References

Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M., and Oldstone, M. B. (1994). Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *Journal of Virology* **68,** 6103–6110.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koffi, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004). HIV vaccine design and the neutralizing antibody problem. *Nature Immunology* **5,** 233–236.

Chang, M. N., Guess, H. A., and Heyse, J. F. (1994). Reduction in the burden of illness: A new efficacy measure for prevention trials. *Statistics in Medicine* **13,** 1807–1814.

Cuzick, J. (1982). The efficiency of the proportions test and the log rank test for censored survival data. *Biometrics* **38,** 1033–1039.

Fisher, R. A. (1932). *Statistical Methods for Research Workers.* Edinburgh and London: Oliver and Boyd.

Follmann, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine* **14,** 1163–1175.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58,** 21–29.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003a). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59,** 531–541.

Gilbert, P. B., DeGruttola, V., Hudgens, M. G., Self, S. G., Hammer, S. M., and Corey, L. (2003b). What constitutes efficacy for a human immunodeficiency virus vaccine that ameliorates viremia: Issues involving surrogate endpoints in phase III trials. *Journal of Infectious Diseases* **188,** 179–193.

Good, I. J. (1955). On the weighted combination of significance tests. *Biometrika* **42,** 264–265.

Graham, B. S. (2002). Clinical trials of HIV vaccines. *Annual Review of Medicine* **53,** 207–221.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802.

Hochberg, Y. and Liberman, U. (1994). An extended Simes test. *Statistics and Probability Letters* **21,** 101–105.

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.

Lachenbruch, P. A. (2001). Comparison of two-part models with competitors. *Statistics in Medicine* **20,** 1215–1234.

Mehrotra, D. V., Li, X., and Gilbert, P. B. (2005). *Dual-endpoint evaluation of vaccine efficacy—Application to a proof-of-concept clinical trial of a cell mediated immunity-based HIV vaccine.* BARDS Technical Report 111, Merck Research Laboratories, Blue Bell, Pennsylvania.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40,** 1079–1087.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43,** 487–498.

The rgp120 HIV Vaccine Study Group. (2005). Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. *Journal of Infectious Diseases* **191,** 654–665.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability of direct and indirect effects. *Epidemiology* **3,** 143–155.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Shih, W. J. and Quan, H. (1997). Testing for treatment differences with dropouts present in clinical trials—A composite approach. *Statistics in Medicine* **16,** 1225–1239.

Shiver, J. W., Fu, T.-M., Chen, L., et al. (2002). Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency virus immunity. *Nature* **415,** 331–335.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.

UNAIDS. (2004). *Joint United Nations Programme for HIV/AIDS. AIDS Epidemic Update 2004.* `http://www.unaids.org/bangkok2004/report.html.`

Received May 2005. *Revised October* 2005.
Accepted October 2005.