Statistical Communications in Infectious Diseases

Volume 3, *Issue* 1 2011 *Article* 4

A Sequential Phase 2b Trial Design for Evaluating Vaccine Efficacy and Immune Correlates for Multiple HIV Vaccine Regimens

Peter B. Gilbert, *Fred Hutchinson Cancer Research Center and University of Washington*
Douglas Grove, *Fred Hutchinson Cancer Research Center*
Erin Gabriel, *University of Washington*
Ying Huang, *Fred Hutchinson Cancer Research Center*
Glenda Gray, *University of the Witwatersrand*
Scott M. Hammer, *Columbia University Medical Center*
Susan P. Buchbinder, *San Francisco Department of Public Health and University of California, San Francisco*
James Kublin, *Fred Hutchinson Cancer Research Center*
Lawrence Corey, *Fred Hutchinson Cancer Research Center and University of Washington*
Steven G. Self, *Fred Hutchinson Cancer Research Center and University of Washington*

Recommended Citation:
Gilbert, Peter B.; Grove, Douglas; Gabriel, Erin; Huang, Ying; Gray, Glenda; Hammer, Scott M.; Buchbinder, Susan P.; Kublin, James; Corey, Lawrence; and Self, Steven G. (2011) "A Sequential Phase 2b Trial Design for Evaluating Vaccine Efficacy and Immune Correlates for Multiple HIV Vaccine Regimens," *Statistical Communications in Infectious Diseases*: Vol. 3: Iss. 1, Article 4.
DOI: 10.2202/1948-4690.1037
Available at: http://www.bepress.com/scid/vol3/iss1/art4

A Sequential Phase 2b Trial Design for Evaluating Vaccine Efficacy and Immune Correlates for Multiple HIV Vaccine Regimens

Peter B. Gilbert, Douglas Grove, Erin Gabriel, Ying Huang, Glenda Gray, Scott M. Hammer, Susan P. Buchbinder, James Kublin, Lawrence Corey, and Steven G. Self

Abstract

Five preventative HIV vaccine efficacy trials have been conducted over the last 12 years, all of which evaluated vaccine efficacy (VE) to prevent HIV infection for a single vaccine regimen versus placebo. Now that one of these trials has supported partial VE of a prime-boost vaccine regimen, there is interest in conducting efficacy trials that simultaneously evaluate multiple prime-boost vaccine regimens against a shared placebo group in the same geographic region, for accelerating the pace of vaccine development. This article proposes such a design, which has main objectives (1) to evaluate VE of each regimen versus placebo against HIV exposures occurring near the time of the immunizations; (2) to evaluate durability of VE for each vaccine regimen showing reliable evidence for positive VE; (3) to expeditiously evaluate the immune correlates of protection if any vaccine regimen shows reliable evidence for positive VE; and (4) to compare VE among the vaccine regimens. The design uses sequential monitoring for the events of vaccine harm, non-efficacy, and high efficacy, selected to weed out poor vaccines as rapidly as possible while guarding against prematurely weeding out a vaccine that does not confer efficacy until most of the immunizations are received. The evaluation of the design shows that testing multiple vaccine regimens is important for providing a well-powered assessment of the correlation of vaccine-induced immune responses with HIV infection, and is critically important for providing a reasonably powered assessment of the value of identified correlates as surrogate endpoints for HIV infection.

KEYWORDS: HIV vaccine efficacy clinical trial, immune correlate of protection, one-way crossover design, surrogate endpoint for HIV infection, two-phase sampling

Author Notes: Peter B. Gilbert, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, and University of Washington, Seattle, WA. Douglas Grove, Vaccine Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA. Erin Gabriel, University of Washington. Ying Huang, Vaccine Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA. Glenda Gray, Perinatal HIV Research Unit, University of

the Witwatersrand, Johannesburg, South Africa. Scott M. Hammer, Division of Infectious Diseases, Columbia University Medical Center, New York, NY. Susan P. Buchbinder, San Francisco Department of Public Health, University of California San Francisco, San Francisco, CA. James Kublin, Vaccine Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA. Lawrence Corey, Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA. Steven G. Self, Vaccine Infectious Disease Division, Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA. The authors thank the participants of the workshop, who provided insightful input into the proposed study design and into avenues of additional research. The NIAID-sponsored workshop, "Alternative Study Design for Early Efficacy Evaluation of HIV Prophylactic Vaccines," was held in Bethesda on January 11, 2011. The authors also thank the workshop organizing committee, especially Elizabeth Adams and Mike Proschan. This research was funded by NIH grant 2 R37 AI054165-08 and by NIH NIAID 5 U01 AI068635. SMH was supported by NIH grant AI069470.

Introduction

Background on Past HIV Vaccine Efficacy Trials, with Emphasis on the Sequential Monitoring Plans.

Five randomized, double-blinded, placebo-controlled preventative HIV vaccine efficacy trials have been conducted, all with HIV infection as a primary endpoint, four of which yielded results on the vaccine efficacy (VE) to reduce the rate of HIV infection [VE = (1 – HR)×100%, where HR is the hazard ratio (vaccine/placebo) of HIV infection diagnosis]. The Vax004, Vax003, and Step trials indicated that VE was zero or very low at best (Flynn et al., 2005; Pitisuttithum et al., 2006; Buchbinder et al., 2008), whereas the RV144 trial provided modest evidence for positive VE (estimated VE = 31%, 95% confidence interval (CI) 1% to 51%, 2-sided p-value = 0.04) (Rerks-Ngarm et al., 2009). RV144 evaluated a prime-boost vaccine regimen, and several products are becoming available that may be combined into novel prime–boost regimens, generating enthusiasm for a follow-up efficacy trial (or trials) that will evaluate multiple such regimens. Here we propose a Phase 2b design for a follow-up trial configured to accelerate the pace of answering key scientific questions and hence to shorten the time until the eventual licensure of an efficacious HIV vaccine. The main features of the proposed design are to evaluate multiple vaccine regimens versus a shared placebo group, adaptive two-stage evaluation of vaccine efficacy against infections occurring proximal or distal to the immunization series, tailored sequential monitoring for optimizing efficiency of vaccine efficacy evaluation, augmented design features to improve the assessment of immune correlates of protection, and head-to-head comparisons of vaccine efficacy among vaccine regimens.

The previous efficacy trials used group sequential designs, wherein an independent Data Safety Monitoring Board (DSMB) periodically reviewed interim results on estimation and inference for VE (Table 1). Vax004 and Vax003 had essentially the same Phase 3 design, whereas Step and Phambili (Gray et al., 2009) (Phambili did not yield a result on VE) had essentially the same Phase 2b design. All four trials evaluated VE at a single interim analysis; Vax004 and Vax003 used O'Brien-Fleming monitoring (O'Brien and Fleming, 1979) to recommend early stopping based on strong evidence for reasonably high efficacy (test H0: VE $\leq$ 30% vs. H1: VE > 30%), whereas Step and Phambili used a customized monitoring procedure to recommend early stopping based on strong evidence for positive efficacy on either the infection endpoint (test H0: VE $\leq$ 0% vs. H1: VE > 0%) or on the set-point viral load co-primary endpoint. At the sole interim analysis Step was also monitored for low efficacy at best (we refer to this as "non-efficacy monitoring"). In particular, conditional power monitoring was

used to recommend early stopping if there was less than a 20% chance to reject the composite null hypothesis of both VE ≤ 0% and no vaccine effect on mean viral load, if in future follow-up the true VE would be 60% and the true viral load effect would be a 1 $\log_{10}$ lower mean viral load in the infected vaccine group compared to the infected placebo group. By only allocating a small part of the overall type I error rate to the interim analysis, this monitoring procedure, similar to the O'Brien-Fleming approach, only recommended stopping based on strong interim evidence. Phambili planned similar non-efficacy monitoring, but the trial was un-blinded before the planned interim analysis (the un-blinding was precipitated by evidence from the Step trial that the vaccine may cause an increased risk of HIV acquisition, Buchbinder et al., 2008).

Table 1. Approaches to Group Sequential Monitoring of HIV Vaccine Efficacy in Past Efficacy Trials.

Efficacy Trial	Monitoring Type	Number and Timing of Interim Analyses	Null and Alternative Hypotheses	Alpha level	Boundary Type
Vax004 Phase 3 1998-2003	Efficacy	1; when 50% infections expected	H0: VE ≤ 30% vs. H1: VE > 30%	0.025	O'Brien-Fleming
Vax003 Phase 3 1999-2003	Efficacy	1; when 50% infections expected	H0: VE ≤ 30% vs. H1: VE > 30%	0.025	O'Brien-Fleming
Step/HVTN 502 Phase 2b 2004-2007	Non-efficacy	1; 30 PP infections[1]	H0: VE ≤ 0% vs. H1: VE > 60%	0.05	Conditional Power < 20%
	Efficacy	1; 30 PP infections	H0: VE ≤ 0% vs. H1: VE > 0%	0.05	Custom
Phambili/ HVTN 503 Phase 2b 2005-2007	Non-efficacy	1; 60 PP infections	H0: VE ≤ 50% vs. H1: VE > 50%	N/A	Conditional Power < 20%
	Efficacy	1; 60 PP infections	H0: VE ≤ 0% vs. H1: VE > 0%	0.05	Custom
RV144 Phase 2b 2004-2009	Harm	Monthly	H0: VE ≥ 0% vs. H1: VE < 0%	0.05	Pocock-type[2]
	Non-efficacy	8; every 6 to 12 months	H0: VE ≤ 0% vs. H1: VE > 50%	N/A	Conditional Power < 10%
	Efficacy	1; 2/3 of follow-up information	H0: VE ≤ 30% vs. H1: VE > 30%	0.025	O'Brien-Fleming

[1]Per-protocol (PP) infections are those diagnosed after the Week 12 visit in volunteers HIV negative at baseline and who received the first two doses of either vaccine or placebo, excluding those who were either diagnosed with HIV infection before or at the Week 12 visit or who violated the protocol on the basis of pre-defined criteria (Buchbinder et al., 2008). The interim analysis was triggered by the 30th PP infection in the primary analysis group of subjects with Adenovirus-5 neutralization titers ≤ 200.

[2]Continuous stopping boundary of Betensky (1998).

DOI: 10.2202/1948-4690.1037

RV144 used O'Brien-Fleming monitoring for reasonably high efficacy (test H0: VE ≤ 30% vs. H1: VE > 30%) at one interim analysis, and also used conditional-power monitoring for non-efficacy at eight interim analyses (every 6-12 months). At each interim analysis the conditional power to reject H0: VE ≤ 0% was calculated under five assumptions about the true VE for the future period of follow-up: (1) VE = 0%, (2) VE = 50%, (3) the current estimate of VE, (4) the current lower 95% confidence limit for VE, and (5) the current upper 95% confidence limit for VE. Stopping was recommended if the conditional power under both assumptions (2) and (3) was less than 10%.

A common feature of the two VaxGen trials and to a lesser extent the Step and Phambili trials is that they either used no monitoring for non-efficacy or conservative monitoring, hence implicitly betting (from a utility perspective) on a reasonable chance for moderate efficacy (Gilbert, 2010), a gamble given the lack of clear scientific rationale (Burton, 2004). In contrast, the proposed design, closer to RV144, uses more aggressive monitoring for non-efficacy, which, had it been applied to the previous three trials that concluded lack of efficacy, would have delivered the conclusion sooner, without incurring an unacceptable risk of prematurely abandoning a promising vaccine candidate such as that identified in RV144. This is illustrated below (see section, "Application of the Proposed Design to Past HIV Vaccine Efficacy Trials").

Summary of Objectives of the Proposed Design.

The previous efficacy trials all evaluated a single vaccine regimen versus placebo. Now that more vaccine regimens are on the near-term horizon for potential efficacy testing, the proposed design evaluates multiple such regimens simultaneously in the same geographic region, sharing a placebo group, with purpose to accelerate the pace of answering key scientific questions about multiple candidate vaccine regimens and hence to accelerate the pace of vaccine development. The primary objective of the design is to expeditiously evaluate VE against HIV infection diagnosed within 18 months of randomization [a parameter we refer to as VE(0-18)] for each vaccine regimen versus placebo, using a sequential monitoring approach fitting to scientific, ethical, and operational considerations. The primary objective focuses on evaluating protection against HIV exposures proximal to the immunization series because the level of protection is plausibly greatest while the vaccine-induced immune responses are at their peak levels, and many immunological parameters wane after the last immunization. The interval 18 months is selected anticipating that the tested vaccine regimens will have HIV envelope protein immunizations at Months 3, 6, and 12. Reasons for counting all infections after randomization rather than only counting infections after a time-point by which full immunity is expected to

accrue include: (1) to assure a fair comparison of vaccine regimens that may have different temporal immunity dynamics; and (2) to obviate the need to select a potentially arbitrary starting time. If issues (1) and (2) are not problematic for the particular vaccine regimens under study, then it would be reasonable to assess VE(6-18) (say) for the primary analysis, albeit as for the analysis of VE(0-18) an intention-to-treat approach is used. Further discussion on this issue is provided in the section, "Intention-to-Treat and Per-Protocol Analysis of VE."

The secondary objectives of the design include: (1) to evaluate durability of vaccine efficacy for each regimen showing reliable evidence for positive VE(0-18); (2) to expeditiously and rigorously evaluate immune correlates of protection if any of the vaccine regimens show reliable evidence for positive VE(0-18); and (3) to compare vaccine efficacy among the vaccine regimens. For secondary objective 1, the durability of vaccine efficacy is evaluated via estimation and inference about the curve $VE(t) = (1 - HR(t))\times 100\%$, where HR(t) is the hazard ratio (vaccine/placebo) of HIV infection diagnosis at time t, ranging from 0 to 36 months post-randomization. For secondary objective 2, immune correlates are evaluated if at least one vaccine regimen shows reliable evidence for positive VE(0-18), with all vaccine regimens included in the assessment, and all available follow-up information included. For secondary objective 3, VE(0-18) is compared among the vaccine regimens, and, if multiple regimens show evidence for positive VE(0-18), durability of VE(t) is compared among the positively efficacious regimens for t ranging between 18 and 36 months.

Secondary objective 1 is important because any vaccine showing positive efficacy proximal to the immunization series merits assessment for the durability of the efficacy, since durability largely influences a vaccine's public health utility (Anderson, Swinton, and Garnett, 1995; Anderson and Garnett, 1996; Abu-Raddad et al., 2007), and, due to data from past HIV vaccine trials showing that many measured vaccine-induced immune responses tend to wane over time, waning efficacy is a ubiquitous concern. Moreover, RV144 motivates this objective, as there was a non-significant trend suggesting that efficacy waned after the first year (Rerks-Ngarm et al., 2009). Secondary objective 2 is important because as soon as there is reliable evidence that a vaccine confers some protective efficacy, it becomes a scientific priority to develop immunological biomarkers that predict the level of VE (one of the "Grand Challenges in Global Health" of the Foundation of the NIH and the Gates Foundation). Such VE-predictive biomarkers would be used as primary endpoints in subsequent Phase I/II trials of refined vaccine candidates, providing a rational basis for iterative improvement of vaccine regimens. There is perception that the one trial showing positive efficacy (RV144) is taking a long time to deliver answers about immune correlates, motivating building planned processes into the proposed design to deliver these answers sooner. Secondary objective 3 is important because head-to-

DOI: 10.2202/1948-4690.1037

head concurrent comparisons of VE within the same trial provides the most rigorous data evidence for decisions about whether and which vaccine regimens to advance to a Phase 3 licensure trial. Furthermore, concurrent assessment of multiple vaccine regimens is expected to shorten the time to a Phase 3 trial compared to separate single-vaccine regimen trials. Additional objectives assess HIV vaccine effects on post-infection endpoints such as viral load; however it is beyond the scope of this article to address approaches for these objectives.

The remainder of this article describes the proposed design and reports on its operating characteristics, with main sections: Description of proposed Phase 2b study design; Sequential monitoring of VE(0-18); Accrual and trial duration for the proposed design implemented in South Africa; Application of the proposed design to past HIV vaccine efficacy trials; Statistical power for assessing an immune correlate of HIV infection; Statistical power for detecting a valuable specific surrogate of protection; Comparing vaccine efficacy among the vaccine regimens; Additional issues; Summary of the proposed design; Other issues of interest that merit further research.

Description of Proposed Phase 2b Study Design

Set-Up of Design.

HIV uninfected volunteers are randomized in equal allocation to a placebo regimen and to between 1 and 3 vaccine regimens, and are followed for up to 36 months for diagnosis of the primary endpoint of HIV infection. While our main interest is in the 2- and 3-vaccine arm trials, we include the 1-vaccine arm trial for comparison. Volunteers receive immunizations at Month 0, 1, 3, 6, and 12 and receive HIV tests monthly starting at Month 0. (A rationale for monthly testing is described below in the section, "Why Monthly HIV Diagnostic Tests?", and has precedent in PrEP trials, e.g., Grant et al., 2010.) We assume that T-cell based prime vaccinations are delivered at the Month 0 and 1 visits (and possibly later visits), and antibody-based envelope protein boosts are delivered at the Month 3, 6, and 12 visits. The trial is event-driven, with the requisite number of HIV infection events in the first 18 months (pooled over a vaccine regimen and placebo) selected such that vaccine regimens with VE(0-18) at least 40% will be identified with high power. Specifically, for each vaccine regimen the design is defined by the characteristic that it has 90% power to reject H0: VE(0-18) $\leq$ 0% if VE(0-18) = 40%, using a 1-sided alpha = 0.025-level log-rank test.

At the end of each vaccine regimen's evaluation, the estimated VE(0-18), 95% CI, and 2-sided p-value, all adjusted for the interim monitoring, will be reported. The reported 95% CI for VE(0-18) is guaranteed to exclude one of the points VE(0-18) = 0% or VE(0-18) = 46%. Thus, the trial will provide reliable

evidence either that VE(0-18) is above 0% or below 46%. For a vaccine regimen that just barely meets the efficacy criterion, the trial will report an estimated VE(0-18) of 30% (Rao-Blackwell adjusted unbiased estimate), 95% CI of 0% to 46%, and 2-sided p-value of 0.05. Each vaccine regimen showing statistically significant positive VE(0-18) will be evaluated for efficacy durability by way of never reaching the non-efficacy boundary described below in the sequential monitoring section. Therefore, for each vaccine regimen the design may be viewed as a two-stage design, wherein vaccine efficacy over 18 months is evaluated in stage 1, and, if and only if positive efficacy is demonstrated, then vaccine efficacy over the extended period of 36 months is evaluated in stage 2. The premise of the two-stage design is that vaccine efficacy is expected to be at least as high proximal to the immunization series as distal. Moreover, the design may be viewed as multiple concurrent two-stage designs, each of which evaluates a vaccine regimen versus placebo, with resource savings accrued via a shared placebo group.

The above approach uses the same type I error rate for each vaccine regimen versus placebo regardless of the number of vaccine arms. Consequently, the risk of any type I errors increases with the number of arms. An alternative design would control the overall type I error rate at 0.025 by using a 1-sided 0.025/M-level test, where M is the number of vaccine arms. This design would require substantially more participants, however, and may be overly stringent, given the trial is not a Phase 3 licensure trial, but rather is a Phase 2b "discovery trial" (Self, 2006; Gilbert et al., 2010) with goals to discover and characterize partially efficacious vaccines and the immune correlates of protection, as well as to provide preliminary comparative assessments of vaccine efficacy.

More Rigorous Evaluation of Immune Correlates via Crossover of Placebo Recipients.

An ultimate goal for HIV vaccine research is development of a measurable characteristic of the vaccine-induced immune response that reliably predicts VE (Plotkin, 2008), a so-called "surrogate of protection (SoP)" or a surrogate endpoint for HIV infection (Qin et al., 2007). In the first tier (least rigorous) of immune correlates assessment, the goal is to discover biomarkers that predict the subsequent rate of HIV infection in the vaccine group(s), named a correlate of risk (CoR). However, a discovered CoR may have no value to predict VE because it may merely correlate with an intrinsic factor such as innate immunity or host genetics that determines whether individuals are more or less naturally resistant to infection (Follmann, 2006; Qin et al., 2007). Recognizing this limitation of the first-tier correlates assessment, statistical approaches have been developed to assess a more rigorous kind of correlate, a second-tier correlate named a SoP,

DOI: 10.2202/1948-4690.1037

defined as a CoR that reliably predicts VE, otherwise known as a partially valid surrogate endpoint for HIV infection (Follmann, 2006; Gilbert and Hudgens, 2008; Gilbert, Qin, and Self, 2008; Qin et al., 2008; Wolfson and Gilbert, 2010). Assessment of a second-tier correlate requires predicting the 'counterfactual' values of the vaccine-induced immunological biomarker for a subset of placebo recipients. As proposed by Follmann (2006), these predictions may be derived based either on (1) Modeling the relationship between baseline subject characteristics and the biomarker (baseline immunogenicity predictor approach, BIP), and/or on (2) Crossing over a subset of uninfected placebo recipients to the vaccine group and directly measuring their vaccine-induced biomarkers (crossover placebo vaccination approach, CRPV). For a given biomarker the second-tier methods yield an estimate of the "VE curve," VE(s), which describes how VE changes with the level of the vaccine-induced biomarker. A biomarker valuable for guiding refinement of a vaccine regimen showing some efficacy in the trial will have VE(s) varying widely across levels of s, for example VE(s) will be near 0 for s near 0 (e.g., "negative" immune response) and VE(s) will be large (e.g., 70-90%) for a large immune response s.

We believe both the BIP and CRPV approaches merit use in the proposed efficacy trial design. In particular, if at least one vaccine regimen demonstrates positive VE(0-18), then we propose to cross-over random samples of uninfected placebo subjects to each vaccine regimen that is advanced to Stage 2. While various time-points of cross-over could be considered, the default approach [originally proposed by Follmann (2006)] is appealing, wherein cross-over occurs at the last study visit (the Month 36 visit in our prototype design). The crossed-over subjects are immunized on the same schedule as when they entered the trial, which is necessary for credibility of the 'time-constancy' assumption, which states that for crossed-over placebo subjects, the measured immune response is the same as it would have been had it been measured approximately three years earlier on the same schedule relative to the first vaccination.

An alternative approach would cross-over subjects at various times starting at the Month 18 visit. The advantage of this approach is that availability of immune response data at multiple cross-over points would facilitate diagnostic tests of the time-constancy assumption mentioned above (Follmann, 2006). However, the disadvantage is that no post-crossover information from these subjects would be used for the analysis of VE(t) for $t > 18$ months. That is, in analyses of VE(t) for $t > 18$ months, the crossed over subjects would be counted in the placebo group only and would be censored at the time of crossover. While this crossover would have no effect on the evaluation of VE(0-18), it would attenuate the statistical power for evaluating VE(t) for $t > 18$ months. More research is needed to determine the optimal fraction of placebo recipients to cross-over, balancing the needs of assessing an immunological surrogate endpoint with

the needs of assessing durability of vaccine efficacy. The default approach that waits until the Month 36 visit to cross-over placebo subjects is appealing given the importance of maximizing power for assessing waning vaccine efficacy. It is also appealing for simplifying the study, avoiding the complexity of multiple random cross-over times.

Sequential Monitoring of VE(0-18)

Sequential Monitoring for Non-Efficacy.

For each vaccine regimen, the proposed design monitors for non-efficacy at several analyses at evenly spaced numbers of infections diagnosed within 18 months pooled over the vaccine group and the placebo group. We require the number of infections n_1 triggering the first interim analysis to be at least 37% of the maximum information, to ensure that a decision to complete a vaccine's evaluation has a minimum level of data support (Freidlin, Korn, and Gray, 2010). In particular, following the suggestion of Freidlin, Korn, and Gray (2010), 37% of the maximum infections was chosen as the first point because, if the estimated VE(0-18) is less than or equal to zero, then the unadjusted/nominal 95% confidence interval for VE(0-18) will exclude the design alternative VE(0-18) = 40% for which the design has 90% power to detect. Because the proposed design requires a maximum of 176 infections within 18 months, this rule equates to the earliest non-efficacy interim analysis taking place at the 65^{th} infection. This approach is an informal way to ensure that, if the reported point estimate indicates non-efficacy, then there will be enough precision about the inference to reliably rule out the design alternative of 40% vaccine efficacy. Completing a vaccine regimen's evaluation prior to this point would be problematic because, given the wide confidence interval, some interpreters of the published result may not be convinced that low efficacy at best was reliably established. This could raise thorny questions about whether additional efficacy trials would be needed, counter to an objective of the design to provide sufficiently definitive evidence about low efficacy such that another efficacy trial would not be needed. Note that with the proposed design the reported monitoring-adjusted 95% confidence interval for VE(0-18) for a weeded-out vaccine regimen is guaranteed to lie below 46%.

To ensure that vaccines with weak efficacy during the ramp-up period of immunity (while the immunizations are being received) but substantial efficacy later are not prematurely weeded out (i.e., the reported 95% confidence interval for VE(0-18) does not lie above 0) based on inter-current infections, we define n_1 as the maximum of 65 and the first infection diagnosis event within 18 months such that at least 20% occurred after the ramp-up period (i.e., post-Month 6 visit).

DOI: 10.2202/1948-4690.1037

Below we show that with this approach the design has less than 20% risk of incorrectly weeding out a vaccine with VE(0-18) = 40% and halved VE during the pre-defined ramp-up period of 0-6 months (see the entry Avg VE(0-18) = 40% in Table 2 Scenario B, where the estimated probability of weed-out is 0.008 + 0.179 = 0.187). If VE(0-18) = 40%, the infection count in the first 18 months when 20% occur post 6 months has median 70, inter-quartile range 58−82, and 10^{th}−90^{th} percentiles 49−92. If VE(0-18) = 0%, the infection count when 20% occur post 6 months has median 79, inter-quartile range 68−92, and 10^{th}–90^{th} percentiles 58−103.

An alternative approach would determine n_1 based on a minimal percentage of person-time at-risk occurring after the ramp-up period. This approach is motivated by two potential down-sides of the infections-based approach: n_1 has relatively high variance, because it depends on the unknown HIV incidence in each study arm; and n_1 depends on the relative level of VE(0-18) during and after the ramp-up period, such that the timing of n_1 could indirectly leak information on vaccine efficacy to individuals outside of the DSMB. However, the infections-based approach has the advantage of defining the milestone based on the information scale for a survival analysis, whereas the person-time at-risk approach could start the analysis based on a small number of infections. Therefore we select the infections-based approach, and in limited simulations we found that the two approaches had very similar false-weed-out rates concordant within 1%. Another potential approach would monitor for non-efficacy at evenly spaced numbers of total infections, and use a weighted log-rank statistic that down-weights infections occurring during the ramp-up period. While this approach could be configured to give satisfactory operating characteristics, it is not clear that this weighting scheme would be desirable for assessing positive efficacy, such that different test statistics may be warranted for testing the two alternative directions. In contrast, the selected approach allows a symmetric monitoring design with the un-weighted log-rank test used for testing in both directions (Emerson and Fleming, 1989).

Table 2. Probabilities (×100%) that the Trial Will Report Each of the Results Efficacy, Potential Harm, Non-Efficacy, and High Efficacy: Scenario (A) [Time-Constant VE(0-18)]*; Scenario (B) [Halved VE in First 6 Months]*.

Scenario A [Time-Constant VE(0-18)]								
Avg VE (0-18)	**Avg RR (1-18)**	**Eff**	**Harm**	**Harm Time**	**Non-Eff**	**Non-Eff Time**	**High-Eff**	**High-Eff Time**
-	3.0	0.0	100.0	6.8 (4.9-9.2)	0.0	14.1 (14.1-14.1)	0.0	- (---)
-	2.5	0.0	99.3	7.6 (5.5-10.5)	0.7	12.8 (11.8-13.9)	0.0	- (---)
-	2.0	0.0	88.9	9.2 (6.2-12.3)	11.1	13.1 (12.3-14.2)	0.0	- (---)
-	1.5	0.0	42.9	10.1 (6.4-13.0)	57.1	13.4 (12.5-14.8)	0.0	- (---)
0%	1.0	2.7	4.2	8.6 (6.1-12.4)	93.0	14.6 (13.1-17.8)	0.0	- (---)
20%	0.8	30.5	1.2	7.4 (5.9-10.5)	68.3	16.7 (13.7-22.0)	0.0	- (---)
30%	0.7	63.0	0.6	7.0 (5.8-10.2)	36.4	18.1 (14.2-23.4)	0.0	17.0 (17.0-17.0)
40%	0.6	89.5	0.2	6.7 (5.8-9.2)	9.9	19.5 (14.5-24.8)	0.4	20.0 (15.5-21.3)
50%	0.5	94.8	0.1	6.8 (5.8-9.1)	1.0	18.1 (14.2-24.8)	4.1	21.0 (16.8-29.7)
60%	0.4	68.1	0.0	6.9 (5.9-8.9)	0.0	20.0 (15.8-21.3)	31.9	22.7 (17.4-29.8)
70%	0.3	14.5	0.0	- (---)	0.0	- (---)	85.5	22.4 (17.1-29.6)
80%	0.2	0.2	0.0	- (---)	0.0	- (---)	99.8	18.8 (13.4-23.8)

DOI: 10.2202/1948-4690.1037

Scenario B [Halved VE in First 6 Months]								
Avg VE (0-18)	**Avg RR (1-18)**	**Eff**	**Harm**	**Harm Time**	**Non-Eff**	**Non-Eff Time**	**High-Eff**	**High-Eff Time**
-	3.0	0.0	96.0	8.6 (6.1-11.2)	4.0	12.3 (11.1-12.9)	0.0	- (---)
-	2.5	0.0	84.9	9.5 (6.3-12.1)	15.1	12.5 (11.7-13.3)	0.0	- (---)
-	2.0	0.0	57.5	10.1 (6.5-12.6)	42.5	12.7 (12.1-13.7)	0.0	- (---)
-	1.5	0.0	22.5	10.2 (6.4-12.9)	77.5	13.2 (12.4-14.4)	0.0	- (---)
0%	1.0	2.7	4.2	8.6 (6.1-12.4)	93.0	15.8 (13.6-21.2)	0.0	- (---)
20%	0.8	25.9	1.9	7.7 (6.0-10.9)	72.3	15.8 (13.6-21.2)	0.0	- (---)
30%	0.7	54.3	1.2	7.3 (5.9-10.8)	44.4	16.1 (13.8-22.0)	0.0	- (---)
40%	0.6	81.3	0.8	7.0 (5.9-9.8)	17.9	15.9 (13.9-21.5)	0.1	21.3 (19.3-24.4)
50%	0.5	92.5	0.6	6.8 (5.8-8.9)	4.6	15.6 (13.8-18.4)	2.3	28.5 (20.2-29.9)
60%	0.4	72.4	0.3	6.6 (5.4-8.7)	0.8	15.3 (14.1-16.7)	26.5	29.1 (21.2-29.2)
70%	0.3	16.2	0.2	6.4 (5.1-7.4)	0.1	15.3 (14.1-16.3)	83.5	25.2 (21.4-29.8)
80%	0.2	0.2	0.2	6.6 (5.0-7.9)	-	- (---)	99.7	24.1 (18.2-29.3)

*****Efficacy (Eff in the third column) is the result that VE(0-18) > 0% with reported 95% confidence interval lying above 0%. Potential Harm (Harm) is the result that the potential harm boundary is reached. Non-efficacy (Non-Eff) is the result that the reported 95% confidence interval for VE(0-18) does not lie above 0%; this occurs if the non-efficacy boundary is reached at an interim analysis or the final analysis for assessing VE(0-18). High efficacy (High-Eff) is the result that the reported 95% confidence interval for VE(0-18) lies above 50%. The Times for the various events show the 50^{th} (10^{th}-90^{th}) percentiles of the number of months until the event is reached.

Once n_1 is determined for a vaccine regimen, the timing of the subsequent analyses for evaluating non-efficacy are defined to satisfy all of the criteria: (1) achieve 90% power to detect VE(0-18) = 40%; (2) use as many analyses as possible up to nine; and (3) evenly space the interim analyses at intervals of at least five infections. Based on these criteria all 9 analyses are scheduled if and only if $n_1 \leq 127$. In the case that VE(0-18) = 40%, there is a > 99.9% chance that all 9 analyses will be scheduled.

Several stopping boundaries were considered, and we select the "P = 0.6 stopping boundary" (Emerson and Fleming, 1989), which is slightly less aggressive than the Pocock (1977) boundary for early stopping, chosen to balance the objectives of rapidly weeding out non-efficacious vaccines and protecting against the false weed-out error mentioned above. The operating characteristics of the non-efficacy monitoring plan are described below (in the section, "Accrual and Trial Duration for the Proposed Design Implemented in South Africa"). Based on expectations for accrual, HIV incidence, and dropout for the proposed design implemented in South Africa (described below) for a vaccine regimen with VE(0-18) = 40%, the median value of n_1 is 75, in which case there are 9 analyses with the last one occurring at $n_{max} = 176$ infections. For $n_1 = 75$, Figure 1 shows the non-efficacy stopping boundary on the scale of the nominal estimated hazard ratio over 18 months [HR(0-18)]; the boundary is reached as soon as an interim estimate of HR(0-18) goes below the boundary.

The Lan-DeMets (Lan and DeMets, 1983) implementation of the stopping boundary is used so as to allow flexibility in the timing and number of analyses. For validity this approach requires that the future analysis times are selected to be independent of the current estimate of VE(0-18) (Betensky, 1998). Given that the interim analyses are fairly frequent and it is not pressing to detect a non-efficacy signal a few months earlier, this assumption is acceptable.

Should Sequential Monitoring for any Vaccine Efficacy be Performed?

A goal of the trial design is to facilitate expeditious assessment of immune correlates for all vaccines showing some efficacy. One technique for helping achieve this is sequential monitoring for positive efficacy (test H0: VE ≤ 0% vs. H1: VE > 0%), and to initiate the immune correlates assessment (i.e., commence measuring the pre-specified candidate immune correlates from infected vaccine-group subjects and from frequency matched uninfected vaccine-group subjects) when the efficacy signal is reached. However, a potential problem with this approach is that, in order to initiate the immune correlates analysis, many individuals would need to know that the positive efficacy signal is achieved (e.g., lab personnel and the managers of specimen processing and shipments), and it

DOI: 10.2202/1948-4690.1037

may be difficult to ensure that dissemination of this knowledge would not damage study conduct (Ellenberg, Fleming, and DeMets, 2002).

Given this potential problem, we expect that a simpler approach may be more effective, wherein for each vaccine regimen the immune correlates assessment is automatically initiated 9-12 months before all of the information is available for evaluating VE(0-18) (i.e., when the last enrolled participant has 6-9 months of follow-up). The immunologic work is only initiated for vaccine regimens that did not earlier reach the non-efficacy boundary, for which some positive efficacy is likely. Vaccines not hitting the non-efficacy boundary will have estimated VE(0-18) at least 20-25% (as demonstrated in Figure 1: for example if the non-efficacy boundary is not reached at 151 infections than the estimated hazard ratio must be less than 0.78, i.e., VE(0-18) must exceed 22%), supporting at least low-level efficacy that would make a correlates analysis worthwhile. This approach would straightforwardly maintain confidentiality, as no one but the independent statistician(s) and DSMB would know whether reliable evidence for positive efficacy had been achieved. Moreover, the known date for a go/no-go decision would help study personnel prepare for the correlates analyses, and this approach may provide results sooner than the interim monitoring-based approach, because the analysis may begin before an efficacy signal would be detected.

Sequential Monitoring for High Efficacy.

While it is unlikely that the prime–boost HIV vaccine regimens under preparation for efficacy testing will confer high levels of protective efficacy, for scientific and ethical reasons it may be prudent to monitor for this event, which, if detected, would lead to un-blinding of participants and reporting of the result (see section "Timing of Reporting of Results and of Un-blinding" for additional discussion on un-blinding). We define "high enough efficacy to warrant un-blinding" as reliable evidence that VE > 50%, operationalized by a log-rank test rejecting H0: VE ≤ 50% vs. H1: VE > 50% at 1-sided 0.025-alpha level. The proposed design tests H0 at three interim analyses, at evenly spaced numbers of arm-pooled infections diagnosed between 0 and 18 months with final number fixed at the median n_{max} if VE(0-18) = 50% (176 in the prototype design). An O'Brien-Fleming stopping boundary is used so as to require strong early evidence for VE(0-18) > 50% (shown in Figure 2). As for the non-efficacy monitoring, the Lan-DeMets (1983) implementation is used so as to allow flexibility in the timing and number of

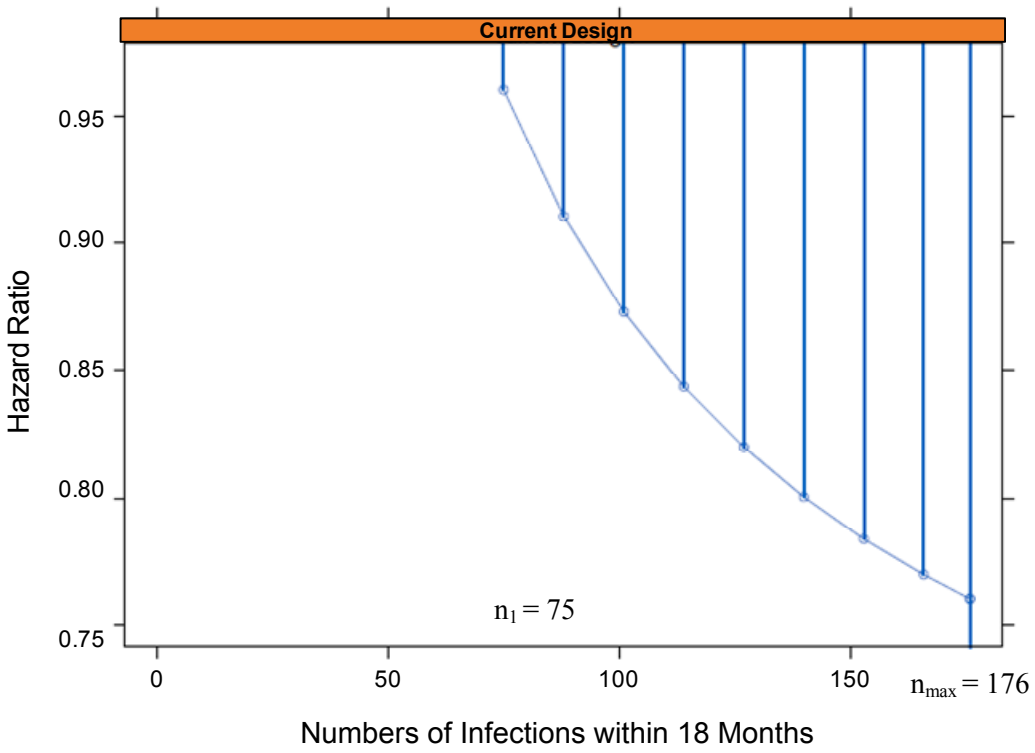

Figure 1. P = 0.6 non-efficacy boundary comparing a vaccine regimen versus placebo, for the scenario where the first interim analysis occurs at n_1=75 infections diagnosed within 18 months. If the final analysis at n_{max} infections is reached before the boundary is crossed, then the reported 95% confidence interval for VE(0-18) will be above 0%.

analyses. Unlike the non-efficacy monitoring, if the VE(0-18) estimate is near the boundary then the DSMB may request an additional interim analysis, in which case the Lan-DeMets implementation could be swapped with Betensky's (1998) continuous stopping boundary to ensure valid type I error control. Figures 3 and 4 show the power curve for detecting VE(0-18) > 50% and the cumulative probabilities of reaching the high efficacy boundary by the four analysis times. For vaccines with VE(0-18) in the range 0-50%, this monitoring has negligible impact on the operating characteristics of the design.

DOI: 10.2202/1948-4690.1037

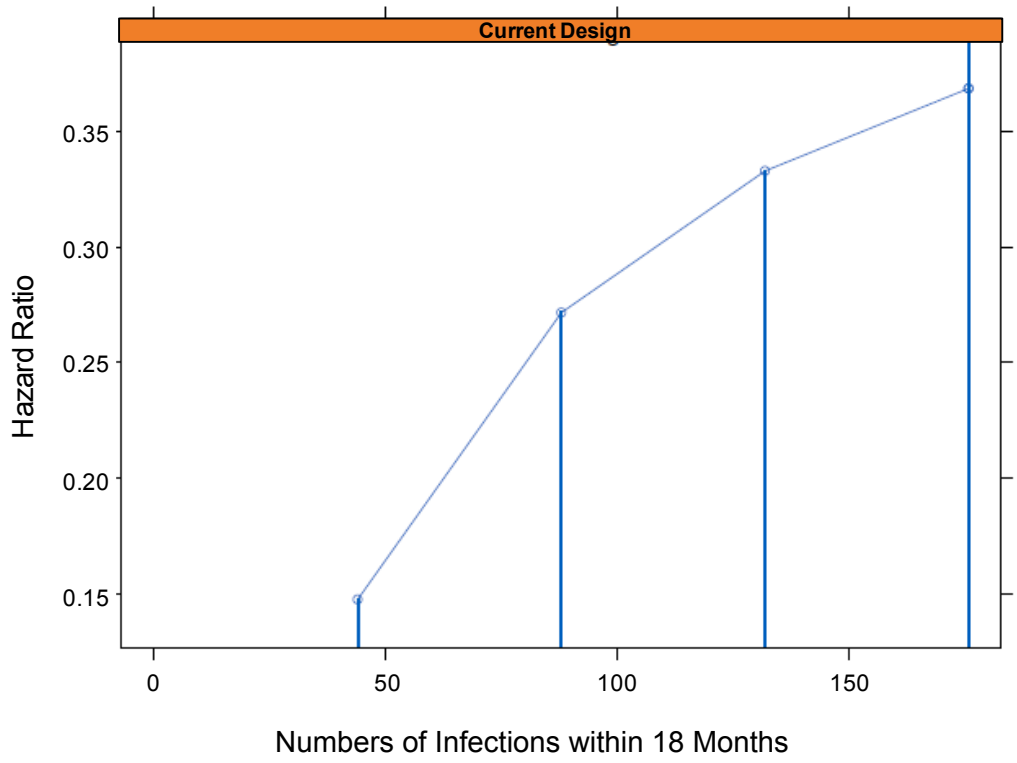

Figure 2. O'Brien-Fleming high-efficacy boundary comparing a vaccine regimen versus placebo, for the scenario where the first non-efficacy interim analysis occurs at n_1=75 infections diagnosed within 18 months such that n_{max} = 176 infections (and the first high-efficacy interim analysis starts at 44 infections).

Sequential Monitoring for Potential Vaccine Harm.

Given the potential vaccine-enhancement of HIV infection risk observed in the Step trial (Buchbinder et al., 2008), it is prudent to closely monitor for VE(0-18) < 0%. To provide maximally close monitoring for each vaccine regimen, the proposed design performs interim analyses after every HIV infection event diagnosed between 0 and 18 months ranging from the 7^{th} to the $n_1{}^{th}$ (pooled over a vaccine regimen and placebo). Similar monitoring was performed by Heyse et al. (2008) in a rotavirus vaccine trial and is being used in an ongoing HIV Vaccine Trials Network (HVTN) trial. Such "continuous" monitoring is performed by an un-blinded statistician (independent from the protocol statisticians) who observes

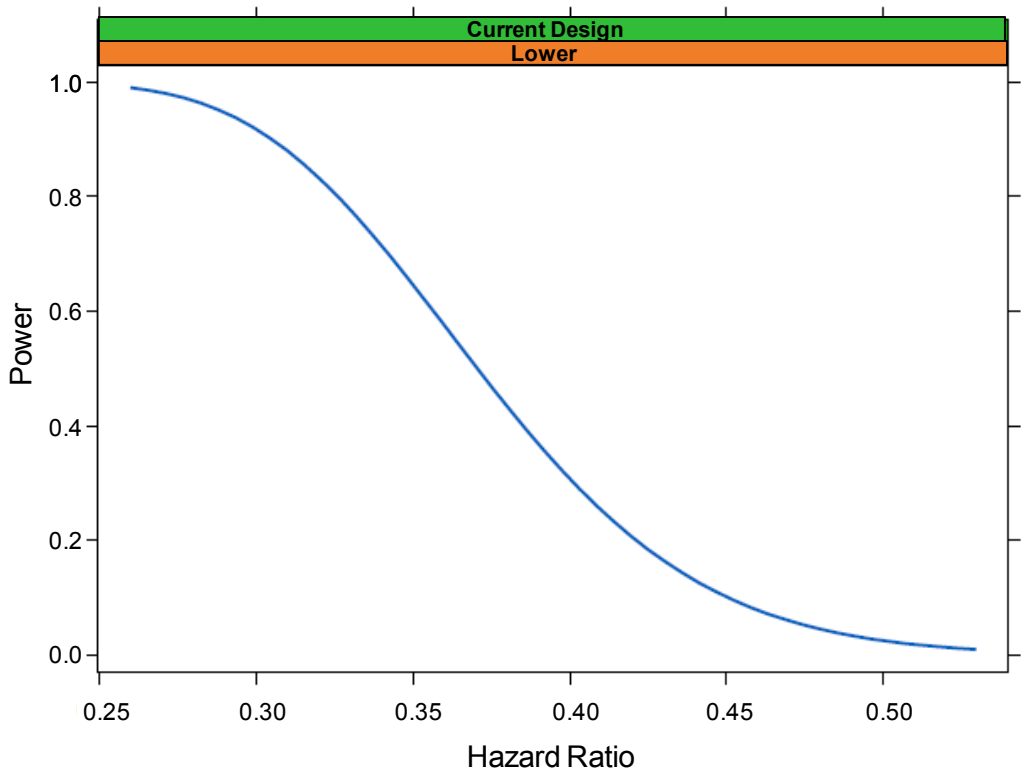

Figure 3. Statistical power for rejecting H0: VE(0-18) ≤ 50% in favor of H1: VE(0-18) > 50% as a function of the true VE(0-18) (1-sided 0.025-level test). Here VE(0-18) = [1 – hazard ratio over the first 18 months] × 100%.

whether, after each confirmed HIV infection event, the stopping boundary is reached. The monitoring applies exact one-sided binomial tests of H0: $p \leq 0.5$ versus H1: $p > 0.5$, where p is the probability that an infected subject was assigned to the vaccine group. Each test is performed at the same pre-specified nominal/unadjusted alpha-level, chosen based on simulations such that, for each vaccine regimen, the overall type I error rate by the 99th arm-pooled infection (i.e., the probability that the potential-harm boundary is reached when the vaccine is actually safe, $p = 0.5$) equals 0.05. The number 99 is selected because, under the null [VE(0-18) = 0%], there is a 90% chance that the non-efficacy monitoring would commence by the 99^{th} infection in the first 18 months ($n_1 \leq 99$). If n_1 is below 99, then the effect is that less than 0.05 overall type I error rate is spent; for

DOI: 10.2202/1948-4690.1037

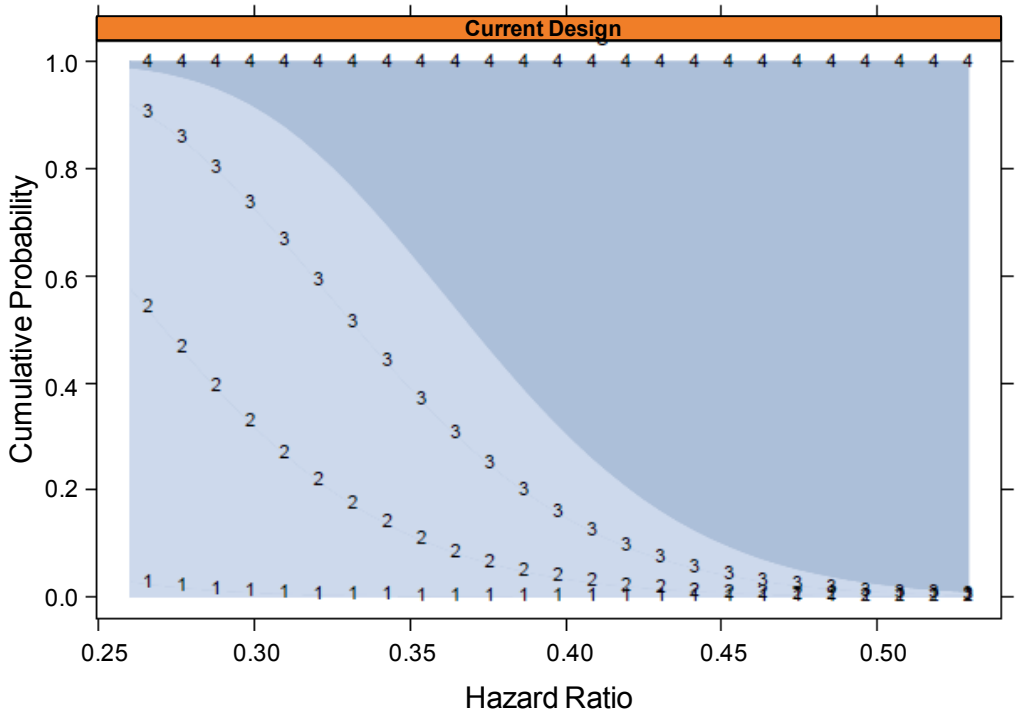

Figure 4. Cumulative probability of reaching the high-efficacy boundary by the first, second, third, and final analyses (lines demarked by 1, 2, 3, 4, respectively), as a function of the true HR(0-18), for the scenario where the first non-efficacy interim analysis occurs at n_1=75 infections diagnosed within 18 months. For example, if the true HR(0-18) = 0.3 [i.e., VE(0-18) = 70%], then there is a 30% chance to reach the boundary by the second analysis and a 75% chance to reach the boundary by the third analysis.

example, with n_1 = 75 the overall error rate is about 0.045. The impact on the potential harm monitoring is a slight loss of power to detect a harmful vaccine. If n_1 exceeds 100, then the tests continue to be applied (using the same critical value), which slightly increases the overall type I error rate during the trial (estimated at 0.0532 for n_1 = 120 and at 0.0558 for n_1 = 140).

Figure 5 shows the potential-harm stopping boundary, and the upper rows of Table 2 describe the power of the monitoring plan to reach the boundary under different HRs > 1. For example, for a vaccine with time-constant HR = 1.5 (50% elevation in the infection hazard rate over the first 18 months) there is a 43% chance to stop before the n_1^{th} infection, and the median stopping time is 10.1 months (Table 2 Scenario A). In addition, if the vaccine doubles the risk of

infection (HR = 2.0), there is a 89% chance to stop before the n_1^{th} infection, and the median stopping time is 9.2 months (Table 2 Scenario A).

The potential-harm boundary is only defined out to the n_1^{th} infection because the non-efficacy boundary serves the function to stop harmful vaccines at all later infection counts, in fact much more aggressively than would an extended harm-boundary [e.g., a vaccine with estimated VE(0-18) < -2% at the first non-efficacy interim analysis is guaranteed to reach the stopping boundary]. An alternative approach to monitoring for potential vaccine-harm would use a repeated generalized likelihood ratio test (Siegmund, 1985, Chapter 4; Wald, 1947) applied at the same analysis times, with potential advantages that the procedure is approximately asymptotically efficient and the critical value is obtained analytically. The boundaries (based on the binomial proportion p) are almost identical to the exact binomial-test-based boundaries (not shown).

The potential-harm monitoring is not intended to reliably establish harm [i.e., VE(0-18) < 0%], as a vaccine regimen could meet the boundary and the reported 95% confidence interval for VE(0-18) would include 0% (although the 90% confidence interval, if constructed correspondent to the testing procedure, would exclude 0%). Rather, the objective is to apply extra caution and prudence for a prevention trial that enrolls healthy volunteers. More discussion may be needed to determine whether this degree of caution is warranted, given that an error to reach the potential-harm boundary for a truly safe vaccine [with VE(0-18) ≥ 0%] may cause undue damage to the HIV vaccine field.

DOI: 10.2202/1948-4690.1037

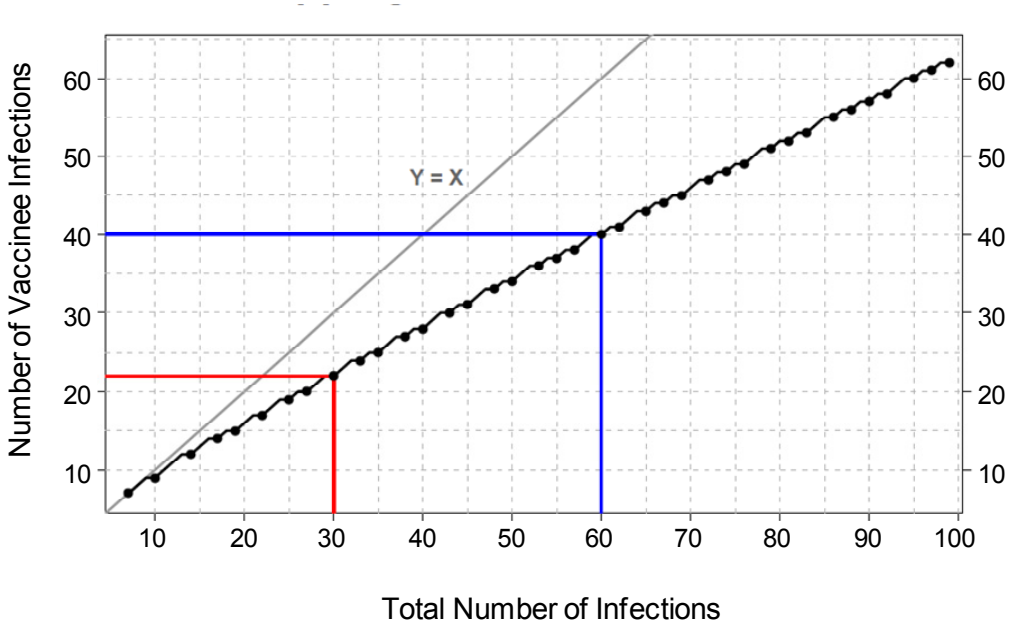

Figure 5. Potential-harm boundary comparing a vaccine regimen versus placebo. For each infection diagnosed within 18 months of randomization from the 7th to the 99th (x) the boundary is reached if at least y of the infections were assigned to the vaccine arm. The red and blue lines illustrate the analyses at the 30th and 60th infection, with stopping boundary ≥ 22 and ≥ 40 of the infections in the vaccine arm, respectively.

Operational Considerations for the Timing of Interim Analyses.

On the surface, the timing of interim analyses is complex, because it is separately determined for each vaccine regimen based on the rate of infection event, and differs across the monitoring types. However, for the purpose of continuous potential-harm monitoring in the current HVTN trial (HVTN 505), the HVTN developed an effective procedure for rapid adjudication of HIV infection events and for automatic generation of monitoring reports after each confirmed infection event. The existence of this system makes straightforward the accommodation of multiple monitoring schedules. In particular, after each adjudicated infection the un-blinded statistician creates the routine reports and notes whether any interim analyses are due, and, if so, whether any boundaries are reached. Reaching a boundary prompts the statistician to immediately notify the DSMB, which may request a more complete analysis that includes secondary endpoints, collated into a report for the next DSMB meeting. Based on this report the DSMB will make

recommendations about continuing or stopping each vaccine regimen. Given the complexity of the pros and cons of continuing or stopping each vaccine regimen, the DSMB might be asked to report to a predetermined Oversight Group as well as the Team given the complexity and implications that may be beyond the DSMB's immediate purview (Ellenberg, Fleming, and DeMets, 2002; Emerson, 2006; Fleming, 2006; Emerson and Fleming, 2010). The Oversight Group includes critical stake-holders, such as representatives of the sponsor, the vaccine manufacturer, and the research group conducting the efficacy trial.

Given that an effective system for accurately and rapidly identifying HIV infection endpoints is in place, it would also be feasible to use continuous monitoring for all of the monitored events, although more work would be needed to delineate the pros and cons.

Monitoring for Operational Futility.

Achieving the primary objectives in a timely manner requires sufficiently high rates of accrual, HIV incidence, and ascertainment of the primary endpoint of HIV infection. Therefore, the design monitors these three types of data, and at each DSMB meeting presents an analysis of the projected time until the final analysis, with a prediction interval to assess uncertainty in the projection. Because the projection method is based only on blinded data (pooling over study groups), and the guidelines for what outcomes constitute operational futility are pre-specified and pre-vetted with various stake-holders including the sponsor, vaccine-manufacturers, DSMB, and experts in the field, the operational futility monitoring poses minimal risk to study integrity and is widely used in clinical research. Developing a statistical approach to projecting operational futility was an important aspect of designing the current small Phase 2 HIV vaccine efficacy trial (HVTN 505). While we consider it beyond the scope of this manuscript to describe details of potential operational futility monitoring plans, it is important to note that such monitoring would be employed.

Accrual and Trial Duration for the Proposed Design Implemented in South Africa

Because the proposed design is event-driven, the required number of subjects to enroll and the anticipated trial duration are estimated based on anticipated rates of accrual, HIV incidence in the placebo group and dropout. We illustrate these calculations for South Africa, where based on HVTN experience we assume: uniform accrual over a 12 month period, with halved accrual in the first 3 months; 4% annual HIV incidence in the placebo group; and 5% annual dropout. Ten thousand trials were simulated, assuming the HIV incidence and dropout rates

DOI: 10.2202/1948-4690.1037

have Poisson distributions, and assuming each vaccine regimen has VE(0-18) = 50% with either (A) constant VE throughout 0−18 months or (B) constant VE throughout 0−6 months at VE(0-6) = 30% and constant VE throughout 6−18 months at VE(6-18) = 60%, both scenarios for which early stopping is unlikely and hence a relatively large sample size N is needed, which should be planned for. In particular, N = 2150 is chosen as the number enrolled (per arm) such that for each vaccine regimen, under either Scenario A or B, there is at least an 85−90% chance that at least n_{max} = 176 infections will be diagnosed within 18 months (combined across the vaccine and placebo arms). In particular, with N = 2150 per group, there is probability 0.025, 0.10, 0.25, 0.50, and 0.75 of 165, 173, 181, 189, and 198 infections diagnosed within 18 months, respectively, and this result is the same for Scenarios A and B. For N = 2000 per group these numbers decrease by about 12 while for a sample size of N = 2250 these numbers increase by about 10.

Table 3. Projected Accrual Rate and Number Enrolled for the Proposed Design.

Number of Study Arms (with One Placebo Arm)	Accrual Per Week During 52 Week Accrual period[1]		Number Accrued Per Study Arm[2]	Total Accrued (N)
	Initial 13 Weeks	Subsequent 39 Weeks		
2	47	95	2150	4300
3	71	142	2150	6450
4	95	189	2150	8600

[1]These accrual rates lead to full accrual at 12 months since the first subject is enrolled, such that the maximum trial duration is 48 months.
[2]Equal allocation of subjects to the study arms.

Based on the 10,000 simulated trials under Scenario A using the sample sizes and accrual rates shown in Table 3, Figure 6a-c shows distributions of the trial duration under different values for true VE(0-18), for trials with 1, 2, or 3 vaccine regimens. Worthless vaccines [with VE(0-18) = 0%] are weeded out (i.e., reach the non-efficacy boundary) within 17 months with 50% probability, and within 20 months with 99% probability (Figure 6a). If a vaccine regimen has VE(0-18) ≥ 40%, then there is at least 82% probability that the regimen will be fully evaluated to the maximum duration of 48 months (Figure 6a). For a trial with 2 or 3 vaccine regimens each with VE(0-18) ≥ 40%, there is at least 93% probability that the trial will reach the full 48 months (Figure 6b,c). Furthermore, if a vaccine regimen has low efficacy in the range 20-30%, then both events of weed-out and continuation to the end are fairly likely. For example, if all vaccine regimens have VE(0-18) = 30%, then a trial with 1, 2, and 3 vaccine regimens

will reach the full 48 months with probability approximately 55%, 67%, and 80% (black dashed lines in Figures 6a-c).

Table 2 shows corresponding information on the probabilities that each individual vaccine regimen reaches each type of stopping boundary, and, if so, how long it takes. Our goal is to have high probability of weeding out vaccines with 0-15% efficacy and low probability of weeding out vaccines with at least 40-50% efficacy. Under either Scenario A and B there is a very low risk that the trial would report a 50% efficacious vaccine as non-efficacious, whereas for a 40% efficacious vaccine this risk is about 10% if VE(0-18) is constant and about 19% if VE(0-18) is halved in the first 6 months (Table 2).

For the design with two vaccine arms, the first with constant VE(0-18) = 20% and constant VE(18-36) = 10% and the second with constant VE(0-18) = 50% and constant VE(18-36) = 25%, Figure 7 shows the distributions of the number of HIV infections diagnosed during the time-intervals 0-36 months, 0-18 months, 0-6 months, 6-18 months, and 18-36 months. The distributions have many outliers due to every type of monitoring bound being reached with at least small positive probability.

DOI: 10.2202/1948-4690.1037

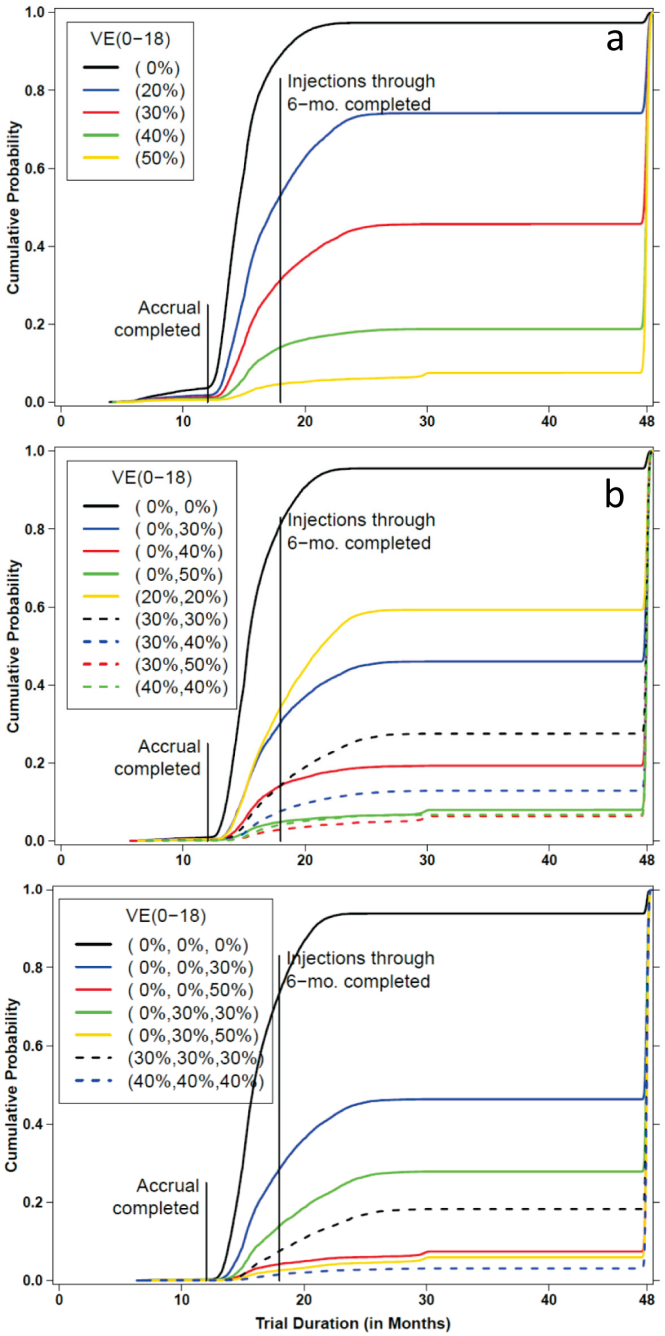

Figure 6. Distributions of the total trial duration for trials with (a) 2, (b) 3, and (c) 4 study groups, all with one placebo group. (a) 1 vaccine regimen versus placebo. (b) 2 vaccine regimens versus placebo. (c) 3 vaccine regimens versus placebo. For trials with multiple vaccine groups, a trial completes once all of the vaccine groups reach the end of evaluation. The calculations for this figure assume the true VE(0-18) is constant over time.

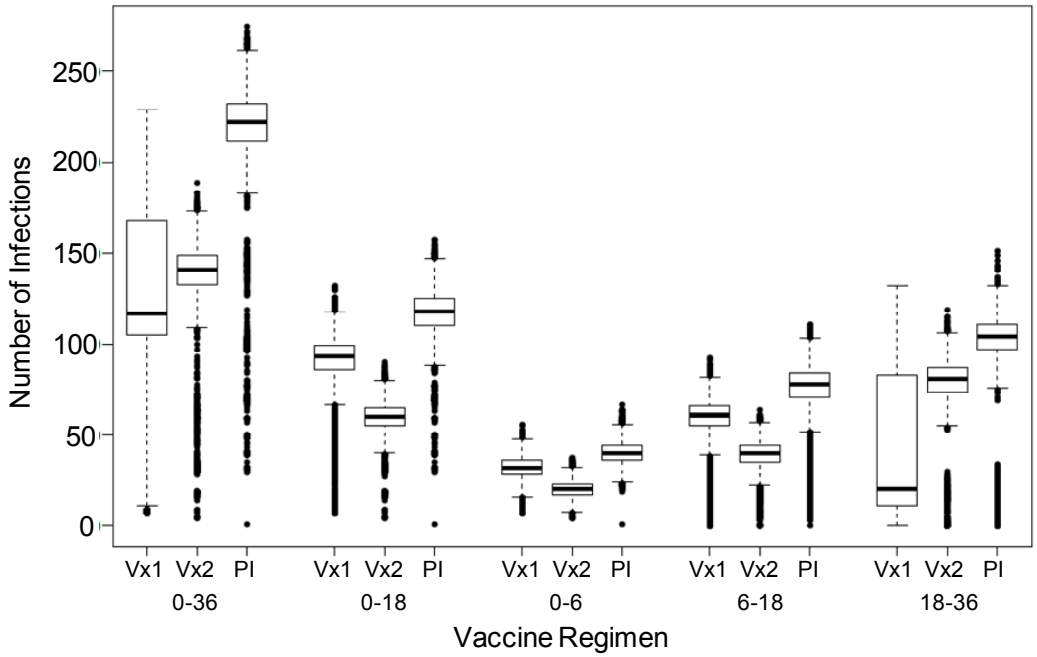

Figure 7. Distributions (box-plots) of the number of infections diagnosed in different time-intervals 0-36 months, 0-18 months, 0-6 months, 6-18 months, and 18-36 months, for the placebo arm and two vaccine arms. The first vaccine has constant VE(0-18) = 20% and constant VE(18-36) = 10% and the second vaccine has constant VE(0-18) = 50% and constant VE(18-36) = 25%.

Application of the Proposed Design to Past HIV Vaccine Efficacy Trials

We applied the proposed 2-arm version (one vaccine versus placebo) of the design to the Vax004, Vax003, Step, and RV144 data-sets. The needed results for determining whether and when any boundaries are crossed are the number of infections triggering the first interim analysis for non-efficacy (which turns out to be 65 for each trial), the infection split after each infection in 0−18 months from the 7^{th} to the 64^{th} (for potential harm monitoring), the estimated HRs over 0−18 months at each of the interim analyses for non-efficacy starting at the 65^{th} infection, and the estimated HRs over 0−18 months at each of the interim analyses for high efficacy. Because Vax003, Step, and RV144 evenly randomized subjects to vaccine or placebo, the proposed boundaries could be directly applied [for Step we analyzed all subjects instead of focusing on the primary analysis cohort− the subgroup with low neutralization levels (≤200) to Adenovirus 5]. However, Vax004 used a 2:1 vaccine: placebo allocation, precluding their direct application. To allow direct application to Vax004, we created 10,000 1:1 allocation data-sets by increasing the placebo group by 33% and decreasing the vaccine group by

DOI: 10.2202/1948-4690.1037

33%, the former achieved by random sampling the placebo group data with replacement and the latter achieved by random sampling the vaccine group data without replacement. All of the needed statistics for checking boundary-crossings were then computed for each of the 10,000 data-sets. A single data-set for analysis was then constructed by using for each statistic the median of the 10,000 statistics; for example, for non-efficacy monitoring, at each interim analysis we use the median of the 10,000 HR(0−18) estimates as the HR(0−18) estimate. This procedure approximately represents the real Vax004 trial because it preserves the expected vaccine efficacy at all time-points and preserves the total statistical information in the data (expected total number of infections).

For each trial, we evaluated infections diagnosed during the first 18 months to determine the time of the first non-efficacy interim analysis and hence n_1 and n_{max}. Hazard ratio estimates were computed (with the proportional hazards model) at each scheduled interim analysis, and were compared to the non-efficacy boundary. In addition, 1-sided Fisher's exact test p-values were compared to the potential-harm boundary after each infection diagnosed within 18 months starting at the seventh, and hazard ratio estimates were compared to the high-efficacy boundary at the scheduled high-efficacy interim analyses. For each trial, SeqTrial software was used to make final inferences about VE(0-18) accounting for all of the monitoring, using the median unbiased estimator of the HR(0-18) with analysis time ordering. None of the trials would have reached the potential-harm boundary or the high-efficacy boundary, though Step came close (Figure 8c).

The results are presented in Figure 8 and Table 4. For all four trials, the first interim analysis occurs at $n_1 = 65$ infections (the earliest allowed) such that the final analysis is scheduled at $n_{max} = 176$ infections, with nine analyses, the first eight evenly spaced at intervals of 15 infections. Vax004, Vax003, and Step reach the non-efficacy boundary at the seventh, first, and first interim analysis, respectively, and a conclusion of low efficacy at best would have been determined about 24, 33, and 9 months sooner than the actual designs that were used. Therefore use of the proposed non-efficacy monitoring approach would have accelerated the delivery of the non-efficacy results to the field, especially for the VaxGen trials. Furthermore, the proposed non-efficacy monitoring would have resulted in completion of the trials before hundreds of subjects would reach the Month 6 visit, hence sparing them from receiving the Month 6 immunization. In particular, for Step 645 of the 1,836 randomized men (35%) would have been spared the recombinant adenovirus vector vaccination at 6 months (Table 4).

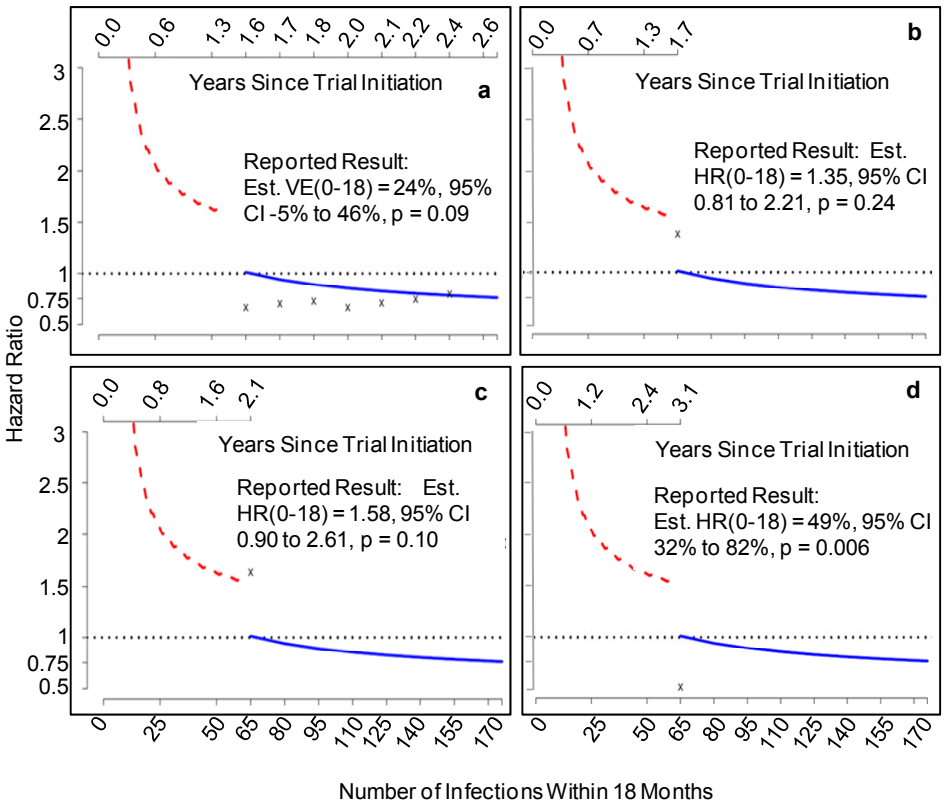

Figure 8. Proposed potential harm, non-efficacy and high efficacy sequential monitoring boundaries applied to the past HIV vaccine efficacy trials. red line = potential-harm boundary; blue line = non-efficacy boundary; x's indicate estimated hazard ratios over the first 18 months for (a) Vax004, (b) Vax003, (c) Step, (d) RV144.

In contrast to the other three trials, RV144 does not reach the non-efficacy boundary, thus indicating some positive efficacy on VE(0-18), such that the trial would have continued to stage 2, assessing vaccine efficacy over the full 36 months. As such, the monitoring plan used for RV144 would have led to similar results as the actual trial design, which is appropriate. In addition, note that of the four previous efficacy trials, Vax004 was approximately the same size as the proposed design, with 171 infections diagnosed within 18 months (compared to our target of 176 infections), whereas the other trials accrued too-few infections within 18 months to meet the infection requirements of the proposed design. This underscores the importance of conducting the proposed design in a high-incidence region.

DOI: 10.2202/1948-4690.1037

Table 4. Application of the Proposed Sequential Design to the Past HIV Vaccine Efficacy Trials.

	Total Randomized and HIV Negative at Baseline		Total Reached Month 6 Visit			Number Infections Diagnosed in First 18 Months		
Trial	**Actual Trial**	**At Boundary Proposed Design**	**Actual Trial**	**At Boundary Proposed Design**	**Time First Analysis Proposed Design**	**Total Actual Trial**	**At First Analysis Proposed Design**	**At Boundary Proposed Design**
Vax004	5,403	5,403 (100%)	5,403 (100%)	5,107 (94%)	1 yr 7mo	171	65	155 (69V: 86P)
Vax003	2,527	2,527 (100%)	2,527 (100%)	2019 (80%)	1 yr 8 mo	104	65	65 (37V: 28P)
Step	1,836	1,771 (96%)	1,836 (100%)	1191 (65%)	2 yrs 1 mo	67	65	65 (40V: 25P)
RV144	16,395	16,395 (100%)	16,395 (100%)	16,165 (99%)	3 yrs 1 mo	67	65	Not Crossed

		Trial Duration		Trial Result	
Trial	**Analysis of Boundary Crossing**	**Actual Trial**	**Proposed Design**	**Actual Trial Est. VE(0-18), 95% CI, 2-sided p-value**	**Proposed Design Est. VE(0-18), 95% CI, 2-sided p-value**
Vax004	7th	4 yrs 6 mos	2 yrs 5 mos	10%, -20% to 33%, p=0.48	24%, -5% to 46%, p=0.09
Vax003	1st	4 yrs 6 mos	1 yr 8 mo	1.03, 0.67 to 1.4, p=0.87	1.35*, 0.81 to 2.21, p=0.24
Step	1st	2 yrs 10 mos	2 yrs 1 mo	1.47, 0.95 to 2.28, p=0.08	1.58*, 0.90 to 2.61, p=0.10
RV144	Not Crossed	5 yrs 0 mos	5 yrs 0 mos	44%, 8% to 66%, p=0.02	49%, 32% to 82%, p=0.006

*Reported as Est. HR(0-18) and 95% CI for HR(0-18)

Table 5. Analysis of VE by Time Interval in the Vax004 Trial.

VE Parameter	Estimated VE*	95% Confidence Interval	p-value
VE(0-3)	-21%	-244% to 57%	0.72
VE(0-6)	31%	-14% to 58%	0.14
VE(0-9)	30%	-3% to 52%	0.07
VE(0-12)	23%	-8% to 46%	0.13
VE(0-15)	17%	-13% to 39%	0.24
VE(0-18)	10%	-20% to 33%	0.48

*Based on a proportional hazards model for infections diagnosed within the specified time-interval.

This exercise also hints at possible low-level vaccine efficacy of the Vax004 vaccine regimen during the first 18 months of follow-up, with estimated VE(0-18) = 24% and p = 0.09. However, the data-set was a pseudo data-set. For the actual Vax004 data-set, Table 5 shows point and confidence interval estimates of VE(0-3), VE(0-6), VE(0-9), VE(0-12), VE(0-15), and VE(0-18), together with p-values. While the point estimates suggest 25%−30% vaccine efficacy during the first 12 months, the results are not statistically significant, and the estimated VE(0-18) is 10% with 95% CI -20% to 33%, p = 0.48. Figure 9 shows a complementary analysis, where vaccine efficacy based on the instantaneous hazard ratio at time t, VE(t), was estimated for all t between 0 and 36 months. Specifically, the vaccine and placebo group hazard functions of infection at time t since entry were separately estimated by nonparametric kernel smoothing (with Epanechnikov kernels) for all t between 0 and 36 months, and then VE(t) was estimated by one minus the ratio of hazard function estimates (vaccine/placebo) at time t. Pointwise and simultaneous 95% confidence intervals were constructed by the method of Gilbert et al. (2002), using the bias-adjustment procedure as described. The bandwidths were chosen to minimize the mean integrated squared error as described in Gilbert et al. (2002). This analysis differs from the analyses of VE(0-3), VE(0-6), VE(0-9), VE(0-12), VE(0-15), and VE(0-18), which evaluated time-averaged hazard-ratios rather than hazard-ratios at particular times.

DOI: 10.2202/1948-4690.1037

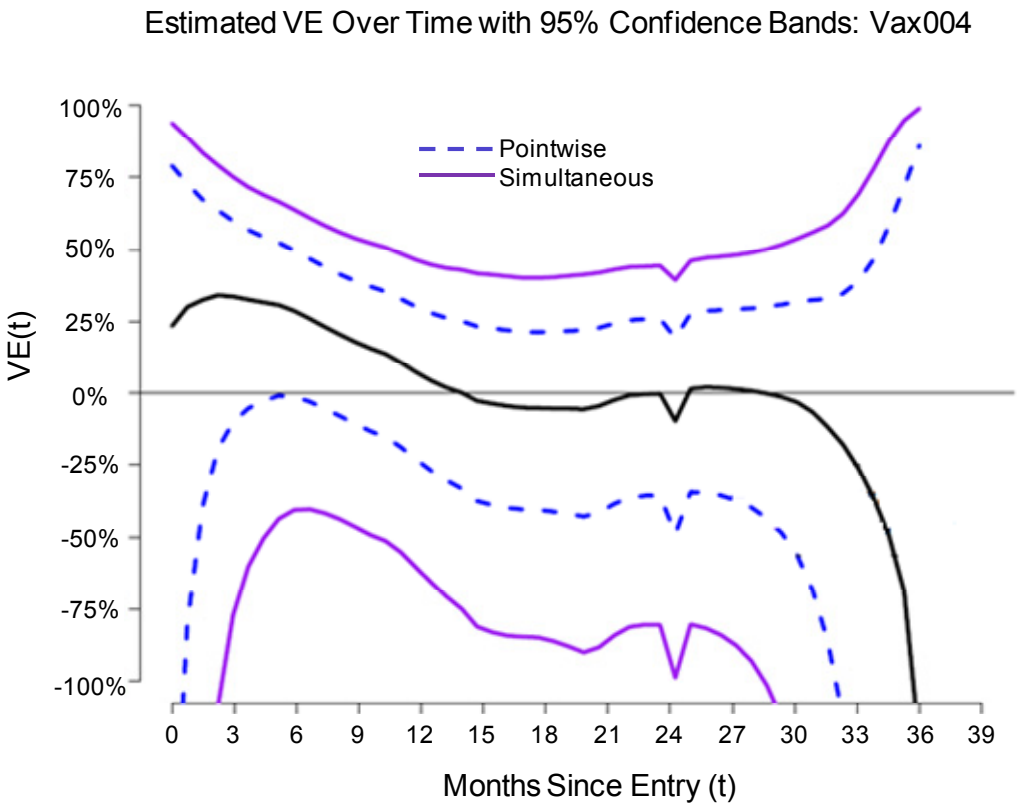

Figure 9. (Nonparametric kernel smoothing estimate of VE(t) for the Vax004 trial data, with 95% pointwise and simultaneous confidence intervals).

Statistical Power for Assessing an Immune Correlate of HIV Infection

Two main types of correlate analyses are conducted among vaccinated subjects, the first of which evaluates immunological measurements at a key fixed time-point (e.g., the Month 6.5 visit, approximate peak immunity) as predictors of HIV infection over a subsequent period of time (e.g., over the next 18 months), and the second of which evaluates time-dependent immune responses as predictors of infection during the next short interval of time extending to the next HIV test. The analyses are complementary, as the former aims to discover correlates that can be measured at a single time-point as close as possible to baseline and hence hold potential as practical surrogate endpoints; the latter addresses the relationship of the immune response near the time of exposure with the acute risk of infection. Given that vaccine-induced HIV antibodies tend to rapidly wane over time, the

analyses could easily yield different answers. The Cox proportional hazards model provides an approach to assessing both types of correlates.

We computed power to assess a normally distributed quantitative HIV-specific immunological measurement taken 2 weeks after the Month 6 visit (referred to as the Month 6.5 visit) as a predictor of the subsequent rate of HIV infection. This assessment is performed only for the vaccine groups, as the immune responses will be negative/zeros for (almost) all placebo recipients. We assume the immunological measurement has no, low, medium, or high noise, (defined as 100%, 90%, 67%, or 50%, respectively, of the inter-subject variance in the measurement being protection-relevant), where the protection-irrelevant variance may stem from a variety of sources including technical assay measurement error and variability in the time between the last immunization and the sample-draw (this time is centered around 14 days with several days of variation). We show power results for the scenario where the hazard rate of HIV infection in all of the vaccine arms pooled follows a proportional hazards model and decreases by the fraction RR per 2 standard deviation increase in the protection-relevant variability of the immunological measurement, where RR is varied from 0.3 to 1.0. For simplicity, the identical proportional hazards model is assumed for each vaccine arm.

For each of the 10,000 simulated trials discussed above for 2-, 3-, and 4-arm trials and constant VE(0-18)=50% for each vaccine arm, we counted as cases vaccine recipients diagnosed with HIV infection between 6.5 and 24 months or between 6.5 and 36 months, and assumed the immune response was measured for 95% of these subjects. Addressing these two time periods evaluates correlates of infection for exposures proximal to the immunization series, and for exposures over the complete follow-up period, respectively. For the proximal time period it would be more consistent with the primary and secondary objectives to assess correlates over 6.5 to 18 months, and our decision to focus on 6.5 to 24 months is due to the greater number of infection events, which largely improves power to detect the same effect size. However, waning of vaccine-induced immunity from 18 to 24 months may imply a smaller plausible effect size for the 6.5 to 24 month analysis.

All vaccine arms were pooled into a single group for analysis, which allows detection of a correlate with a mechanism that is common across the vaccine regimens. To create a control group of uninfected vaccine recipients, we selected a random sample of vaccine recipients that tested HIV negative at the Month 6.5 visit and completed follow-up with an HIV negative test at the terminal Month 36 visit. This sample was chosen to provide a 5:1 ratio of uninfected to infected vaccine recipients in 6.5−24 or 6.5-36 months, which provides approximately 83% efficiency compared to an approach that would measure the immune response from all controls. For each data-set, a 1-sided Wald test

DOI: 10.2202/1948-4690.1037

(alpha = 0.025) in a proportional hazards model was used to test whether the hazard rate decreases with measured immune response level. To account for the two-phase/case-cohort sampling of immune responses, the Borgan et al. (2000) estimator II was used. Power was computed as the fraction of simulation runs with 1-sided p-value bounded by 0.025. Table 6 shows the number of vaccine recipients from which we expect to have the measured immune response available.

Table 6. Number of Vaccine Recipients with Immune Response Measured at Month 6.5 Visit and Hence Used in the Evaluation of an Immunological Correlate of Risk, for Vaccine Regimens with Time-Constant VE of 50%*.

Sample Size for Analysis Counting Infections Diagnosed Between 6.5 and 24 Months			
Number of Vaccine Arms	**Expected Number Infections Diagnosed 6.5−24 Months with Immunological Data**	**Number Uninfected Vaccinee Controls (5:1 Ratio)**	**Total Number of Immunological Measurements**
1	53	265	318
2	106	530	636
3	159	795	954
Sample Size for Analysis Counting Infections Diagnosed Between 6.5 and 36 Months			
Number of Vaccine Arms	**Expected Number Infections Diagnosed 6.5−36 Months with Immunological Data**	**Number Uninfected Vaccinee Controls (5:1 Ratio)**	**Total Number of Immunological Measurements**
1	87	435	522
2	174	870	1044
3	261	1305	1566

Figures 10a-f show power curves for the 24 scenarios defined by the number of vaccine arms, assay noise levels, and time-period 6.5−24 or 6.5−36 months for diagnosing infections. Benchmarks for realistically-detectable effect sizes (RRs) are indicated on the plots, based on estimates observed in Vax004 for which there was an estimated 0.45 RR per 2 SD increase in the $\log_{10}$ 50% MN neutralization titer (Gilbert et al., 2005) and an estimated 0.61 RR per 2 SD increase in the percent viral inhibition as measured by an antibody-dependent cell-mediated viral inhibition (ADCVI) assay (Forthal et al., 2007). The four plotted benchmarks are the estimated RRs per 2 SD protection-relevant variability (x-axis scale) that result under each of the four scenarios that the assay had noise-level equal to one of our supposed levels. The results show that assay-noise

attenuates power, and that all of the designs have adequate power to detect a correlate with strength of the MN neutralization titer in Vax004, whereas the 3- and 4-arm designs but not the 2-arm design have adequate power to detect an ADCVI-like correlate. Power increases with the number of vaccine regimens.

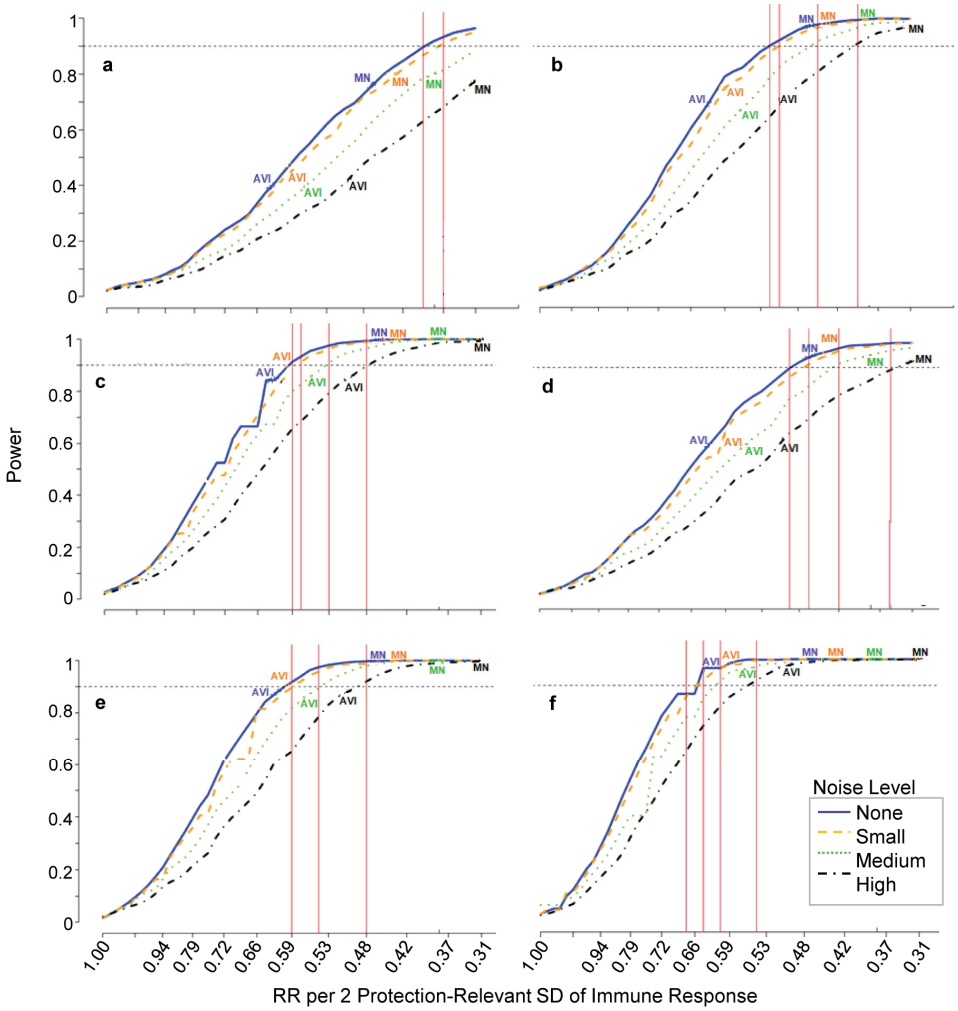

Figure 10. Power curves for the probability of detecting an immunological correlate of risk, assuming that all tested vaccines have true VE(0-18) = 50% and at least one achieves positive efficacy. A Noise Level of None, Small, Medium, and Large assumes the protection-irrelevant assay-variability is 0%, 50%, 67%, or 100% as large as the protection-relevant assay-variability, respectively. Dashed line indicates 90% Power. (a, b, c) include HIV infections diagnosed between 6.5 and 24 months post-randomization, for (a) 1, (b) 2, and (c) 3 vaccine regimens. (d, e, f) include HIV infections diagnosed between 6.5 and 36 months post-randomization, for (d) 1, (e) 2, and (f) 3 vaccine regimens.

DOI: 10.2202/1948-4690.1037

Statistical Power for Detecting a Valuable Specific Surrogate of Protection

As described above, for immunological measurements discovered to be CoRs it is of interest to evaluate their value as specific SoPs. A CoR with surrogate value will have VE(s) varying in s; therefore, we evaluate the power of the proposed trial design to reject the null hypothesis of a useless surrogate [H0: VE(s) = VE] versus the alternative hypothesis of a biomarker with some surrogate value [H1: VE(s) varies in s]. We base the calculations on the parametric method for estimating VE(s) initially developed by Follmann (2006) and later extended by Gilbert and Hudgens (2008) to accommodate 2-phase sampling and assay censoring limits.

Power is calculated using 1,000 trials simulated the same as above using the no measurement error scenario, with additional data generated for allowing the BIP, CRPV, and BIP+CRPV designs. Similar to the above, we assess power for infections diagnosed in the periods 6.5−24 months and 6.5−36 months, pooling infections across all the vaccine regimens, and assuming each vaccine has time-constant VE= 50% through 36 months. The additional generated data are as follows: (1) a BIP W is simulated in all trial participants who reach the month 6.5 visit HIV negative, such that W and S have a bivariate normal distribution each with mean 2 and variance 1 and correlation 0.8; (2) for placebo recipients HIV negative at the terminal visit at 36 months, 10 times more than the number of placebo recipients infected over the first 36 months are crossed over to the vaccine arm and have S measured; (3) the time between month 6.5 and infection diagnosis in the placebo arm follows an exponential distribution with annual incidence of 4%; and (4) the time between the month 6.5 visit and infection diagnosis in the vaccine arm conditional on S and W follows an exponential distribution with hazard rate beta10 + beta11 S, with beta10 chosen such that VE = 50% at all follow-up times and beta11 chosen such that S is inversely correlated with the infection hazard in the vaccine group and either: (i) VE(s) = VE for all s; (ii) VE(0) = 25% and VE(4) = 75%; or (iii) VE(0)=0% and VE(4)=90%. These scenarios reflect biomarkers with no surrogate value, moderate surrogate value, and high surrogate value, respectively, and the corresponding true curves are illustrated in Figure 11. Note that this set-up assumes availability of subject characteristics highly predictive of S (linear correlation 0.8, which is plausible based on the correlation of 0.85 observed between hepatitis A vaccine titers and hepatitis B vaccine titers (Czeschinski et al., 2000) and power would be less if such characteristics were not available. For simplicity, for each scenario (i)−(iii), the same true coefficients beta10 and beta11 are assumed for each vaccine arm. It would also be of interest to evaluate scenarios where the VE(s) curve differed among the vaccine regimens.

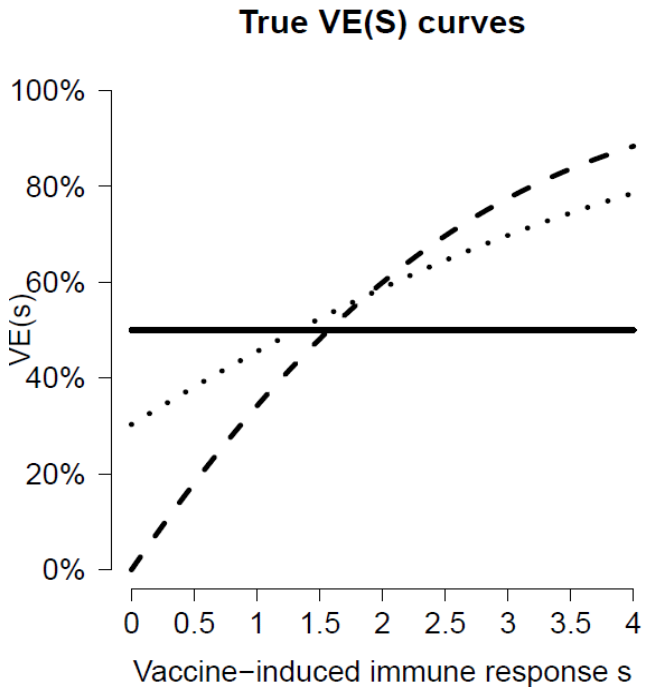

Figure 11. Three true VE(s) curves under which power is calculated for rejecting H0: VE(s) = VE in favor of H1: VE(s) varies in s [solid horizontal line = null scenario VE(s) = VE; dotted line = moderate surrogate value scenario; dashed line = high surrogate value scenario].

Table 7 shows the power estimates for these curves. The simulations confirm that the tests for all three designs have nominal size of 0.05. For a trial with one vaccine regimen, power is moderate to detect even high surrogate value; for the BIP + CRPV design power is 58% and 71% for follow-up through 24 and 36 months. This shows that greater statistical information is needed to assess a surrogate endpoint than to assess a correlate of risk, a point well known in the surrogate endpoint assessment literature. Increasing the number of vaccine arms substantially increases power, for example for the BIP + CRPV design there is power of 77% and 84% to detect high surrogate value for 2-vaccine and 3-vaccine arm trials over 24 months of follow-up. This illustrates that an important function of studying multiple vaccine regimens in the same trial is to improve the resolution of the degree to which a correlate of risk has value as a surrogate endpoint. This advantage is accrued only if the immunological predictor of VE is common among the multiple vaccine regimens, which is most likely to occur if the vaccine regimens have the same (or very similar) mechanism of protection.

DOI: 10.2202/1948-4690.1037

Given difficulty in assuring a common mechanism, it is prudent to carry out the surrogate endpoint analysis separately for each vaccine regimen, although power is limited as shown here. The efficacy trial may evaluate the same protein boost within each tested vaccine regimen, which would support plausibility of a common mechanism.

The power calculations also show that the designs with BIP provide much greater power than the CRPV design. This is expected because an excellent BIP was assumed, such that for the BIP and BIP + CRPV designs vaccine recipients outside the phase-2 sample and placebo recipients have considerable information about S; whereas in contrast for the CRPV design vaccine recipients outside the phase-2 sample have no information about S and infected placebo recipients have no information about S. Furthermore, for the CRPV design uninfected placebo subjects outside of the phase-2 sample have no information about S, and when the calculations were repeated using complete sampling of S for uninfected placebo recipients, power for the CRPV design improved considerably (results in Table 8). For example, for 3 vaccine arms and 24 month follow-up power to detect an excellent surrogate increases from 20% to 33%.

Table 7. Power for Testing that an Immunological Biomarker has Some Surrogate Endpoint Value: H0: VE(s) = VE versus H1: VE(s) varies in s [Sub-Sampling]*.

	Analysis Counts Infections Diagnosed Through 24 Months			Analysis Counts Infections Diagnosed Through 36 Months		
True VE(s)	**1Vac Arm**	**2Vac Arms**	**3Vac Arms**	**1Vac Arm**	**2Vac Arms**	**3Vac Arms**
	BIP + CRPV Design					
VE(s) = 0.50 (Null)	0.03	0.04	0.05	0.04	0.04	0.05
VE(s) = mod increase	0.27	0.35	0.41	0.43	0.58	0.68
VE(s) = large increase	0.58	0.77	0.84	0.71	0.86	0.94
	BIP Design					
VE(s) = 0.50 (Null)	0.04	0.05	0.05	0.04	0.05	0.05
VE(s) = mod increase	0.34	0.44	0.48	0.57	0.69	0.71
VE(s) = large increase	0.71	0.84	0.89	0.86	0.94	0.96
	CRPV Design					
VE(s) = 0.50 (Null)	0.05	0.04	0.05	0.03	0.06	0.04
VE(s) = mod increase	0.1	0.11	0.13	0.15	0.17	0.2
VE(s) = large increase	0.15	0.17	0.20	0.22	0.3	0.34

*For the BIP + CRPV and BIP designs the BIP has linear correlation 0.8 with the immunological biomarker S. For the BIP + CRPV and CRPV designs, among placebo recipients uninfected at closeout, 10 times more than the number of placebo recipients infected over the first 36 months are crossed over to the vaccine arm and have S measured.

In addition, the power calculations in Table 7 show that the BIP design provides slightly higher power than the BIP + CRPV design. This result is counter-intuitive given that CRPV provides additional information under the assumption (which was made) that uninfected placebo recipients with immune response measured after crossover equals the immune response 6.5 months after randomization. Part of the explanation comes from the fact that an excellent BIP was used, such that it is not surprising that no improvement is conferred. In fact, Follmann's (2006) simulation study for the case of complete-sampling showed no efficiency improvement moving from BIP to BIP+CRPV when the linear

DOI: 10.2202/1948-4690.1037

correlation between the BIP and S exceeds 0.8. Moreover, for the case of complete-sampling the simulations were repeated for a modestly predictive BIP with linear correlation 0.25, and for a single vaccine power was 48% for BIP+CRPV design and 33% for the BIP design. This demonstrates that CRPV indeed augments power when only a modestly predictive BIP is available. Another part of the explanation comes from the fact that CRPV was only administered for a phase-2 sub-sample of uninfected placebo recipients; when complete CRPV sampling was used the power between the designs equalized, and sometimes power for the BIP + CRPV design exceeded that for the BIP design (Table 8).

Table 8. Power for Testing that an Immunological Biomarker has Some Surrogate Endpoint Value: H0: VE(s) = VE versus H1: VE(s) varies in s [Complete Sampling]*.

	Analysis Counts Infections Diagnosed Through 24 Month			**Analysis Counts Infections Diagnosed Through 36 Months**		
True VE(s)	**1Vac Arm**	**2Vac Arms**	**3Vac Arms**	**1Vac Arm**	**2Vac Arms**	**3Vac Arms**
	BIP + CRPV Design					
VE(s) = 0.50 (Null)	0.04	0.05	0.04	0.05	0.05	0.04
VE(s) = mod. increase	0.35	0.45	0.53	0.58	0.73	0.79
VE(s) = large increase	0.81	0.91	0.95	0.9	0.97	0.99
	BIP Design					
VE(s) = 0.50 (Null)	0.05	0.05	0.06	0.05	0.05	0.04
VE(s) = mod. increase	0.41	0.49	0.52	0.61	0.71	0.76
VE(s) = large increase	0.83	0.91	0.94	0.91	0.97	0.99
	CRPV Design					
VE(s) = 0.50 (Null)	0.05	0.04	0.06	0.05	0.05	0.05
VE(s) = mod. increase	0.13	0.14	0.14	0.18	0.24	0.25
VE(s) = large increase	0.22	0.28	0.33	0.36	0.44	0.47

*For the BIP + CRPV and BIP designs the BIP has linear correlation 0.8 with the immunological biomarker S. For the BIP + CRPV and CRPV designs, all placebo recipients uninfected at closeout are vaccinated and have S measured.

We think that a full explanation is achieved by noting that the parametric method we used for the BIP + CRPV design uses different sets of samples to accomplish the two main estimation steps, i.e., the estimation of the conditional distribution of S given W, and the maximization of the estimated likelihood. Specifically, only samples with S and W measured in the vaccine group contribute to the former step, whereas samples with W measured in both the vaccine and placebo groups contribute to the latter step. This conjecture is partly supported by the fact that the BIP + CRPV design performs slightly better than the BIP design when we enter the information about the true conditional distribution of S into the parametric method. Moreover, in ongoing unpublished work, we are developing a nonparametric method based on a discretized W for estimating VE, which allows the information from crossed-over placebo subjects to contribute to the estimation of the conditional distribution of S. With inclusion of this extra information, we are finding that the BIP + CRPV design always provides greater efficiency than the BIP design.

Whereas the BIP and BIP + CRPV approaches require some modeling assumptions linking the risk of disease under each treatment assignment to S and other covariates, the CRPV approach can advantageously be implemented without making such assumptions. Indeed, Follmann (2006) developed nonparametric tests for any surrogate value based on the CRPV design. While appealing, we expect the BIP and BIP + CRPV designs to be most useful in practice, because the availability of a good BIP largely improves statistical power compared to the CRPV only design.

DOI: 10.2202/1948-4690.1037

Table 9. Power for Comparing VE(0-18) Between Two Vaccine Regimens[1].

Vaccine 1		Vaccine 2				
VE(0-18)	Expected Number Infections 0-18 Months	VE(0-18)	Expected Number Infections 0-18 Months	Hazard Ratio (0-18) Vaccine 2 vs. 1	Power of Log-Rank Test	Power of Log-Rank Test[2]
0%	88	30%	72	0.70	0.51	0.45
0%	88	40%	66	0.60	0.80	0.78
0%	88	50%	59	0.50	0.95	0.94
0%	88	60%	50	0.40	0.99	0.99
15%	81	40%	66	0.71	0.56	0.56
15%	81	50%	59	0.59	0.87	0.87
15%	81	60%	50	0.47	0.98	0.98
15%	81	70%	41	0.35	>0.995	>0.995
30%	72	50%	59	0.71	0.51	0.51
30%	72	60%	50	0.57	0.87	0.87
30%	72	70%	41	0.43	0.99	0.99
30%	72	80%	29	0.29	>0.995	>0.995
45%	62	60%	50	0.73	0.39	0.39
45%	62	70%	41	0.55	0.85	0.85
45%	62	80%	29	0.36	>0.995	>0.995

[1]2-sided 0.05 level log-rank test, using all available blinded follow-up information between 0 and 18 months. In particular, if at least one vaccine regimen achieves positive efficacy [i.e., the reported 95% confidence interval for VE(0-18) lies above 0%] and no vaccine regimens reach the potential-harm boundary, then all vaccine regimens have the full 18 months of follow-up. Similarly, if no vaccine regimens achieve positive efficacy but none reach the potential-harm boundary, and none reach the non-efficacy boundary until the final analysis, then all vaccine regimens have the full 18 months of follow-up. If no vaccine regimens achieve positive efficacy and at least one hits the potential-harm or non-efficacy boundary before the final analysis, then all vaccine regimens have whatever follow-up information through 18 months is available up to the time the last regimen is weeded out.

[2]The same test except that rejection of the null hypothesis requires both that the log-rank test reject and that the superior vaccine regimen achieves positive efficacy VE(0-18) > 0%.

Comparing Vaccine Efficacy Among the Vaccine Regimens

Power to Compare VE(0-18) Among the Vaccine Regimens.

Power for testing equality of VE(0-18) between two vaccine arms was evaluated in two ways, each of which uses all available blinded follow-up information through 18 months. The first way uses a standard log-rank test wherein the null hypothesis is rejected if the 2-sided p-value is less than 0.05. The second way is more stringent, wherein the null hypothesis is rejected if both the 2-sided p-value is less than 0.05 and the vaccine regimen showing superiority has VE(0-18) > 0% [based on the reported 95% confidence for VE(0-18) interval lying above 0%]. The two approaches give similar power, with slightly smaller power for the latter method if one of the vaccines has zero efficacy (Table 9).

The proposed design has high power to distinguish vaccines with 30% versus 60% VE(0-18) (power = 87%) and moderate power to distinguish vaccines with 30% versus 50% VE(0-18) (power = 51%) (Table 9).

Probability of Correctly Selecting the Vaccine Regimen with Highest VE(0-18).

In contrast to the above power results, under the objective to select-and-advance a high-performing vaccine regimen to a subsequent efficacy trial (perhaps Phase 3), without requiring reliable evidence for superiority of the advanced vaccine regimen, the design is adequately large for moderate differences in VE(0-18). In particular, suppose selection is based on the estimate of VE(0-18); for 3-arm and 4-arm designs, Table 10 shows probabilities that the truly best vaccine will be correctly selected under different scenarios for true VE(0-18) values. The design has high probability to select the best vaccine, especially if a tolerance limit of 10% VE is allowed for what constitutes a meaningful difference.

DOI: 10.2202/1948-4690.1037

Table 10. Probabilities of Correctly Selecting the Vaccine Regimen with Highest True VE(0-18)**.**

2 Vaccine Regimens			
VE(0−18) % (Vx1, Vx2)	**Prob at least 1 vaccine achieves positive efficacy[1]**	**Prob select best vaccine[2]**	**Prob select best vaccine within 10% tolerance[3]**
(0, 40)	0.81	0.81	0.81
(20, 40)	0.83	0.79	0.79
(30, 40)	0.87	0.71	0.87
(30, 50)	0.96	0.91	0.91
(40, 50)	0.98	0.80	0.98
(40, 60)	0.99	0.80	0.80
(45, 60)	>0.995	0.74	0.74
(50, 60)	>0.995	0.64	>0.995
(50, 65)	>0.995	0.51	0.51
3 Vaccine Regimens			
VE(0-18) % (Vx1, Vx2, Vx3)			
(0, 0, 40)	0.81	0.81	0.81
(0, 20, 40)	0.83	0.80	0.80
(0, 30, 40)	0.88	0.71	0.87
(20, 20, 40)	0.85	0.79	0.79
(29, 30, 40)	0.89	0.71	0.86
(0, 30, 60)	0.99	0.73	0.73
(0, 45, 60)	>0.995	0.68	0.68
(30, 30, 60)	0.99	0.72	0.72
(40, 50, 60)	>0.995	0.60	0.77

[1]A vaccine achieves some positive efficacy if the potential-harm boundary and non-efficacy boundary are never reached, such that the reported 95% confidence interval for VE(0-18) lies above 0%.

[2]This column shows the probability that the vaccine regimen with the highest estimated VE(0-18) both achieves positive efficacy and has the highest true VE(0-18).

[3]This column shows the probability that the vaccine regimen with the highest estimated VE(0-18) both achieves positive efficacy and has true VE(0-18) no more than 10 percentage points lower (on the additive scale) than the vaccine regimen with the highest true VE(0-18); e.g., if vaccines 1, 2, and 3 have true VE(0-18) of 20%, 30%, 40%, then selecting either vaccine 2 or 3 (but not vaccine 1) meets the criterion.

Additional Issues

Why Monthly HIV Diagnostic Tests?

The rationale for frequent HIV testing is to improve the assessment of immune correlates. The monthly schedule of HIV testing will allow catching 50-80% of the infected subjects in the acute-phase (antibody-negative phase) of infection, before HIV has undergone significant evolution, albeit some T cell escape may occur in the early weeks post-HIV acquisition (Goonetilleke et al., 2009). This allows analysis of the originating HIV sequences in the majority of infected subjects, thereby allowing a 'sieve analysis' to be conducted, which is a method of identifying how the vaccine efficacy on HIV acquisition depends on the genetics of the transmitting/founder HIV sequences relative to the insert HIV sequences represented in the tested vaccine (Gilbert, McKeague, and Sun, 2008); in particular to identify HIV amino acid sites and sets of sites in antibody epitopes or T cell epitopes that have an elevated rate of mismatch to the insert sequences in vaccine versus placebo recipients.

Sieve analysis is intrinsically tied with the evaluation of immune correlates of protection, as two sides of the same coin. Specifically, on the one hand, if VE > 0% and a sieve effect (i.e., elevated rate of amino acid mismatches to the insert sequence for vaccine versus placebo sequences) is detected, then the implication, given the fact the trial is randomized and double-blinded, is that vaccine-induced immune responses to certain HIV epitopes must have caused the protection. Therefore the detected sieve effect leads to follow-up explorations to identify measurable immune responses that capture (at least partially) these protective responses and thereby have some validity as surrogate endpoints for HIV infection. For example, identification of a sieve effect in 7 particular HIV antibody epitopes generates the hypothesis that the sum of neutralization levels to these 7 targets matched to the vaccine insert sequence would have high surrogate value.

On the other hand, sieve analysis is very useful for validating the degree to which an immunological measurement is a valid surrogate endpoint. To illustrate, suppose VE > 0% and the candidate surrogate, S, is a summary measure of the magnitude and breadth of neutralizing antibody titers to a panel of pseudo-viruses constructed from acute-phase HIV isolates from infected placebo recipients. If S has surrogate value to predict VE, it must be the case that protein differences to the vaccine-insert are larger in infected vaccine than placebo recipients; this logically follows because genetic mutations in antibody epitopes are known to effect neutralization levels. Therefore, sieve analysis is a tool for corroborating the surrogate value of S as a SoP. However, this sieve analysis would not be possible with infrequent HIV diagnostic testing such as the semi-annual schedule

DOI: 10.2202/1948-4690.1037

used by the previous efficacy trials, given that too-few infected subjects would be caught in the acute-phase to afford an assessment of the vaccine effect on transmitted sequences.

In addition, the sieve analysis may be directly incorporated into the surrogates assessment described above, by estimating the VE(s) curve with the endpoint definition restricted to HIV infection with a strain within a certain threshold of genetic distance to the vaccine-insert. This analysis would be repeated for a range of thresholds. Greater variation in the VE(s) curve for thresholds closer to the insert-sequence would support the value of the immune biomarker as a surrogate endpoint.

Intention-to-Treat and Per-Protocol Analysis of VE.

Vaccine efficacy trials commonly assess VE in the intention-to-treat (ITT) cohort, which is all randomized subjects, as well as in the modified intention-to-treat (MITT) cohort, which is the subset of the ITT cohort that are later discovered to not have been HIV infected at baseline. Because blinded procedures are used for ascertaining baseline infection status, the MITT analysis has the same validity from randomization as the ITT analysis, such that the MITT analysis is generally preferred, given that it assesses vaccine efficacy in HIV uninfected persons. In addition, given the ubiquitous concern that a vaccine may not confer protection until all or at least some of the immunizations are received, most vaccine efficacy trials also assess vaccine efficacy in the sub-cohort that receives all of the immunizations and are disease-free after the immunization series; this sub-cohort may be referred to as the per-protocol (PP) cohort (Horne, Lachenbruch, and Goldenthal, 2001). All of the past HIV vaccine efficacy trials assessed VE in both the MITT and PP cohorts, with the MITT assessment the primary analysis in each case (Gilbert et al., 2010).

As stated above, the MITT analysis is primary because the comparator groups are guaranteed to have balanced prognostic factors on average due to randomization and double-blinding, such that the analysis assesses the causal effect of assignment to vaccine. In contrast, the standard analytic approach to assessing PP VE applies the same method as used for the MITT analysis, which compares HIV infection incidence between the subgroups of vaccine and placebo recipients that are observed to qualify for the PP sub-cohort. However, these comparator sub-cohorts are subsets of randomized subjects, such that the analysis is susceptible to possible post-randomization selection bias (Rosenbaum, 1984; Robins and Greenland, 1992; Frangakis and Rubin, 2002), hence making the results difficult to meaningfully interpret. To improve upon this standard analysis of VE in the PP cohort, an analytic method that adjusts for measured factors that simultaneously predict HIV infection and PP sub-cohort membership (such

factors cause the selection bias) should be applied (e.g., Lu and Tsiatis, 2008; Tsiatis et al., 2008; Zhang, Tsiatis, and Davidian, 2008; Moore and van der Laan, 2009; Zhang and Gilbert, 2010), which in addition to correcting for bias can improve statistical power by leveraging prognostic factors. Moreover, because some simultaneously predictive factors may be unmeasured, the sensitivity of results to such factors should also be investigated, following the paradigm described in Scharfstein, Rotnitzky, and Robins (1999). Therefore, in our proposed design we assess VE in the MITT cohort for the primary analysis and conduct a causal sensitivity analysis of PP VE for the secondary analysis, wherein the answer is reported as a range of point estimates and a corresponding union of 95% confidence intervals (a so-called "sensitivity interval"), which account for a spectrum of potential levels of post-randomization selection bias (Shepherd, Gilbert, and Lumley, 2007).

Timing of Reporting of Results and of Un-blinding.

With respect to reporting the results, the proposed design has two stages: for stage 1, results are reported on VE(0-18); and for stage 2 [which occurs if and only if at least one vaccine regimen achieves positive efficacy for VE(0-18)], results are reported on the durability of VE between 18 and 36 months. For stage 2 the issues are simple: all vaccine arms advanced to stage 2 plus the placebo arm continue blinded follow-up until the last enrolled subject has 36 months of follow-up, at which time the final analysis is conducted and the results reported.

The issues are more complicated for stage 1, with the approach to un-blinding dependent upon which boundaries are reached. As soon as a vaccine arm reaches a conclusion [either by reaching the potential-harm boundary, the non-efficacy boundary, the high efficacy boundary, or completing the evaluation of VE(0-18) without reaching a boundary], the result is reported. This conveys the result to the field as expeditiously as possible. If a vaccine arm completed its evaluation by reaching the potential-harm boundary, then the arm would be immediately un-blinded, given the ethical warrant to inform participants of the potential harm caused by exposure to the vaccine. The other study arms would continue blinded. If a vaccine arm reaches the high efficacy boundary, then the placebo group is immediately un-blinded and offered this vaccine. If it is the single vaccine arm design, then the sole vaccine group is also un-blinded. However if it is the multiple vaccine arm design, and at least two vaccine arms are still being evaluated, then the blind is maintained for all of the vaccine arms, which allows continuing accrual of data for comparing vaccine efficacy head-to-head among the vaccine regimens. Furthermore, if a vaccine arm reaches the high efficacy boundary, it may be worth continuing the vaccine's evaluation out to 36 months. While a rigorous assessment of durability of VE will likely be impossible

DOI: 10.2202/1948-4690.1037

(given that the contemporaneous comparator placebo group is being offered the vaccine), the additional follow-up may nonetheless provide useful data about the vaccine, which would be difficult to collect in follow-on studies. Further thought is needed on this issue, and on whether it is also warranted to offer subjects assigned to the other vaccine arms the highly efficacious vaccine.

Next we consider the scenario wherein a vaccine arm completes its evaluation by reaching the non-efficacy boundary. In this case, blinded follow-up under the original HIV diagnostic testing schedule would continue either until all other vaccine arms are weeded out, or, in the case that at least one vaccine arm achieves positive efficacy, until all enrolled subjects have 18 months of follow-up. This continued blinded follow-up would contribute information to the analyses of safety, VE(0-18) (including comparisons with other vaccine regimens), and immune correlates of protection. If, alternatively, the arm were un-blinded then the post-un-blinding data would be excluded from the main analyses of vaccine efficacy and of immunological surrogate endpoints, given that the un-blinding may lead to imbalances in HIV prognostic factors between the vaccine and placebo groups (and between vaccine arms), which could not be confidently corrected for statistically due to the inability to accurately measure HIV risk behavior and exposure. Given the scientific benefit accrued from maintaining the blind and the absence of evidence of harm caused to participants, it seems ethical to maintain blinding for subjects assigned a vaccine regimen shown to have low efficacy at best.

For operational reasons, ideally all study arms would be un-blinded at the same time, as un-blinding one study arm could compromise follow-up for the participants assigned to the other arms. As discussed above, by dividing the trial into two stages the design does not achieve this, as vaccine arms reaching a stopping boundary will be un-blinded once the evaluation of VE(0-18) is completed, whereas vaccine arms not reaching a stopping boundary will be un-blinded once stage 2 is completed (expected at least 18 months later). While one approach would keep vaccine arms reaching the non-efficacy boundary blinded all the way through stage 2, this seems like a poor use of resources, given that non-efficacy over 18 months is expected to predict non-efficacy from 18-36 months, such that it is prudent to complete the evaluation of non-efficacious vaccines at 18 months. Thus, our approach makes the un-blinding as simultaneous as ethically warranted within each stage. As discussed above, for stage 2 a completely simultaneous un-blinding is achieved, whereas for stage 1, if no vaccine arms reach the potential-harm boundary then a completely simultaneous un-blinding is achieved. The informed consent process would describe the events that would trigger un-blinding, and the approach to un-blinding would be vetted with local Institutional Review Boards and the DSMB.

In summary, the whole study is un-blinded at the first event of: (1) the last of the vaccine regimens is weeded out, either by reaching the potential-harm boundary or the non-efficacy boundary; (2) the last of the vaccine regimens reaches the high efficacy boundary; (3) the last enrolled subject reaches 36 months of follow-up, for the case that neither event (1) nor (2) occurs. For event (1), the trial has maximum duration of 18 months beyond the last enrolled subject, and minimum duration the time at which either the last weeded-out vaccine regimen reaches the potential-harm boundary or accrues n_1 infections diagnosed within 18 months.

What Does Completing a Vaccine Regimen for Non-Efficacy Entail?

As described above, upon reaching the non-efficacy boundary, the primary result on VE(0-18) would be reported, thus providing data as expeditiously as possible. Figure 6 (a) shows that, by the time a vaccine regimen reaches the non-efficacy boundary, accrual is very likely to be complete, in which case weeding out a regimen would not spare enrollees, all of whom would have received at least one immunization. On the other hand a substantial fraction of enrollees will likely have not yet completed the immunization series, such that ceasing vaccinations upon reaching a non-efficacy boundary would spare immunizations. For example, at the median stopping time of a vaccine with VE(0-18) = 0%, approximately 3000 of the 4300 enrollees (pooled over a vaccine arm and placebo) would have completed the immunization series through Month 6 and approximately 1800 through Month 12. Moreover, regardless of the number of immunizations spared, it still may be warranted to cease immunizations at the time of reaching the non-efficacy boundary, as the primary question about VE(0-18) would have been answered. Furthermore, if accrual lags behind the planned accrual, then this approach may spare a great deal of immunizations and substantially decrease the total enrollment. Lastly, if a vaccine regimen reaches the potential-harm boundary then a large number of enrollments and immunizations would likely be spared. Therefore, the proposed design ceases immunizations and accrual to vaccine arms if and when they reach a non-efficacy boundary.

Equal Versus Unequal Allocation to the Vaccine and Placebo Groups.

The design equally allocates subjects to each study group, which is inefficient for the two- and three-vaccine arm trials, for which the efficient design would randomize more subjects to the placebo arm. The rationale for equal allocation is to increase the information for the second and third secondary objectives to evaluate immunological correlates of infection rate in the vaccine groups and to compare vaccine efficacy among the vaccine regimens. Equal allocation results in

DOI: 10.2202/1948-4690.1037

efficiency loss for the primary objective in exchange for efficiency gain for key secondary objectives. This reflects the premise of the design that development of immune correlates of protection and head-to-head comparisons of vaccine efficacy are priorities for HIV vaccine research. More research is needed to thoroughly define the trade-offs of the equal-versus-unequal allocation approaches.

Accommodation of Pre-Exposure Prophylaxis (PrEP) and for Other HIV Prevention Interventions.

Recently an efficacy trial in men who have sex with men in the Americas (mostly South America) demonstrated that daily oral PrEP use [fixed-dose combination tenofovir disoproxil fumarate (TDF) and emtricitabine (FTC)] provided an estimated 44% reduction in the incidence of HIV infection compared to placebo (Grant et al., 2010). Moreover, the incidence rate appeared especially low in men with detectable PrEP drug levels, suggesting that the PrEP efficacy is higher for adherent subjects. Because the PrEP drugs TDF and FTC are approved and some vaccine trial participants may take PrEP, it is relevant to consider how the design accommodates PrEP use. Moreover, several other efficacy trials of PrEP are ongoing, such that it is prudent to plan for how the trial design will respond to future results that will become available before or during the trial.

The baseline approach to accommodating PrEP does not alter the primary analysis, as it is intention-to-treat and compares HIV incidence among the vaccine and placebo groups while disregarding PrEP use. The event-driven design set-up is also unaltered, such that with or without PrEP the same numbers of HIV infections trigger the interim and final analyses. However, once the required numbers of events are fixed, PrEP use impacts the anticipated sample size needed to achieve the required number of infections in a timely manner via impact on the background HIV incidence. For example, if 10% PrEP use occurs, and we assume that PrEP users have a 50% reduction in incidence, then the sample size would need to be increased by approximately 5% ($0.05 = 0.10 \times 0.50$) in order to deliver results within the same time-frame as the baseline scenario (no PrEP use). Alternatively, if all participants are offered PrEP and 80% accept it, then the sample size would need to be increased by approximately 40% ($0.40 = 0.80 \times 0.50$).

Given the difficulty to predict the degree of PrEP use, the trial would monitor PrEP use through self-report questionnaires and PrEP drug level measurement. The enrollment target could be adjusted based on this monitoring; such an adaptation would pose minimal risk to study integrity because it is based on blinded data and a deterministic plan could be pre-specified for what data lead to what kinds of trial expansions. There is also uncertainty in the degree of PrEP

efficacy, and this is addressed through the operational futility monitoring; the level of PrEP efficacy will affect the background HIV incidence, and the lower it is the more likely the operational futility guidelines will be met. The operational futility monitoring is primarily based on rates of accrual, HIV infection, and dropout during the study, regardless of the amount of PrEP use or PrEP efficacy.

It is relevant to evaluate whether PrEP is expected to enhance or diminish vaccine efficacy for trial design set-up, as this would impact the maximum plausible effect size VE, and hence could result in powering the trial for a different effect size. Currently the data on potential interaction of vaccines and PrEP are too scant to warrant altering effect size assumptions.

A second approach to accommodating PrEP use would offer a voluntary second randomization to PrEP or to PrEP placebo. This would form three analysis strata: subjects assigned PrEP, subjects assigned PrEP placebo, and subjects who declined the second randomization. The primary analyses would be intention-to-treat similar to the above, the difference being they would be stratified. For each regimen HIV incidence would be compared between vaccine and placebo within each of the three strata separately, and then aggregated into one overall estimate of VE; for example, assuming the same VE within each strata and using strata-specific baseline hazards in the Cox proportional hazards model. This analysis is valid because randomization and double-blinding guarantee balance in HIV prognostic factors within each stratum. While an interaction of PrEP and vaccine would complicate the interpretation, the assessment of the common VE still has useful interpretation as vaccine efficacy averaging over the three strata.

This primary analysis does not explicitly account for data on PrEP use or PrEP adherence, because of complications in achieving valid inferences adjusted for post-randomization intermediate variables that are subject to measurement error. However, secondary analyses using causal inference method would evaluate vaccine efficacy while subjects are actually using and not using PrEP. Additional secondary analyses would compare efficacy among each of the individual arms (Vaccine + PrEP, Placebo + PrEP Placebo, Vaccine + PrEP Placebo, Placebo + PrEP Placebo). A third approach would power the trial to compare efficacy among these individual arms, implicating a larger trial would be needed. These considerations for accommodating PrEP use are also relevant for use of other HIV prevention approaches. Accommodating microbicides may be particularly relevant given the recent report of a partially efficacious microbicide (point estimate of 39% reduction in HIV incidence compared to placebo) in the CAPRISA 004 Phase 2b efficacy trial of tenofovir gel conducted in South Africa (Karim et al., 2010).

DOI: 10.2202/1948-4690.1037

Summary of the Proposed Design

The proposed design has the following main features:

- Multiple vaccine regimens versus a shared placebo group
  - The design evaluates vaccine efficacy of multiple vaccine regimens using a shared placebo group in the same geographic region, and is akin to multiple Phase 2b two-arm vaccine versus placebo trials conducted simultaneously. The simultaneous evaluation of multiple vaccine regimens accelerates learning about vaccine efficacy compared to sequential two-arm trials.
  - The simultaneous evaluation of multiple vaccine regimens improves the assessment of immune correlates of protection compared to a two-arm design, both by increasing the amount of statistical information and by facilitating greater variability in vaccine-induced immune responses.
  - The design is well-powered to detect large differences but not moderate differences in VE(0-18) among the vaccine regimens. The design provides high probability of correctly selecting the vaccine regimen with the highest VE(0-18) within a 10% tolerance limit.
  - These advantages still attain if the vaccine regimens are initiated at different calendar times, as long as a concurrent placebo group is always used (see discussion in the section below, "Other Issues of Interest That Merit Further Research.") However, precision for comparing VE among vaccine regimens may be reduced given the need to adjust for potential secular effects using the placebo group HIV incidence data.

- Two-stage evaluation of vaccine efficacy
  - Vaccine efficacy of each vaccine regimen is evaluated in two stages, first over 18 months, and, if there is reliable evidence for VE(0-18) > 0%, over an additional 18 months. This design is efficient because assessment of durability of vaccine efficacy becomes a priority if, and only if, there is evidence for vaccine efficacy for infections occurring relatively soon after the immunization series.

- Improved assessment of immune correlates of protection
    - Whereas even the design with one vaccine regimen has reasonable power for detecting correlates of HIV infection rate, testing multiple vaccine regimens is necessary for providing reasonably high power for assessing the surrogate value of identified correlates. Using identified correlates as immunogenicity study endpoints for follow-up, Phase I/II HIV vaccine trials may provide an unreliable basis for comparing and selecting vaccine regimens by their future protective efficacy. In contrast, vetting identified correlates by their estimated surrogate value provides a more rigorous basis for selecting follow-up study endpoints that are more likely to reliably predict protective efficacy in future efficacy trials.
    - Availability of baseline subject characteristics predictive of an immunological biomarker is crucial for enabling well-powered assessment of the biomarker as a surrogate endpoint.
    - The design provides for early initiation of the immune correlates assessment for promising vaccine regimens while maintaining confidentiality of the results and double-blind format.
    - The design uses frequent HIV testing, with advantages to improve the immune correlates assessment and the sieve analysis, as well as to allow assessment of the vaccine effect on acute-phase viral load. In addition, as PrEP use may increase over time, frequent HIV testing may also help in preventing drug resistance through PrEP use.

- Sequential monitoring for efficient evaluation of vaccine efficacy
    - The design sequentially monitors each vaccine regimen for potential harm to elevate the rate of HIV infection, for non-efficacy and high efficacy. The design also conducts operational futility monitoring, to stop the trial if accrual, HIV incidence, or study quality metrics are inadequate.
    - The potential-harm monitoring plan is maximally vigilant by assessing a vaccine-associated elevation in HIV acquisition risk after each HIV infection.
    - The non-efficacy monitoring plan is designed to be as aggressive as possible while explicitly guarding against prematurely weeding out efficacious vaccine regimens with low efficacy during the initial immunizations. In particular, the monitoring plan is configured to bound the risk of weeding out

DOI: 10.2202/1948-4690.1037

a vaccine regimen with VE(0-18) = 40% and halved efficacy during the first six months at 20%.

- For each tested vaccine regimen, the design will yield a result about vaccine efficacy over the first 18 months [VE(0-18)] that reliably distinguishes between greater than 0% efficacy versus less than 46% efficacy (i.e., the reported 95% confidence interval adjusting for the monitoring will exclude 0% or 46%). Therefore, weeded-out vaccine regimens will have reliable evidence that the vaccine efficacy is bounded by 46%. These statements are not absolute in that there is a tiny probability (< 1%) that a vaccine regimen would reach the potential-harm boundary very early (with the observed infection rate much higher in the vaccine than placebo arm) and the reported 95% confidence interval would cover both 0% and 46%.

Other Issues of Interest That Merit Further Research

- Research is needed to identify the most appropriate methods for distinguishing waning VE from lack of waning VE in the presence of heterogeneous HIV exposure. For an identified method, power calculations are needed. Related research is needed for estimation of VE over time.
- Power calculations are needed for comparing durability of vaccine efficacy [e.g., VE(18-36)] among vaccine regimens.
- Power calculations are needed for assessing the vaccine effects on acute-phase viral load and on viral load at the HIV infection diagnosis visit.
- Additional research is needed to ensure that the method for inference and estimation of VE(0-18) and of related vaccine efficacy parameters appropriately account for all of the sequential monitoring that is conducted.
- This article assumed that all of the vaccine regimens were started simultaneously. While this would be ideal, practical considerations may necessitate staggered start times of one or more vaccine regimens. This event would require an extension of the total accrual period for the shared placebo group, to ensure concurrent randomized controls for all assessments of VE(t), and to provide a valid basis for comparing VE(t) among the vaccine arms. Research is needed to determine the impact of staggered vaccine start times on the ability to achieve the primary and secondary study objectives.
- The statistical method for the primary and secondary analyses of VE(0-18) and of VE(t) may be made more efficient by leveraging baseline subject

characteristics predictive of HIV infection, as discussed above in the section, "Intention-to-Treat and Per-Protocol Analysis of VE." While valid published methods are available for accomplishing this task, research is needed to compare the methods to identify one with optimal characteristics for the context of the proposed trial design.

DOI: 10.2202/1948-4690.1037

Appendix

Background

Elizabeth M. Adams
Division of AIDS, National Institute of Allergy and Infectious Diseases, Bethesda, MD

On January 11th, 2011, the National Institute of Allergy and Infectious Diseases (NIAID) and Office of AIDS Research (OAR), the National Institutes of Health, sponsored a statistical workshop entitled "Alternative Study Design for Early Efficacy Evaluation of HIV Prophylactic Vaccines." The overarching goal of this technical workshop was to have focused discussion and provide constructive criticism on the study design discussed in the cited paper by Gilbert et al.
Simultaneous advancement of multiple HIV vaccines into phase III efficacy testing would require commitment of extraordinary human and capital resources as well as an expansive and experienced multi-site and multi-country clinical infrastructure. It is unlikely the committed funders of HIV vaccine development and the world community would be able to support such an effort, which is anticipated to be further complicated by the requirements of providing an expanding prevention package that may include microbicides and pre-exposure prophylaxis [PREP]. Hence the need for methodologies to make early HIV vaccine efficacy evaluation more efficient in a scientifically valid and rigorous manner is clear.

The proposed trial design in the cited paper by Gilbert et al. offers one approach to early efficacy or phase IIB evaluation should two or more candidate HIV vaccines be ready for simultaneous testing. The following topic areas focused the discussion of the trial design paper during the workshop: sequential monitoring in the proposed trial and alternative approaches; issues surrounding evaluation and comparison of vaccine efficacy; design and analysis for evaluating immune correlates of protection; impact of additional design and analysis considerations on the proposed trial (e.g., waning vaccine efficacy, advancing to phase III testing, etc.,); and operational considerations related to implementing the proposed trial design. As the emphasis of the workshop was statistical and technical rather than programmatic in nature, specific HIV vaccine study products, the HIV vaccine pipeline, the state of the HIV vaccine clinical research agenda, etc., were not addressed.

Invited attendees included academicians and statisticians with vaccine industry background, government regulators with experience in adaptive designs, and clinical trialists who have implemented or provided oversight for various adaptive design studies or been involved in phase III studies and DSMB interactions. The comments and recommendations from participants fell into

three general categories: 1- those on the design for fine tuning; 2- those based on participants' past experience from vaccine and adaptive design trials in order to help facilitate study implementation; and 3- those related to design considerations for potential follow-on studies. Attendees were invited to prepare talks for the workshop and provide commentary after the workshop. A few of the attendees (Michael Proschan, James P. Hughes, James Dai, Ivan S. F. Chan, Dean Follmann, Anneke Grobler, Gavin J. Churchyard, and Glenda Gray) generously contributed written commentary following the workshop, and their comments are included here.

Some commentaries are written by authors in their capacity as NIH employees, but the views expressed here do not necessarily represent those of the NIH.

DOI: 10.2202/1948-4690.1037

Commentary on Monitoring Aspects of Gilbert et al.

Michael Proschan
Biostatistical Research Branch, Division of Clinical Research, NIAID, Bethesda, MD.

The primary comparisons are of 18-month vaccine efficacy (VE) of each of 2-3 vaccines compared to placebo, with no adjustment for multiple comparisons. Durability, correlates of protection, and comparisons of vaccines with each other are secondary comparisons for vaccines showing benefit over 18 months. The sample size of approximately 2150 patients per arm ensures approximately 176 HIV infections for each pairwise comparison of vaccine with placebo. This trial monitors for high efficacy, non efficacy, and harm. The boundaries and timing for these three distinct goals are quite different.

High Efficacy

Monitoring for high efficacy occurs after approximately 44, 88, 132, and 176 infections, and the boundaries are very difficult to cross for two reasons. The first is that the statistical test is designed to prove that the VE is greater than 50%, rather than just 0%, using a one-tailed test at α=0.025. The second reason that early stopping will be very difficult is the selected boundary. A spending function dictates the cumulative amount of type I error rate to spend by different fractions of trial completion such that 0.025 is spent by the end of the trial. The properties of this approach depend on the spending function selected. The one selected spends very little early and yields boundaries similar to the O'Brien-Fleming (1979) procedure, namely quite high early on. The combination of the chosen boundary and the use of a null VE value of 50% instead of 0% require the splits shown in Table 1 to stop for high efficacy. For instance, the boundary requires that at least 42 of the first 44 infections occur in the placebo arm to stop at the first interim look. Making it difficult to stop early for high efficacy is good for several reasons. One is that un-sustained trends are not uncommon early in a clinical trial. Also, because HIV vaccines have not been very successful thus far, there is reason for skepticism that a new vaccine would be wildly effective. Subjects in a vaccine trial do not yet have the disease, so there is no ethical imperative to put everyone on a vaccine, as there would be to put cancer patients on a successful treatment, for example. Also, subjects are likely to be receiving care that is at least as good as what they would be receiving outside the trial. Still, the level of evidence required here may present challenges to a Data and Safety Monitoring Board (DSMB). How many DSMBs seeing a 41-3 split at the first interim analysis would argue that the trial question had still not been answered? In summary, it is a good idea to make early stopping for high efficacy

difficult by limiting the number of interim efficacy looks and using steep boundaries; nonetheless, the boundary at the first interim efficacy look may be a bit too steep.

Table 1. Level of evidence required to declare high efficacy at each of 4 interim looks.

	Look 1	**Look 2**	**Look 3**	**Look 4**
# Infections	44	88	132	176
Required Split	42-2	72-16	101-31	131-45

Non Efficacy

Monitoring for non-efficacy differs in several ways from high-efficacy monitoring. First, it does not begin until approximately 75 infections occur for each pairwise comparison. This is to ensure that there is a sufficient amount of data before abandoning a potentially useful vaccine. From this point on, monitoring occurs more frequently than the monitoring for high efficacy—up to 9 times using the Emerson-Fleming (1989) procedure. The idea is to choose from between two hypotheses: 1) there is no vaccine efficacy or 2) the vaccine efficacy is 46% or better. The procedure ensures a 2.5% chance of each error—declaring a benefit when the true VE is 0%, or declaring no benefit when the true VE is 46%. There is a design parameter called p that determines how difficult it is to stop early for non-efficacy. A value of p close to 0 results in steep early boundaries, while p close to 1 makes it easier, to cross the non-efficacy boundary early. The value selected, p=0.6, produces boundaries that are not overly steep early on. This translates to observing a hazard ratio of about 0.96 at the first analysis after about 75 infections. After that, it becomes progressively easier to stop for non-efficacy. For instance, after about 140 infections, non-efficacy is declared even if the observed hazard ratio is about 0.80 (i.e., 20% observed VE). Gilbert et al. (2011) showed that although the selected monitoring plan is fairly aggressive at weeding out unfavorable vaccines, there is only a 20% chance of falsely weeding out a vaccine with 40% VE overall, but which is only half as effective in the first 6 months.

An alternative useful tool when contemplating stopping for non-efficacy is conditional power. One could compute the conditional probability of demonstrating VE>40% at the end of the trial, given the current data, under various assumptions about the true VE. If this conditional power is low enough (e.g., 15% or lower), one could stop for non-efficacy. Similarly, one could compute the probability of showing that VE>0% at the end of the trial, given the

DOI: 10.2202/1948-4690.1037

current data. Conditional power is very natural and meaningful to clinicians, and one can compute it under a variety of assumptions about VE, not just the static value of 46%. One may determine that a smaller VE of 40% might still be meaningful, especially if there are few side effects. Moreover, the 2.5% error rate of the Emerson-Fleming procedure holds when the boundaries are treated as binding, but Data and Safety Monitoring Boards seldom treat any boundary as binding, especially futility boundaries. Conditional power is regarded as an additional useful tool to help make decisions rather than as a binding boundary.

I support the main features of the proposed non-efficacy monitoring, namely starting only after about a third or more of the total number of infections are observed, and subsequently looking fairly frequently with an eye toward dropping inferior arms. The Emerson-Fleming method is very reasonable. I would supplement it with conditional power calculations to aid the decision about whether to stop for non-efficacy.

Harm

Monitoring for harm occurs much more frequently, beginning with the 7^{th} infection and occurring after each subsequent infection. Monitoring early and often for harm makes sense because harm, when it occurs, can often be seen early in a clinical trial. Underlying the method is the fact that, given a new infection and in the absence of any true harm, that infection should be equally likely to be in the vaccine or placebo arm. The idea is to continually test (after each new infection) whether the probability that the infection came from the vaccine arm exceeds 0.5. The boundary is selected so that the probability of ever falsely declaring harm is 0.05. Using a one-sided test at alpha=0.05 instead of 0.025 makes perfect sense because one does not want to require a very stringent level of evidence in order to declare harm. It also makes sense to monitor frequently, because one does not want to wait too long to find that a vaccine is harmful. However, there is a tradeoff between frequency of monitoring and steepness of the boundary; the more frequently we monitor, the stronger the evidence must be at a given look to stop for harm. Continuous monitoring after the 7^{th} infection means that the p-value needed to declare harm is about 0.005. Therefore, it may be preferable to try to find a happy medium between very frequent monitoring with high boundaries and less frequent monitoring with lower boundaries. For instance, safety monitoring after every 8^{th} infection yields a p-value criterion closer to 1% (about 0.009) than 0.5%. At the same time, the monitoring is frequent enough to detect harm early if it exists.

Summary

While I have small disagreements about some specifics of the monitoring plan, I agree completely with the principles underlying it: making it difficult to stop early for high efficacy, monitoring early and often for harm, and deferring non-efficacy analyses until about one third of the total infections accrue. I also find the monitoring plan to be extremely well thought out and well evaluated in terms of its properties. The authors used simulation to show that a truly harmful vaccine with a relative risk of 3 has nearly a 100% chance of being declared harmful, a truly neutral vaccine with a relative risk of 1 has about a 93% chance of crossing the non-efficacy threshold, and a truly beneficial vaccine with relative risk of 0.5 has about a 95% chance of demonstrating that VE>0%. They also showed how the proposed monitoring plan would have shortened some previous HIV vaccine trials. Gilbert et al. are to be commended for their very thoughtful design of an HIV vaccine trial.

DOI: 10.2202/1948-4690.1037

Commentary on Gilbert et al.

James P. Hughes[1] and James Dai[2]
[1] Department of Biostatistics, University of Washington, Seattle, Washington
[2] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington

One of the objectives of the trial design described in Gilbert et al. [1] is to estimate waning vaccine efficacy, defined as $VE(t) = 100x(1 - HR(t))$ for $t > 18$ months. Since this can be measured only on individuals who are uninfected at 18 months, a naïve estimate of the waning VE conditions on a post-randomization event (infection at 18 months) may be influenced by selection bias [2]. In this note, we discuss assumptions and approaches to estimating the causal vaccine efficacy $VE(t)$, $t > 18$.

Shepherd et al [3] discuss causal estimation of post-randomization outcomes in a closely related context. They use the principal stratification approach of [4] to estimate the causal effect of an HIV vaccine on time to progression to AIDS, an outcome that can only be observed among those who become infected with HIV. This same framework can be extended to the problem at hand. Using the same notation as [3], let Z_i denote randomization assignment (1 = vaccine, 0 = placebo) for subject i, S_i denote the infection status at 18 months (1 = infected, 0 = not infected), T_i and C_i are the post 18 month infection and censoring times (i.e., $T_i = C_i = 0$ at 18 months), respectively, $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i < C_i)$. Note that (T_i, C_i, Y_i and Δ_i) are only defined if $S_i = 0$. For each individual, we also define the counterfactual (or potential) outcomes ($S_i(z)$, $T_i(z)$, $C_i(z)$, $Y_i(z)$, $z = 0,1$) as the outcomes that would have been observed had subject i been randomized to placebo or intervention, respectively. Of course, only one of the potential outcomes (corresponding to the actual randomization assignment) is observed for each participant but, conceptually, it is useful to think of each individual as belonging to one of the four principal strata defined by the (counterfactual) infection status at 18 months (S(0), S(1)), namely, (0,0), (0,1), (1,0), (1,1). These strata may be thought of as the "not infected," "harmed," "protected" and "infected" strata, respectively.

We assume, as usual in the time-to-event context, independent censoring: $C_i(z) \perp T_i(z) \mid S_i(z)$ for $z = 0,1$. It is also common in the context of vaccines to also assume monotonicity: $S_i(1) \leq S_i(0)$ for all i. This assumption eliminates the (0,1) principal stratum and effectively assumes that no individual is put at increased risk of infection by the vaccine. In light of the results of the recent STEP trial [5] this assumption may be untenable in general. However, since the issue of waning vaccine efficacy is only of interest for a vaccine which is clearly efficacious in the initial 18 months, the assumption is more reasonable in the current context

(although, of course, overall efficacy does not imply efficacy in every possible subgroup). Therefore, parallel to [3], we will assume monotonicity.

Using this notation, one can define the causal average vaccine efficacy (cAVE) between 0 and 18 months as $cAVE_0(18) = 1 - P(S(1) = 1)/P(S(0) = 1)$ and, under randomization, this quantity can be estimated directly from the observed data using standard methods. The causal effect of the vaccine after 18 months can only be estimated among the subset of participants who survive to 18 months (i.e., those for whom $S_i = 0$). However, as noted above, this comparison is subject to selection bias (unless the vaccine has no efficacy in the first 18 months, i.e., no (1,0) stratum). To avoid selection bias the waning vaccine effect is defined in terms of the individuals in the (0,0) strata only. Let $F^0_{(0,0)}(\tau) = P(T(0) \leq \tau | S(0) = 0, S(1) = 0)$, $F^1_{(0,0)}(\tau) = P(T(1) \leq \tau | S(0) = 0, S(1) = 0)$ and similarly for the other principal strata. Then, following [3], one can define the survival causal effect (SCE) between 18 and 18 + τ months as

$$(1) \qquad SCE_{18}(\tau) = F^0_{(0,0)}(\tau) - F^1_{(0,0)}(\tau).$$

Similarly, one can define the causal average vaccine efficacy between 18 and 18 + τ months as

$$(2) \qquad cAVE_{18}(\tau) = 1 - F^1_{(0,0)}(\tau)/ F^0_{(0,0)}(\tau).$$

Under monotonicity, $F^0_{(0,0)}(\tau)$ is identifiable from the data and may be estimated using the usual Kaplan-Meier estimator of $P(T \leq \tau \mid Z = 0, S = 0)$. However, the corresponding observable quantity for vaccine arm participants, $P(T \leq \tau \mid Z = 1, S = 0)$, is a mixture of participants in the (0,0) and (1,0) principal strata:

$$(3) \qquad P(T \leq \tau \mid Z = 1, S = 0) = P(S(0) = 0 \mid S(1) = 0) * F^1_{(0,0)}(\tau) + (1 - P(S(0) = 0 \mid S(1) = 0)) F^1_{(0,0)}(\tau).$$

Under randomization and monotonicity, the mixing proportion, $P(S(0) = 0 \mid S(1) = 0)$, can be estimated from the data as $RS = P(S = 0 \mid Z = 0) / P(S = 0 \mid Z = 1)$, the relative survival at 18 months in the placebo arm compared to the vaccine arm.
Then, by (3) and since, $0 \leq F^1_{(0,0)}(\tau) \leq 1$,0, $F^1_{(0,0)}(\tau)$ is bounded by max $(((P(T \leq \tau \mid z = 1, S = 0) - (1\text{-}RS))/RS), 0)$, min $(((P(T \leq \tau \mid z = 1, S = 0))/RS),1)$. Further, (3) can be rewritten as

$$(4) \qquad F^1_{(0,0)}(\tau) = (P(T \leq \tau \mid z = 1, S = 0))/(RS + (1\text{-}RS)\varphi(\tau))$$

where $\varphi(\tau) = F^1_{(1,0)}(\tau)/F^1_{(0,0)}(\tau)$ quantifies the selection bias - the relative proportion infected (between time 18 and τ), if randomized to the vaccine arm, among

DOI: 10.2202/1948-4690.1037

participants in the "protected" stratum compared to participants in the "not infected" stratum. If $\varphi(\tau) = 1$, there is no selection bias and the observed estimate of $P(T \leq \tau \mid Z = 1, S = 0)$ is unbiased for $F^1_{(0,0)}(\tau)$ (so $cAVE_{18}(\tau)$ may be estimated unbiasedly from observed data). Unfortunately, $\varphi(\tau)$ cannot be identified from the data. Thus, one must do a sensitivity analysis to understand the behavior of $F^1_{(0,0)}(\tau)$ and, hence, $cAVE_{18}(\tau)$, across a range of values of $\varphi(\tau)$.

Shephard et al. [3] describe a sensitivity analysis that is similar in spirit to the above. The sensitivity parameter $\varphi(\tau)$ can be related to the sensitivity analysis in [3] (which is based on $P(S(0) = 1 \mid S(1) = 0, T(1) \leq \tau)$) by noting that $Odds(S(0) = 1 \mid S(1) = 0, T(1) \leq \tau) = \varphi(\tau) * Odds(S(0) = 1 \mid S(1) = 0)$. Shepherd et al. [6] relax the monotonicity assumption although the resulting sensitivity analysis becomes more complicated as it involves three parameters rather than just one as above. Conversely, it may be reasonable in the present context to assume that the "no harm" or monotonicity assumption continues through time τ. Effectively, this implies that $F^1_{(0,0)}(\tau) \leq F^0_{(0,0)}(\tau)$ and sets a tighter upper bound on the sensitivity analysis in (4) (and implies an upper bound of 1 for $cAVE_{18}(\tau)$). An interesting alternative to sensitivity analyses would be a Bayesian analysis that puts a prior on the sensitivity parameter(s) and bases inferences on the posterior distribution of the causal estimate of interest.

Conditional on $\varphi(\tau)$, $cAVE_{18}(\tau)$ is a function of observable quantities. In the context of estimating $SCE_{18}(\tau)$, Shepherd et al. [3, 6] discuss the problem of variance estimation of $\hat{F}^0_{(0,0)}(t)$, and $\hat{F}^1_{(0,0)}(t)$. We speculate that the delta method could be used to obtain variances for $\log(\hat{F}^0_{(0,0)}(\tau))$ and $\log(\hat{F}^1_{(0,0)}(\tau))$, thereby, confidence intervals for $cAVE_{18}(\tau)$.

References

Gilbert, P.B., et al., *A Sequential Phase 2b Trial Design for Evaluating Vaccine Efficacy and Immune Correlates for Multiple HIV Vaccine Regimens.* Statistical Communications in Infectious Diseases, 2011.

Rosenbaum, P.R., *The consequences of adjustment for a concomitant variable that has been affected by the treatment.* Journal of the Royal Statistical Society, Ser. A, 1984. **147**: p. 656-666.

Shepherd, B.E., P.B. Gilbert, and T. Lumley, *Sensitivity Analyses Comparing Time-to-Event Outcomes Existing Only in a Subset Selected Postrandomization.* J Am Stat Assoc, 2007. **102**(478): p. 573-82.

Frangakis, C.E. and D.B. Rubin, *Principal stratification in causal inference.* Biometrics, 2002. **58**(1): p. 21-9.

Buchbinder, S.P., et al., *Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial.* Lancet, 2008. **372**(9653): p. 1881-93.

Shepherd, B.E., P.B. Gilbert, and C.T. Dupont, *Sensitivity Analyses Comparing Time-to-Event Outcomes Only Existing in a Subset Selected Postrandomization and Relaxing Monotonicity.* Biometrics.

DOI: 10.2202/1948-4690.1037

Commentary on the Evaluation of Immune Correlates in Gilbert et al.

Ivan S.F. Chan
Late Development Statistics, Merck Research Laboratories
North Wales, PA

The evaluation of immune correlates is very critical in vaccine development. If an immune marker that reliably predicts the clinical outcome is identified, it can be used as a "surrogate endpoint" to measure vaccine efficacy in clinical trials so that more efficient evaluation of new vaccines or manufacturing process changes can be performed based on immune correlates instead of disease endpoints, which can only be obtained from large-scale, often costly and lengthy, field efficacy trials. In vaccine literature, most common methods for immune correlate evaluation have primarily focused on identifying a threshold level of immune response (called "protective level") that correlates with disease protection [1, 2]. More recently, statistical models focusing on the whole antibody titer distribution [3] and the general approach proposed by Prentice [4] for surrogate endpoint validation have been applied to evaluate immune correlates in vaccine trials [5].

In the proposed HIV trial, Gilbert et al. designed a 2-tier approach for assessing the immune correlates using a rigorous statistical framework [6, 7] with the goal to (1) identify HIV immune markers as potential correlates of risk (CoR), and (2) to evaluate whether a CoR can be used as a surrogate of protection (SoP) that reliably predicts vaccine efficacy. This approach is novel and more rigorous compared with the methods commonly used in vaccines (mostly focusing on CoR evaluation). Of note, demonstration of a good SoP requires showing the relationship between a CoR and the disease outcome is the same in the vaccine group as in the placebo group. Therefore, the SoP evaluation is more difficult than the CoR evaluation and generally requires a large sample size. Because no detectable HIV immunity is expected from placebo recipients, the tier-1 analysis of CoR is proposed to be performed only in the vaccine recipients using a case-control method. This design is appropriate as placebo recipients will not contribute information to the correlation analysis. The proposed 5-to-1 matching is also reasonable as it preserves high statistical efficiency compared with complete sampling. The power analysis shows that the study has adequate power for detecting a HIV immune CoR assuming an anticipated vaccine efficacy of 50%, particularly with at least 2 vaccine regimens included in the study. It also suggests that the likelihood of identifying potential immune correlates will be greatly improved by including multiple regimens in the trial. The statistical precision and power might be further enhanced if one can identify good baseline prognostic factors that can be used to stratify the population for selecting the matching controls.

The absence of any HIV immunity in placebo recipients also poses great challenges for the tier-2 evaluation of SoP. As noted in the literature [3, 8], when placebo recipients do not have immunity, the Prentice method is not applicable for surrogate endpoint validation due to the confounding of vaccination status and immune response status. To overcome this problem, the trial proposed two novel approaches of predicting the 'counterfactual' values of the vaccine-induced immunity in placebo recipients using baseline immunogenicity predictors (BIP) and crossover placebo predictors (CRPV) based on the causal inference framework [9, 10]. The BIP approach will rely on the ability to identify a good baseline predictor of HIV immunity as the trial design assumes a correlation of 0.8, which may be optimistic. It would be useful to understand the impact on power if the predictor has lower correlation (e.g., 0.4 or 0.5) with HIV immune responses. The CRVP approach seems to have much lower power than the BIP approach, which may be due to the fact that only a subset of placebo recipients are selected to crossover at the end of the trial. In addition, it is a bit unclear as to why the combined approach of using both CRVP and BIP predictors actually has lower power than the BIP approach alone. The size of the proposed study offers some hope for the SoP evaluation if there is a very good BIP predictor and a large change of vaccine efficacy as a function of immune marker. The likelihood of a successful SoP evaluation will be increased by having long-term follow-up (at least 36 months) and by incorporating multiple vaccine regimens.

Overall, Gilbert and colleagues have done an excellent job in designing this phase IIb study. The well thought-out approach to the evaluation of immune correlates is novel and scientifically rigorous, offering reasonably high power for identifying potential HIV immune correlates of risk and some hope in validating their surrogacy for protection. The design can be readily applied to subsequent large-scale phase III efficacy trials for a more definitive evaluation of immune correlates, especially for the SoP evaluation which requires a large sample size.

References

Siber, G. (1997). Methods for estimating serological correlates of protection. *Developments in Biological Standardization* 89, 283-296.

Siber, G.R., Chang, I., Baker, S., Fernsten, P., O'Brien, K.L., Santosham, M., Klugman, K.P., Madhi, S.A., Paradiso, P., Kohberger, R. (2007). Estimating the protective concentration of anti-pneumococcal capsular polysaccharide antibodies. *Vaccine* 25, 3816-3826.

DOI: 10.2202/1948-4690.1037

Chan, I.S.F., Li, S., Matthews, H., Chan, C., Vessey, R., Sadoff, J., and Heyse, J. (2002). Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine*, 21, 3411-3430.

Prentice, R. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8, 431-440.

Kohberger, R.C., Jemiolo, D. And Noriega, F. (2008). Prediction of pertussis vaccine efficacy using a correlates of protection model. *Vaccine* 26, 3518-3521.

Qin, L., Gilbert, P., Corey, L., McElrath, J. and Self, S. (2007). A framework for assessing an immunological correlate of protection in vaccine trials. *The Journal of Infectious Disease* 196, 1304-1312.

Gilbert, P., Qin, L. and Self, S. (2008). Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine*, 27, 4758-4778.

Chan, I.S.F., Wang, W.W. and Heyse, J.F. (2003) Vaccine clinical trials. *Encyclopedia of Biopharmaceutical Statistics*, 2nd Edition, Marcel Dekker, NY, 1005-1022.

Follmann, D. (2006). Augmented designs to assess immune response in vaccine trials. *Biometrics* 62, 1161-1169.

Gilbert, P. and Hudgens, M. (2008). Evaluating candidate principle surrogate endpoints. *Biometrics* 64, 1146-1154.

Commentary on Gilbert et al

Anneke Grobler
Dorris Duke Medical Research Centre, CAPRISA, Nelson Mandela School of Medicine University of KwaZulu-Natal, Durban, South Africa

Gilbert et al proposes a study design for a screening trial to evaluate efficacy of multiple vaccine regimens using a shared placebo group. This is akin to multiple phase 2b two-arm vaccine versus placebo trials conducted simultaneously. The goal of the design is to provide a large screening trial where the decision can be made as to which of the possible vaccines should be advanced to phase 3 testing. The proposed design should work well in settings where one vaccine is clearly superior to others or where one, or more, vaccine(s) is clearly harmful and should be discontinued. The design is superior to various parallel arm studies to determine which vaccine should be advanced; but only if several vaccines are in the same developmental stage at the same time.

The stopping rules proposed by Gilbert et al are extensive, with good operating characteristics. They propose different stopping rules and sequential monitoring for high efficacy, potential harm, operational futility or no-efficacy; each done at different time points and with different boundaries.

In this commentary I want to focus on one aspect of the stopping rules; stopping for high efficacy. The proposed design is excellent for screening out unsuccessful vaccine candidates, while protecting against the premature discarding of a vaccine which requires multiple immunizations before reaching optimum effectiveness. However, this design might not be the fastest route to licensure should one of the vaccine candidates prove to be highly efficacious. Since the goal of the proposed design is to identify THE candidate for a Phase 3 trial; the study (all arms) is stopped once one vaccine passes a certain, very high, boundary for high efficacy. I want to highlight two potential issues with this approach.

It is assumed by drug developers that two trials, both with a statistically significant beneficial effect, are required to license a drug. It might be difficult to justify a second trial for preventing an outcome associated with high morbidity or mortality [1]. Medicine regulatory authorities are willing to consider the results of a single, well conducted trial for licensure if the observed strength of evidence is comparable to that obtained from two independent trials [2]. Simplistically, this is interpreted as a trial with a very small p-value (one-sided less than 0.025*0.025=0.000625).

The current design therefore raises the following question: What if the vaccine is very efficacious but the screening trial is stopped prior to reaching what was called the level of evidence required by two trials? This could require

DOI: 10.2202/1948-4690.1037

another trial to be done; maybe in the absence of equipoise. A possible suggestion is to use the stopping boundaries suggested by Gilbert et al; but instead of stopping the study as suggested once one vaccine is regarded to be highly efficacious, all arms but the highly efficacious arm and the placebo arm are stopped, and possibly increase the sample size in both these arms and continue the screening trial to some predetermined number of HIV-events in order to provide one trial which could possibly lead to licensure. What I propose is essentially a phase 2 screening trial that moves seamlessly into a phase 3 trial, targeting licensure level of evidence, if the effectiveness of one of the vaccines appears to be above a predetermined threshold. This way the information already collected in the phase 2 study can be used in the larger phase 3 trial.

Consider a related example; a design where a two-arm parallel phase 3 trial has 80% power to detect a 50% efficacious vaccine after 66 HIV infections. An interim analysis is done, at say 44 events. At this time point, the trial can be adapted. Only one adaptation is allowed, and only at this one time point. If the effectiveness estimate falls within a certain predetermined range, the study expands to a predetermined number of HIV-events; larger than the original planned number (66). If the observed effectiveness is larger than this specific range, the study does not expand; because enough evidence is expected to be available at the planned end of the study. If the observed effectiveness at this interim review is below this range, it is regarded as unlikely that expansion to the larger trial would lead to registration level of evidence and the trial continues to the original planned number of events. A design such as this does not inflate alpha if an appropriate interim effectiveness range is chosen, increases power slightly, especially when true effectiveness of the vaccine is high, and leads to study expansion about two thirds of the time. This may alleviate the need for a confirmatory study without committing excessive resources from the outset to a trial with a vaccine of unknown efficacy [3]. Type I error would be much harder to control when one chooses amongst several vaccine candidates to move forward than it was in this example from a two-arm study. This proposed design should be adapted for several vaccines in the context of a screening trial.

Unfortunately, any study design is only as good as the practical limitations within which it can be implemented. The proposed design would only be useful if many vaccine candidates were available for screening at the same time and a choice had to be made to take the one most promising vaccine to confirmatory testing. Any adaptive design is only as adaptive to real time events as the timelines of the data allow. Very tight data management processes would need to be in place to evaluate all vaccines for harm immediately after each event occurred. This would need immediate reporting of every single HIV-infection to the database. Something that is much easier on paper than done in a multisite trial, possibly with several laboratories.

I propose one small modification to the screening trial, which allows a phase 2b trial to roll seamlessly into a phase 3 trial that targets the strength of evidence of two trials. The phase 3 component is modest in size, and is only implemented if the interim effectiveness estimate suggests that the vaccine is much more effective than originally thought. This could accelerate the development process substantially while maintaining equipoise.

References

Fleming TA, Richardson BA. Some design issues in trials of microbicides for the prevention of HIV infection. J Infect Dis. 2004; 190: 666-674.

Transcript of the Food and Drug Administration (FDA) Antiviral Drugs Advisory Committee Meeting, 20 August 2003 (Bethesda, MD).

Grobler A, Taylor D. Adaptive design considerations in planning the CAPRISA 004 study. Oral presentation at the Microbicides 2010 conference, Pittsburgh, PA.

DOI: 10.2202/1948-4690.1037

Commentary on Gilbert et al., - Operational aspects from a South African perspective

Gavin J Churchyard[1] and Glenda Gray[2]
[1] Aurum Institute for Health Research, Klerksdorp, South Africa
[2]Perinatal HIV Research Unit, University of the Witwatersrand, Johannesburg, South Africa

Trial site issues

Adaptive HIV vaccine trials bring with them a number of operational challenges. Monthly study follow up visits are required for HIV testing to enable real time monitoring of the number of HIV infection endpoints. Estimates of HIV incidence are based on data from current HIV prevention trials that follow up patients three times monthly. Ethically, participants would need to receive harm reduction counseling at each monthly study visit, which may reduce HIV incidence below that observed with less frequent follow up visits. Thus the sample size may need to be increased to accommodate a lower HIV incidence. Intensive study follow up visits may also be associated with a greater loss to follow up than a less frequent visit schedule. The ongoing VOICE microbicide trial is currently following up women monthly and will provide insight into what impact more frequent study visits may have on HIV incidence. Monitoring for operational futility will however identify a lower than expected HIV incidence and thus enable sample size to be increased appropriately.

Adaptive trial designs require the clinical trial sites to be able to rapidly stop the trial or to increase the number of participants recruited. Based on the experience from the Phambili study which was evaluating the Merck trivalent rAd5 HIV vaccine, it is possible to rapidly stop a trial [Gray et al. Current Opinion AIDS, 2010]. Increasing the number of participants recruited if required is not likely to be a challenge. The adaptive trial design requires a high, uniform enrollment rate, which, based on the Phambili experience is possible.

Financial and human resource management of adaptive trial designs is more complex as trials may go to completion or be stopped prematurely if a boundary for non-efficacy, harm or futility is met. If a trial is stopped prematurely it will be challenging to maintain site infrastructure and key staff. A strategy to retain experienced staff and maintain trial infrastructure if a trial is stopped early is required as it is very expensive and time consuming to rebuild capacity once it is lost. Staff will need focused training on unique aspects of adaptive trial designs, such as unblinding and cross-over studies.

Participants

During the informed consent process it will be important to inform potential participants that there will be monthly follow up visits with harm reduction counseling and blood draws, and that the trial may be stopped prematurely.

PreExposure prophylaxis with tenofovir tablets in men who have sex with men and tenofovir gel in women have both been shown to be effective in preventing HIV infection in high risk individuals [Karim, Science, 2010]. Tenofovir for use as PreP or as a microbicide gel has not been approved for use in South Africa and is therefore not currently available through the public sector or medical schemes. Prior to regulatory approval for these indications, adaptive HIV vaccine trials should consider doing a second randomization to PreP or a microbicide gel. The acceptance of PreP or microbicide gel is likely to be high [Karim, Science, 2010].

Regulators and community advisory boards (CABs) will need to be consulted extensively on adaptive trial designs, and be able to provide input on the pre-specified adaptations proposed. For post-hoc adaptations, regulators will need to establish the capacity to review applications for adaptations rapidly. Regulators may have to work closely with the DSMB and the study team. CABs will play an important role in communicating results to the community, particularly if the trial is terminated for harm or non-efficacy. Although there are operational challenges to doing adaptive trials, all can be addressed with education and ongoing support.

References

Abdool Karim Q, Abdool Karim SS, Frohlich JA, Grobler AC, Baxter C, Mansoor LE, Kharsany AB, Sibeko S, Mlisana KP, Omar Z, Gengiah TN, Maarschalk S, Arulappan N, Mlotshwa M, Morris L, Taylor D; CAPRISA 004 Trial Group. Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. Science. 2010 Sep 3;329(5996):1168-74.

Gray G, Buchbinder S, Duerr A.Overview of STEP and Phambili trial results: two phase IIb test-of-concept studies investigating the efficacy of MRK adenovirus type 5 gag/pol/nef subtype B HIV vaccine. Curr Opin HIV AIDS. 2010 Sep; 5(5):357-61.

DOI: 10.2202/1948-4690.1037

Comments regarding immune correlate analysis for multiple HIV vaccine regimens

Dean Follmann

First off, I want to thank the authors for providing a thoughtful and thorough proposal to rapidly and efficiently evaluate multiple HIV vaccine regimens (Gilbert, Grove Gabriel, Gray, Self, Kublin, and Corey, 2011). It was an interesting and thought provoking read. It seemed like every aspect of trial conduct had been scrutinized with fresh eyes and the most modern methods applied. In this discussion, I want to briefly review the design and then focus on the immune correlates analysis.

The proposal differs in many ways from previous HIV vaccine trials:

- Frequent monitoring for harm, non-efficacy, and high efficacy. This monitoring is more frequent (e.g., every infection for harm) and encompassing (three different processes) than is usually done.
- Rapid reaction following identification of a promising vaccine to
    - Identify immune responses that predict protection
    - Vaccinate some placebo uninfected subjects with the promising vaccine
    - Evaluate whether the vaccine benefit observed over the first 18 months continues over the next 18 months
- Simultaneous evaluation of multiple vaccine candidates

Why the difference? The standard phase III trial has been around for a long time and is the ideal experiment to conduct when a promising vaccine candidate has been identified and one is willing to make a substantial bet that the vaccine will be successful and lead to licensure. For such trials, monitoring for non-efficacy is often not done, immune correlates are not a substantial focus and a single vaccine is evaluated. The proposed design reflects the current HIV vaccine zeitgeist where there is no candidate with a good bet for high VE, but promising signals from RV144 have sparked hope that guided vaccine improvement may lead to a licensable product.

To help guide development, there is keen interest in identifying immunological responses that track with infection risk. Consider Figure 1, which is a hypothetical curve linking risk of infection to an immune response measured in the vaccine group. Such a figure suggests that if a future vaccine could induce immune responses of around 5, the vaccine might be nearly perfect. In the nomenclature Qin, Gilbert, Corey, McElrath, and Self (2007) use, such an immune response is a correlate of risk of (CoR). However, on the basis of such data it is not really possible to know whether such a response is truly causative. It

could be that volunteers who produce more of an immune response are those who have better innate immunity, thicker mucous, or even less of a libido. If one is willing to dismiss these possibilities, then one might proceed to modify a vaccine to achieve more of this specific immune response. However, it seems prudent to further investigate whether the CoR is truly a proper target.

To help such investigation it is appealing to evaluate multiple vaccine regimens that work through similar mechanisms. With such information, confidence can be gained that putative correlates are reliable. For example, if a particular antibody assay readout of x is reliably associated with a low infection rate of 1%, no matter what vaccine is used, then there is greater confidence that the antibody readout is a useful correlate. For this kind of thinking to hold, however, it is important that the vaccine regimens are similar enough but not identical. For example, if modestly successful vaccine A worked by antibodies and modestly successful vaccine B worked by T cells, the two might be difficult to combine statistically. However, if vaccines worked through antibodies, they should be more readily combinable even if they target different sites (e.g., different parts of the viral envelop) or different mechanisms, e.g., antibody dependent cell mediated cytotoxicity (ADCC) or neutralization antibodies, provided a functional assay was used. If binding assays were used, the vaccines might not be combinable. For example, a readout for vaccine A might be zero on an assay that reads high for vaccine B and vice versa. Qualitatively different regimens can still be useful, for example, if regimen A worked better than regimen B, one might try to improve A rather than B, but it is unclear how to combine them in a single model.

DOI: 10.2202/1948-4690.1037

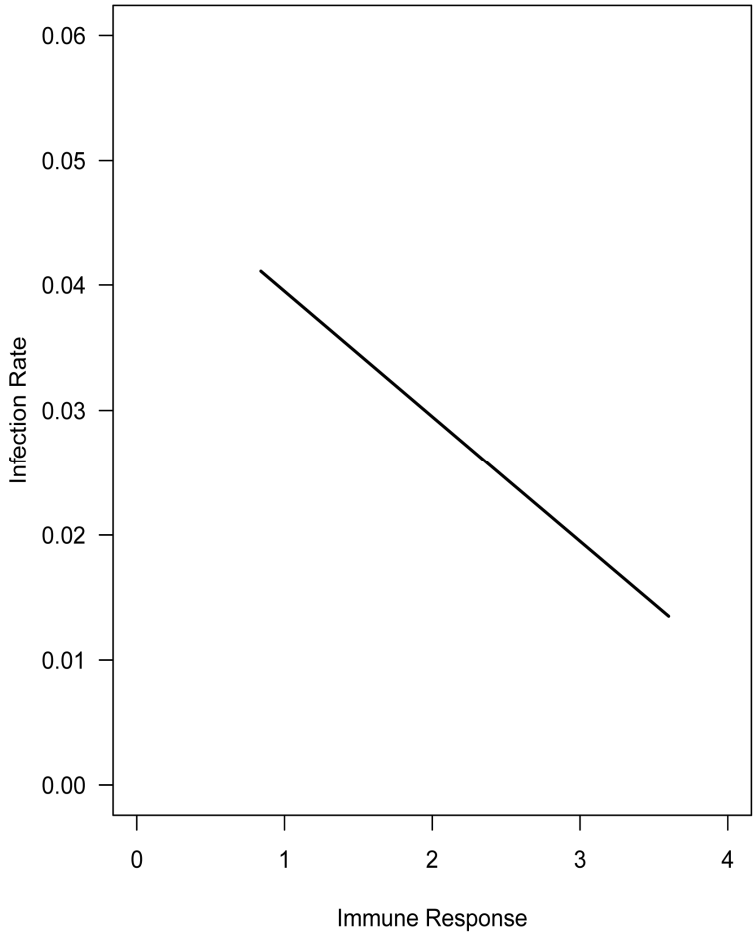

Figure 1. True infection rate as a function of vaccine induced immune response measured from a single vaccine.

To further explore the analyses that one can accomplish with similar vaccines, consider Figures 2 and 3. Figure 2 would be reassuring that a readout of *x* produces the same result whether from vaccine A or B. More formally such data allow us to examine whether an analogue for Prentice's criterion for surrogacy was met. Figure 3 indicates an additional kind of possibility that might dampen our enthusiasm for the readout *X* as being a promising target. The key point here is that we wouldn't know if Figure 2, 3 or something else held unless we evaluated multiple vaccine regimens that induced immune response *X*.

In Gilbert et al (2011), power calculations are done to evaluate correlates of risk where the association between immune response and risk of infection is

similar to that seen in Vax004. Overall, these power calculations are reassuring in that risk correlations similar to Vax004 have good power of being detected under the proposed design. While probably not essential, it might be interesting to calculate power where the relationship between immune response and infection risk is similar to that seen in influenza where there is a very strong relationship (Qin, et al, 2007). Presumably power would be better under such a scenario.

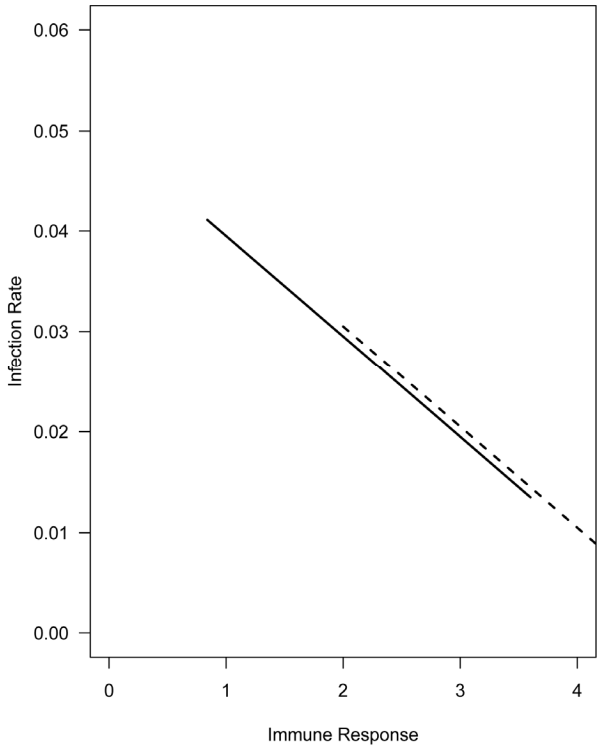

Figure 2. True infection rate as a function of vaccine induced immune response measured from two vaccines with different immunogenicity but similar relationships.

DOI: 10.2202/1948-4690.1037

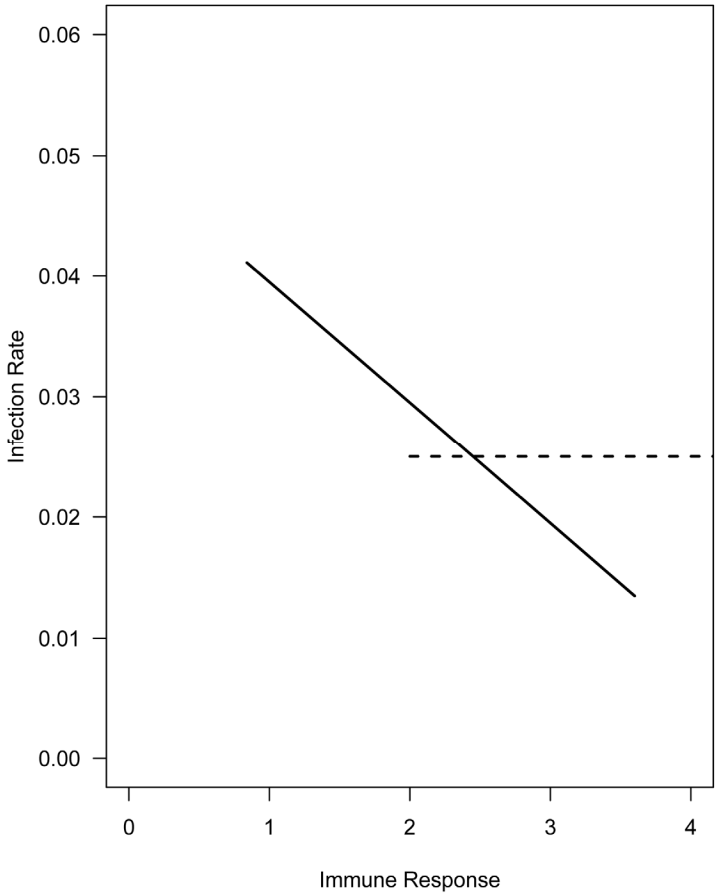

Figure 3. True infection rate as a function of vaccine induced immune response measured from two vaccines with different immunogenicity and different relationships.

Table 1. Death rates in the clofibrate and placebo arms of the Coronary Drug Project Trial broken down by self-reported pill count.

	<80% Pills	**>80% Pills**	**Overall**
Placebo	28.2	15.1	19.4
Clofibrate	24.6	15.0	18.2

Another kind of maneuver that can help one gain comfort about the role of an immune correlate is best motivated by an old trial in cardiovascular diseases that had a surprising result. The Coronary Drug Project was a clinical trial

conducted in the 1960s and 1970s to assess the efficacy and safety of 5 lipid lowering drugs in 8,341 men who had a prior MI (The Coronary Drug Project Research Group, 1980). Table 1 shows the death rates for the placebo group and the clofibrate arm broken down by self-reported pill counts. Looking at the clofibrate group alone, it seems that clofibrate pill consumption has a substantial benefit. However, this benefit is mirrored in the placebo group. So it seems that patients who choose to take pills do better than patients who don't bother and that clofibrate molecules don't really have a causative role in reducing the death rate. Without a placebo group, one might have been easily mislead about the effect of clofibrate on mortality.

Table 2. Relative risk of HIV infection by immune response quartile in the vaccine group.

	Immune Response Quartile			
	Weak	**Modest**	**Good**	**Best**
Placebo	?	?	?	?
Vaccine	1.00	0.43	0.34	0.29

This example illustrates a potential concern for HIV vaccine trials where we are searching for an immune response that has a causative role on infection. Immune response is measured post-randomization but only in the vaccine group. So the structure is worse than the CDP example as immune response to the HIV vaccine cannot be measured in the placebo group. One might be tempted to argue that vaccine trials are different and that such a peculiar result could not happen. But the VaxGen North American trial elevates this concern. Table 2 reports a strong decrease in the relative risk of infection as a function of immune response quartile of the vaccine group. However, overall the infection rates were nearly the same, suggesting that perhaps a similar relationship would be obtained in the placebo group, if only we could have determined what immune response the placebo volunteers would have had, had they received vaccine.

This concern motivates two design aspects that try to infer, more or less, what immune response the placebo volunteers would have had, had they received vaccine. The first maneuver is the baseline immunogenicity predictor or BIP. This is best illustrated by Figure 4. A baseline variable(s) is measured in both groups prior to randomization and this variable has a strong relationship between measured immune response in the vaccine group. Because baseline variables are equally distributed between the placebo and vaccine groups by randomization, one can apply the regression relationship that is estimated in the vaccine group to

DOI: 10.2202/1948-4690.1037

the placebo group. One can then regress the infection rate on measured immune response in the vaccine group and on imputed immune response in the placebo group. If the infection risk by imputed immune response is flat in the placebo group, we have greater comfort that the immune response is truly causative. Thus the big black dot at 2.5 represents a baseline predictor measured on a placebo patient. To impute what his HIV immune response would have been, we read off the predicted value from the regression line obtaining the big open dot.

The second maneuver is to cross over some number of uninfected placebo participants at the end of the trial. We then assume that the HIV immune response measured at the end of the trial is about what they would have had at the start of the trial. This allows us to fill in the first row of Table 3. By randomization we know that there should be about equal numbers of vaccine and placebo volunteers in each of the immune response quartiles and can thus approximately fill in the second and third rows with the italicized number. Table 3 illustrates a situation where the immune response does not seem causative.

Table 3. Trial where Crossover Placebo Vaccination identifies a misleading immune response in a trial of 800 volunteers. Italicized numbers are approximate.

		Immune Response Quartile				
Group	**Outcome**	**Weak**	**Moderate**	**Good**	**Best**	**Total**
Placebo	Uninfected	40	60	80	100	280
	Infected	*60*	*40*	*20*	*0*	120
	Total	*100*	*100*	*100*	*100*	400
Vaccine	Uninfected	70	85	90	95	340
	Infected	30	15	10	5	60
	Total	100	100	100	100	400

Gilbert et al (2011) evaluate power for each maneuver separately and the combination under various scenarios. Crossover placebo vaccination alone has poor power. This is partly due to crossing over only a fraction of the placebo uninfected subjects; but even when all placebo uninfected subjects are crossed over, power only improves from 20% to 33% for a scenario with a large effect. Additionally, there is a suggestion that it is only necessary to cross over several fold the number of placebo infected subjects to achieve power close to crossing over all the patients. This thinking is true for case-control studies, but probably does not apply here. In Follmann (2006), the variance of parameters of interest was proportional to the number of placebo uninfected subjects that were crossed over, unlike the case-control setting.

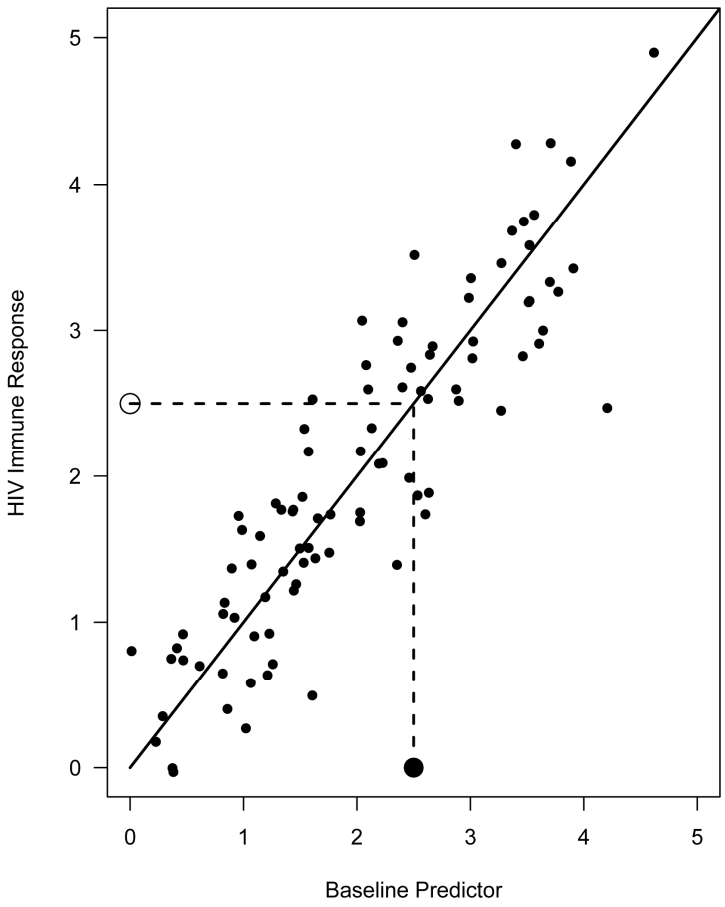

Figure 4. Scatterplot of the relationship between a baseline predictor and HIV immune response both measured in the vaccine group and a regression line. In the placebo group the counterfactual HIV immune response is unobservable, but can be imputed based on using the regression line. By randomization, we know this relationship, estimated in the vaccine group, must equally apply to the placebo group.

As noted by the authors, it is puzzling that the combination of CRPV and BIP results in less power than when BIP is used alone. Note that a simpler approach could be used where CRPV and BIP are used separately to attain parameter estimates. One could then bootstrap the entire procedure to obtain estimates of the variances of each and their covariance, say $\hat{(\beta)}_C$, $\hat{(\beta)}_B$, $\hat{var}((\beta)^\wedge_C)$, $\hat{var}((\beta)^\wedge_P)$, $\hat{cov}((\beta)^\wedge_P, (\beta)^\wedge_C)$. One could then form a new estimate $\hat{(\beta)}$ as a linear combination

$$(1) \qquad \hat{(\beta)} = \omega\hat{(\beta)}_C + (1 - \omega)\hat{(\beta)}_P.$$

DOI: 10.2202/1948-4690.1037

One would choose ω to minimize the variance of $\hat{\beta}$. Assuming that the estimates were each unbiased, a Wald test based on $\hat{\beta}$ could do no worse than taking ω = 0 and fashioning a Wald test of $\hat{\beta}_P$ alone. So it seems hard to believe that the likelihood method of combining both bits of information would do worse than BIP alone. It might be that one of the estimates is biased under these scenarios, and requires a much larger sample size to be unbiased. Perhaps the peculiar behavior disappears with a 10 or 100 fold increase in sample size. It also might be that a likelihood based test would perform better than a Wald test.

The power evaluations for the SoP assume a very substantial correlation of 0.8 between a measured immune response and a baseline predictor. This is probably quite optimistic and it would be sensible to additionally investigate power with smaller correlations such as 0.4. Under such settings, CRPV may be more important. Additionally, it would be sensible to try to identify such a correlate before the trial is launched.

With a continuous outcome Y, and a baseline covariate W, clinical trials will sometimes use a regression model of analysis of covariance ANCOVA to estimate the treatment effect. That is, the following model is estimated

$$(2) \qquad Y_i = \beta_0 + \beta_1 Z_i + \beta_2 W_i + e_i$$

where Z_i is the treatment indicator for the ith patient and e_i the error variance, assumed normal with mean 0 and variance σ^2. A test of the treatment effect H_0: $\beta_1 = 0$ in this model can be much more efficient than the two sample t-tests as the variance of Y_i can be much larger than the variance of e_i.

To explore whether such an approach might help the power of the CRPV design, a small simulation was conducted. Let Y be the infection indicator, X be the immune response to the HIV vaccine (not observed in the placebo group), Z the vaccine indicator, and W a good predictor of infection measured at baseline. We assume that

$$(3) \qquad P(Y = 1) = \Phi\,(\beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 W_i)$$

where X_i is missing in the vaccine group.

We estimate the coefficient of X_iZ_i under two models that either incorporate or ignore W_i:

- The properly specified model probit (2) that incorporates W_i. Here the coefficient for X_iZ_i is β_3
- A probit model that ignores W_i. Note that the parameters of this model are given by

$$(4) \qquad P(Y = 1) = \int \Phi(\beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 W_i)\, \varphi(W; 0\ \sigma^2) dW$$

$$= \Phi[(\beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i)/\sqrt{(1+ (\beta_4\sigma)^2)}]$$

Thus the coefficient for X_iZ_i here is attenuated and equals

$$(5) \qquad \beta_3/\sqrt{(1+ (\beta_4\sigma)^2)}.$$

We generated 1,000 clinical trials under (3) with $\beta_0\beta_1\beta_2\beta_3$ = (-1.405, -0.086, 0, -0.36), as in the Causation scenario of Follmann (2006), and set β_4 = -0.4. Thus W_i is a good predictor of risk, slightly stronger than the immune response. The parameters of the two models were estimated using maximum likelihood.

Table 5 indicates that the variance of the two approaches is very similar. The estimate based on the model that incorporates W is unbiased for β_3, while the estimate based on the model that ignores W is unbiased for (5). Disappointingly, there seems to be no efficiency gain for the parameter of interest by using a model that incorporates risk.

Table 4. Estimates of β_3 using Crossover Placebo Vaccination using two models: one that incorporates a strong predictor of risk (W) and one that ignores W

	Method	
Simulation	**Incorporate *W***	**Ignore *W***
Average	-0.3606	-0.3351
Variance	0.0318	0.0329

DOI: 10.2202/1948-4690.1037

Table 5. Estimates of β_3 using Crossover Placebo Vaccination using two models: one that incorporates a strong predictor of risk (*W)* and one that ignores *W*. The strength of the predictor is reflected by β_4.

	Method		
β_4	**Simulation**	**Incorporate W**	**Ignore W**
0	Average	-0.3528	-0.3644
	Variance	0.0534	0.0506
-0.4	Average	-0.3606	-0.3351
	Variance	0.0318	0.0329
-0.8	Average	-0.3632	-0.2845
	Variance	0.0189	0.0204

References

The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. NEJM. 1980. 303:1038-1041.

Follmann, D. Augmented designs to assess immune response in vaccine trials. Biometrics. 2006. 62:1161-1169.

Gilbert PB, Grove D, Gabriel E, Gray G, Self SG, Kublin J, and Corey L. A sequential phase 2b trial design for evaluating vaccine efficacy and immune correlates for multiple HIV vaccine regimens. 2011. Stat Comm Infec Dis.

Qin L, Gilbert PB, Corey L, McElrath MJ, and Self SG. A framework for assessing immunological correlates of protection in vaccine trials. 2007. JID. 196:1304-1312.

References

Anderson RM, Garnett GP. Low-efficacy HIV vaccines: potential for community-based intervention programmes. Lancet **1996**; 348: 1010–1013.

Anderson RM, Swinton J, Garnett GP. Potential impact of low efficacy HIV-1 vaccines in populations with high rates of infection. Proc Biol Sci **1995**; 261:147–151.

Abu-Raddad L, Boily M-C, Self SG, Longini IM Jr. Analytic insights into the population level impact of imperfect prophylactic HIV vaccines. J Acq Imm Def Synd Hum Retrov **2007**; 45:454–467.

Betensky, RA. Construction of a continuous stopping boundary from an alpha spending function. Biometrics **1998**; 54:1061−1071.

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime Data Analysis **2000**; 6: 39–58.

Buchbinder SP, Mehrotra DV, Duerr A, et al. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): A double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet **2008**; 372:1881–1893.

Burton DR, Desrosiers RC, Doms RW, et al. A sound rationale needed for phase III HIV-1 vaccine trials. Science **2004**; 303:316.

Czeschinski PA, Binding N, Witting U. Hepatitis A and hepatitis B vaccinations: immunogenicity of combined vaccine and of simultaneously or separately applied single vaccines. Vaccine **2000**; 18: 1074-1080.

Ellenberg S, Fleming T and DeMets D. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. John Wiley & Sons, Ltd., West Sussex, England, **2002**.

Emerson, S.S. Issues in the use of adaptive clinical trial designs. Statistics in Medicine **2006**; 25: 3270-3296.

DOI: 10.2202/1948-4690.1037

Emerson S, Fleming TR. Adaptive methods: Telling 'the rest of the story.' Journal of Biopharmaceutical Statistics **2010**; 20:1150-1165.

Emerson SS, Fleming TR. Symmetric group sequential test designs. Biometrics **1989**; 45:905–923.

Freidlin B, Korn EL, Gray R. A general inefficacy interim monitoring rule for randomized trials. Clinical Trials **2010**; 7:197-208.

Fleming, T.R. Standard *versus* adaptive monitoring procedures: a commentary. Statistics in Medicine **2006**; 25: 3305-3312.

Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF, and the rgp120 HIV Vaccine Study Group. Placebo-controlled trial of a recombinant glycoprotein 120 vaccine to prevent HIV infection. J Infect Dis **2005**; 191:654–665.

Follmann D. Augmented designs to assess immune response in vaccine trials. Biometrics **2006**; 62:1161–1169.

Forthal D, Gilbert P, Landucci G, Phan T. Recombinant gp120 vaccine-induced antibodies inhibit clinical strains of HIV-1 in the presence of Fc receptor-bearing effector cells and correlate inversely with HIV infection rate. Journal of Immunology **2007**; 178:6596-6603.

Frangakis CE, Rubin DB. Principal stratification in causal inference. Biometrics **2002**; 58:22-29.

Gilbert PB. Some design issues in Phase 2b prevention trials for testing efficacy of products or concepts. Statistics in Medicine **2010**; 29:1061-1071.

Gilbert PB, Hudgens MG. Evaluating candidate principal surrogate endpoints. Biometrics **2008**; 64:1146-1154.

Gilbert, PB, McKeague I, Sun Y. The two-sample problem for failure rates depending on a continuous mark: an application to vaccine efficacy. Biostatistics **2008**, 9:263-276.

Gilbert PB, Peterson M, Follmann D, Francis D, Gurwith M, Heyward W, Hudgens M, Jobes D, Popovic V, Self S, Sinangil F, Burke D, Berman P. Immunologic responses to rgp120 vaccine correlate with the incidence of HIV-1 infection in a Phase 3 preventive HIV-1 vaccine trial. Journal of Infectious Diseases **2005**; 191:666–677.

Gilbert PB, Qin L, Self SG. Evaluating a surrogate endpoint at three levels, with application to vaccine development. Stat Med **2008**; 27:4758–4778.

Gilbert P, Wei LJ, Kosorok MR, Clemens JD. Simultaneous inference on the contrast of two hazard functions with censored observations. Biometrics, 2002; 58:773-780.

Goonetilleke, N, Liu, MKP, Salazar-Gonzalez, JF, et al. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. Journal of Experimental Medicine **2009**; 206: 1253-1272.

Gray GE, Bekker L, Churchyard GJ, et al. Did unblinding affect HIV risk behaviour and risk perception in the HVTN503/Phambili study? Retrovirology **2009**; 6 (Suppl 3): P209.

Heyse, JF, Kuter, BJ, Dallas, MJ, Heaton, P for the REST Study Team. Evaluating the safety of a rotavirus vaccine: the REST of the story. Clinical Trials **2008**; 5: 131–139.

Horne AD, Lachenbruch PA, Goldenthal KL. Intent-to-treat analysis and preventive vaccine efficacy. Vaccine **2001**; 19:319-326.

Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika **1983**; 70:659-663.

Lu X, Tsiatis AA. Improving the efficiency of the log-rank test using auxiliary covariates. Biometrika **2008**; 95:679--694.

Karim QA, Karim SS, Frohlich JA, Grobler AC, Baxter C, Mansoor LE, Kharsany ABM, Sibeko S, Misana KP, Omar Z, Gengiah TN, Maarschalk S, Arulappan N, Mlotshwa M, Morris L, Taylor D and on behalf of the CAPRISA 004 Trial Group. Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. Science **2010**; 3:1168-1174.

DOI: 10.2202/1948-4690.1037

Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. Stat Med **2009**; 28:39–64.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics **1979**; 35: 549-556.

Pitisuttithum P, Gilbert P, Gurwith M, et al. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J Infect Dis **2006**; 194:1661–1671.

Plotkin SA. Vaccines: Correlates of vaccine-induced immunity. Clin Infect Dis **2008**; 47:401−409.

Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika **1977**; 64:191-199.

Qin L, Gilbert P, Corey L, McElrath J, Self S. A framework for assessing immunological correlates of protection in vaccine trials. J Infect Dis **2007**; 196:1304–1312.

Qin L, Gilbert P, Follmann D, Li D. Assessing surrogate endpoints in vaccine trials with case-cohort sampling and the Cox model. Annals of Applied Statistics **2008**, 2:386-407.

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med **2009**; 361:2209–2220.

Robins JM, Greenland S. Identifiability and exchangeability of direct and indirect effects. Epidemiology **1992**; 3:143--155.

Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. J Roy Stat Soc Ser A **1984**; 147:656--666.

Self SG. Issues in the design of HIV vaccine efficacy trials. In: *Accelerating AIDS Vaccine Development: Challenges and Opportunities* (Patricia Kahn, Ian Gust and Wayne Koff, editors). Horizon Scientific Press, Norfolk: United Kingdom, **2006**.

Siegmund D. Sequential Analysis: Tests and Confidence Intervals. Springer-Verlag, New York, **1985**.

Scharfstein DO, Rotnitzky A, and Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association **1999**; 94: 1096-1146.

Shepherd B, Gilbert P, Lumley T. Sensitivity analyses comparing time-to-event outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. Journal of the American Statistical Association **2007**; 102:573-582.

Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in Medicine **2008**; 27:4658--4677.

Wald A. Sequential Analysis. **1947**; Wiley, New York.

Wolfson J, Gilbert P. Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. Biometrics **2010**; 66:1153-1161.

Zhang M, Gilbert P. Increasing the Efficiency of Prevention Trials by Incorporating Baseline Covariates. Statistical Communications in Infectious Diseases **2010**; 2. DOI: 10.2202/1948-4690.1002 Available at: http://www.bepress.com/scid/vol2/iss1/art1

Zhang M, Tsiatis AA, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics **2008**; 64; 707–715.

DOI: 10.2202/1948-4690.1037