Published in final edited form as:
Stat Med. 2010 May 10; 29(10): 1061–1071. doi:10.1002/sim.3676.

SOME DESIGN ISSUES IN PHASE 2B VERSUS PHASE 3 PREVENTION TRIALS FOR TESTING EFFICACY OF PRODUCTS OR CONCEPTS

Peter B. Gilbert
Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington, 98109, U.S.A

SUMMARY

After one or more Phase 2 trials show that a candidate preventive vaccine induces immune responses that putatively protect against an infectious disease for which there is no licensed vaccine, the next step is to evaluate the efficacy of the candidate. The trial-designer faces the question of what is the optimal size of the initial efficacy trial? Part of the answer will entail deciding between a large Phase 3 licensure trial or an intermediate-sized Phase 2b screening trial, the latter of which may be designed to directly contribute to the evidence-base for licensing the candidate, or, to test a scientific concept for moving the vaccine field forward, acknowledging that the particular candidate will never be licensable. Using the HIV vaccine field as a case study, we describe distinguishing marks of Phase 2b and Phase 3 prevention efficacy trials, and compare the expected utility of these trial types using Pascal's decision-theoretic framework. By integrating values/utilities on (1) Correct or incorrect conclusions resulting from the trial; (2) Timeliness of obtaining the trial results; (3) Precision for estimating the intervention effect; and (4) Resources expended; this decision framework provides a more complete approach to selecting the optimal efficacy trial size than a traditional approach that is based primarily on power calculations. Our objective is to help inform the decision-process for planning an initial efficacy trial design.

Keywords

Clinical trial; Decision analysis; HIV vaccine; Intermediate-sized efficacy trial; Licensure; Microbicide; Phase 2b versus Phase 3

1. INTRODUCTION

The traditional sequence of clinical trials for developing a preventive vaccine entails Phase 1, 2, and 3, where there is a large jump in resources moving from Phase 2 to 3, typically in the order of a 10–20-fold jump in sample size and a 2-fold jump in trial duration. Recently the HIV vaccine field has moved to a different paradigm wherein Phase 2 trials are followed by intermediate-sized, "Phase 2b" efficacy trials. Here we describe this evolutionary history, with goals to describe some of the key characteristics that distinguish Phase 2b from Phase 3 vaccine efficacy trials, and to scaffold a framework for deciding the optimal efficacy trial type and size for following a successful Phase 2 trial. While we use HIV vaccine efficacy trials as an illustrative example throughout, the discussion also pertains to efficacy trials of non-HIV vaccines and of other preventative interventions such as microbicides.

Correspondence to: Peter Gilbert, Vaccine Infectious Disease Institute, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington, 98109, U.S.A. Phone: (206) 667-7299, pgilbert@scharp.org.

In Section 2 we summarize the history of Phase 2b and 3 HIV vaccine efficacy trials, and define and describe the objectives of these trial types. In Section 3 we suggest some distinguishing marks between these trial types. In Section 4 we conduct an analysis comparing the expected utility of two possible decisions: to follow a Phase 2 trial with a 2b trial versus with a Phase 3 trial. The decision analysis also compares the expected utility of various 2b designs indexed by the total number of study endpoints. This analysis can be viewed as a sequel to the comparative decision framework considered by [1]. In Section 5 we discuss concluding points and open questions that bear further research.

2. HISTORY OF PREVENTIVE HIV VACCINE EFFICACY TRIALS

2.1 History of Phase 3 Efficacy Trials

Initially, the HIV vaccine field focused on the development of recombinant glycoprotein (rgp) envelope vaccines, which were designed to prevent HIV infection by stimulating anti-HIV neutralizing antibodies. Phase 1 and 2 trials in the nineties showed that VaxGen, Inc.'s rgp candidate vaccine (named AIDSVAX) could induce antibodies that neutralized strains of HIV grown in the laboratory, and challenge studies showed that AIDSVAX prevented a small number of chimpanzees from being infected with these strains genetically matched to the vaccine strain. Due to the inability of AIDSVAX to neutralize "primary" HIV strains that expose people in real life [2], and to the fact that the challenge strain was "wimpy" (selected to be as easy as possible to protect against), in 1994 most scientists believed AIDSVAX could not protect against HIV infection. This belief was reflected in the National Institute of Allergy and Infectious Diseases' (NIAID's) decision in 1994 to not fund efficacy trials of AIDSVAX. Despite AIDSVAX's low plausibility of efficacy, with private financing VaxGen conducted two parallel Phase 3 efficacy trials of AIDSVAX from 1998 and 2003, in men who have sex with men in North America and The Netherlands [3] and in intravenous drug users in Thailand [4] (Table 1).

With VE the vaccine efficacy, defined as one minus the hazard ratio (HR) (vaccine group/placebo group) of HIV infection in study volunteers who are initially HIV negative, the VaxGen trials were powered to test the null hypothesis $H_0 : VE \leq .3$ versus $H_1 : VE > .3$. Following communications with the U.S. FDA on criteria for a licensable HIV vaccine, a null of 30% efficacy was used instead of 0% because mathematical models suggest that substantial efficacy is required to support that deployment of a vaccine will confer public health benefit in terms of reduced numbers of new HIV infections and AIDS deaths (e.g., [5]). In the North American and Thailand VaxGen trials 368 and 211 volunteers acquired HIV infection, respectively, with $\widehat{VE}=.07$ (95% CI-.17 to .24) and $\widehat{VE}=.01$ (95% CI-.31 to .24). Therefore, with high precision these Phase 3 trials showed that AIDSVAX did not prevent HIV infection.

In general, licensure of a treatment or vaccine in the U.S. requires "robust and compelling" evidence of a favorable benefit-to-risk profile [1]. The traditional criterion for establishing this evidence for an HIV vaccine– which was operative in the VaxGen trials– is two placebo-controlled Phase 3 trials each showing $VE > .3$ at the standard significance level of a 1-sided p-value $\leq .025$ [6]. However, for candidate vaccines with low pre-test plausibility of efficacy, this approach is a risky gamble, because it expends great resources for a long-shot hope for success. Based on this concern, HIV vaccine scientists devised an alternative path to licensure, which follows successful Phase 1/2a trials and animal challenge trials with a small screening efficacy trial, which has been called a Phase 2b intermediate-sized efficacy trial [7].

2.2 History of Phase 2b Efficacy Trials

Three Phase 2b preventive HIV vaccine efficacy trials have been conducted (Table 1); generally Phase 2b trials are designed to accrue between a quarter and one-third the number of events that would be assessed in a Phase 3 trial. Results are available from one of these Phase 2b studies, the “Step trial” [8]. Step enrolled men who have sex with men and women, of whom 82 and 1 acquired HIV infection. In the 1,836 randomized men the estimated HR was 1.5 (95% CI 1.0 to 2.4), raising the hypothesis that the vaccine increased susceptibility to acquisition of HIV infection. Moreover, the HR was particularly elevated in uncircumcised men and in men with prior exposure to Adenovirus serotype 5 (Ad5), the vaccine vector that carried the HIV genes. The result of potential harm in men with prior Ad5 exposure has had a major impact on the HIV vaccine field, leading to a research climate in which investigators are reluctant or precluded from testing any vector-based HIV vaccine in persons who have prior exposure to the vector. For example, individuals with prior exposure to vaccinia are not enrolled into current trials of modified vaccinia ankara-vector based candidate vaccines. Furthermore, trials of Adenovirus vector-based candidate vaccines for non-HIV pathogens (e.g., malaria) have been stopped. The after-effects of the Step trial show that a small Phase 2b trial, and an unplanned subgroup analysis with 38 primary endpoint events at that, can majorly impact vaccine research programs. The question arises as to whether a small 2b trial provides a sufficient evidence base for supporting the magnitude of impact. Decision analysis as illustrated in Section 4 provides a framework for addressing this type of question.

2.3 Objectives and Definitions of Phase 2b and 3 Efficacy Trials

The objective of a Phase 3 trial is to characterize clinical efficacy and clinical safety of a specific vaccine product, to directly support a licensure decision. As such, during the design of a Phase 3 trial the vaccine developer communicates closely with the appropriate regulatory agency to define what observed effects on what endpoints will be sufficient to justify licensure. The discussions will include a decision on whether one or two Phase 3 trials will be required to support licensure. A single Phase 3 trial may suffice if the disease is rare and Phase 3 trials are very costly; however the standard of statistical evidence is more stringent for a single Phase 3 trial than for each of two separate Phase 3 trials. For example, the criterion for vaccine licensure based on a single Phase 3 trial may require a 1-sided *p*-value less than $.025^2 = .000625$ or $.025^{1.5} = .004$ for supporting a beneficial vaccine effect on the primary efficacy endpoint [1,6].

A Phase 2b screening trial is harder to define, as there are different types with different objectives. We distinguish two types– the first (a “product 2b”) is a small version of a Phase 3 trial of a specific product, which may directly support licensure of the product; and the second (a “test-of-concept (TOC)” 2b) aims not to generate evidence supporting licensure of a specific product, but rather to test the merit of a vaccine concept. For both 2b trial types the main objective is to support a decision to either (1) establish that the product or concept is useless, leading to termination of its clinical testing; or (2) establish that the product or concept is plausibly clinically efficacious, leading to its further clinical efficacy testing. For product 2b trials for which the primary endpoint is a clinical endpoint or a validated surrogate endpoint, the decision guideline may have two additional possible decisions: (3) establish “1 Phase 3-level evidence” for efficacy, leading to a second Phase 3 trial; or (4) establish “2 Phase 3-level evidence” for efficacy, leading to licensure of the product. Fleming and Richardson [1] described this kind of 4-level decision guideline for product 2b trials, and applied it in the design of the HPTN 035 trial that tested the efficacy of two microbicide gels to decrease the rate of HIV infection [9]. When choosing the optimal initial efficacy trial design among various product 2b designs and Phase 3 designs (addressed in Section 4), the evidence levels specified in (3) and (4) typically influence the decision less than other factors, especially if there is only a remote chance to establish 2 Phase 3-level evidence.

3. DISTINGUISHING MARKS OF PHASE 2B AND PHASE 3 EFFICACY TRIALS

Beyond their different objectives, differences in study population, endpoints, and analysis distinguish 2b from 3 effcacy trials. For a Phase 3 trial, it is desirable to select a study population as heterogeneous as possible within the bounds of plausible efficacy, to robustly support safety and to support the broadest indication for licensure. In contrast, for 2b trials that aim to weed out a product or concept that typically has low pre-test plausibility of efficacy, it is logical to select a study population that offers the best chance for finding some efficacy. This is why the Step trial was initially designed to evaluate the vaccine in individuals with low prior immunity to Adenovirus 5; these volunteers had the highest HIV-specific immune responses and were thought to be the most easily protected by vaccine.

3.1 Distinguishing Mark: Study Endpoint

Because licensure of a product is supported by evidence of a favorable benefit-to-risk profile, Phase 3 trials use a primary endpoint that is either clearly clinically significant (e.g., symptomatic disease plus isolation of the infecting pathogen) or is a surrogate endpoint that is reasonably well-validated. HIV infection is an example of a validated surrogate for the clinical endpoint AIDS. For product 2b trials with potential to support licensure of the tested product, use of a clinical endpoint or validated surrogate is important. On the other hand, for TOC 2b trials that are not designed to support licensure decisions, there is more flexibility to consider the use of surrogate endpoints that are less completely validated. In particular, if the use of a surrogate endpoint allows the trial to be done much more quickly and cheaply than the use of a clinical endpoint, then one may accept a small risk of a false negative result (i.e., the scenario where there is no beneficial vaccine effect on the surrogate but there is on the clinical endpoint) in order to quickly screen the candidate for plausible efficacy.

The size of the pool of available candidate products to potentially screen for efficacy impacts the pros and cons of using an unvalidated surrogate endpoint. If the pool is very small, with one or two viable products or prototypes, then it is paramount to use reliable endpoints, as a false negative result that leads to the discarding of an auspicious approach is the most costly error. If the pool is large, on the other hand, it may be more important to rapidly screen a large number of candidates, in which case it may be acceptable to miss some promising products for the sake of rapidly advancing a subset of auspicious candidates. For the HIV vaccine field with a small pool of available candidates, the Step, Phambili, and RV 144 (Table 1) designs all used the validated surrogate primary endpoint of HIV infection, although they also used the unvalidated co-primary surrogate endpoint of viral load measured 1–5 months after HIV infection diagnosis.

3.2 Distinguising Mark: Balance of Type 1 and 2 Errors

Related to endpoints, the importance of controlling false positive versus false negative errors depends on the type of trial. For hypothesis tests in Phase 3 and product 2b trials that can influence licensure, stringent false positive error control is paramount (e.g., using a traditional 1-sided $p \leq .025$), as it is a very costly error to license an ineffective product. For TOC 2b trials that will not support licensure decisions, again the size of the pool of available candidates influences the pros and cons. If the pool is very small, then a false negative error is the worst mistake, whereas a false positive result would incur the smaller cost of leading to a larger efficacy trial, which would, with high likelihood, correctly establish the uselessness of the product. Therefore for TOC 2b trials in fields with lean product pipelines, the traditional design approach that controls the type 1 error rate at 1-sided .025 and the type 2 error rate at .10 or .20 may be unfitting to the scientific objectives, and it may be prudent to control the type 2 error at least as stringently as the type 1 error rate (e.g., with 1-sided α = .10 and power of 95%). These considerations also apply to hypothesis tests in product 2b trials that evaluate uselessness versus plausible efficacy, as illustrated in Section 4.

In the following section we consider product 2b trials but not TOC 2b trials.

4. COMPARATIVE EXPECTED UTILITY OF PRODUCT PHASE 2B AND 3 EFFICACY TRIALS

In this section we compare the expected utility of various product Phase 2b vaccine efficacy trial sizes to that of a prototype Phase 3 trial, as an initial efficacy trial for following a successful Phase 2 trial. Our decision-analysis is motivated by the observation that traditional power calculations are indequate for providing quantitative guidance about the relative merits of these designs. Traditional power calculations ignore or insufficiently account for important considerations including beliefs about the likely level of VE, the relative importance of certain correct or incorrect conclusions from the trial, the timeliness of the study results, and the resources expended.

A failure to account for pre-test beliefs about the distribution of VE can result in a trial design with a surprisingly low chance of achieving the hoped-for result. Specifically, if a candidate vaccine has a low pre-test plausibility of efficacy (a ‘long shot’), then the probability of detecting an efficacy signal at a certain alternative hypothesis is much less than the power. To illustrate this, consider a 90 endpoint prototype 2b trial, which has 90% power to reject $H_0 : VE \leq 0$ at 1-sided $\alpha = .10$ if the true $VE = .42$. Suppose $\log(1 - VE)$ (i.e., the log HR) is normally distributed with mean μ and standard deviation σ, with $\mu = -.105$ chosen to reflect a pre-test guess that the vaccine has minimal efficacy of $VE = .10$ (.10 = 1-exp(−.105)), and $\sigma = .34$ chosen to reflect a pre-test guess that there is only a 30% probability that the true VE is at least .42. In this case of modest expectations the probability of rejecting H_0 and thus declaring efficacy is only .35.

4.1 Guideline for Decision-Making in a Product 2b Trial

We formulate the product 2b trial similarly as Fleming and Richardson [1], who specified a decision guideline such that the analysis of the sole primary endpoint, HIV infection, leads to one of several possible decisions about what to do next with the product. In our formulation, every possible value of the estimated VE leads to exactly one of the possible decisions to (1) Declare harm; (2) Declare useless, (3) Declare plausibly efficacious and advance the product to a Phase 3 trial; (4) Declare efficacious at one Phase 3 .025-level evidence; and (5) Declare efficacious at two Phase 3 $.025^2 = .000625$-level evidence. This decision framework is the same as in [1] except that the possibility to declare harm is added, and we require a smaller p-value for declaring two Phase 3 level evidence ($.025^2$ compared to $.025^{1.5}$). Either outcome (1) or (2) would lead to discontinuation of product development. In practice information on the vaccine effect on other endpoints would be taken into account to inform decision-making; however for clarity we focus on the case that the guideline is based on a single endpoint.

A guideline for making the decision can be defined by three 1-sided p-values and by pre-set significance thresholds for what constitute significant results. The three p-values are for (A) Testing harm ($H_0 : VE \geq 0$ vs $H_1 : VE < 0$), with p-value p_{harm}; (B) Testing plausible efficacy ($H_0 : VE \leq 0$ vs $H_1 : VE > 0$), with p-value p_{plaus}; and (C) Testing efficacy ($H_0 : VE \leq .3$ vs $H_1 : VE > .3$); with p-value p_{eff} . Figure 1 shows a particular guideline, which we use throughout this section. It illustrates that a guideline is defined equivalently in terms of *p*-value thresholds or confidence limit thresholds. With this guideline for a prototype Phase 2b trial with 90 infections, Table 2 illustrates the advantage of using a $p_{plaus} \leq .10$ significance threshold for declaring plausibility of efficacy versus an alternative guideline that would use the traditional .025 significance threshold. If the vaccine is promising such that $VE = .3$, then the traditional guideline provides a 38% chance of moving the candidate vaccine to further testing (i.e., hitting decisions (3), (4), or (5)), whereas the selected guideline provides a 65% chance of moving the candidate ahead.

As a comparator to the Phase 2b design, we consider a prototype Phase 3 design with 300 total infections. The decision-framework is the same as for the Phase 2b design except that the *p*-value significance thresholds for decisions (1)–(3) are all set at the traditional .025 level, and the decision to "declare plausibly effcacious" is replaced with "declare low efficacy." Figure 1 shows the confidence interval formulation of this selected Phase 3 guideline.

For a fixed number of total HIV infections *n*, there is a one-to-one correspondence between *p*-value thresholds that define the decision guideline and cut-point values of estimated VE that distinguish the five possible decisions. This one-to-one correspondence is established using the asymptotic distribution of the log HR estimate (see page 394 of [10]); the cut-point estimated VE values are determined once *n* and the *p*-value significance thresholds are specified. For example, Table 3 shows the cut-points that result from a Phase 2b design with 90 infections and from a Phase 3 design with 300 infections. For different values of true VE, the table shows the probability of declaring each of the five decisions. As expected, for the larger trial there is a greater probability of making a correct decision; for example if the true $VE = 0$ then the correct decision would be to declare the vaccine useless, which will be done with probability .85 for the 2b trial and probability .95 for the Phase 3 trial.

4.2 Decision-Theoretic Analysis of Comparative Expected Utility

Now we perform an elementary decision theoretic analysis to compare the expected utility of the above product 2b design with *n* total infections to the above Phase 3 design with 300 infections. There is a vast literature on decision analysis; for example [11–13] are introductory textbooks for biomedical applications. We use the original decision theory framework developed by Blaise Pascal in Paris gambling parlors in the 1640s; Ian Hacking described Pascal's famous Wager argument [14] as "the first well-understood contribution to decision theory [15]."

In a decision problem, the state of reality, and the decision of an agent, determine an outcome for the agent. Each possible outcome is assigned a utility, which represents the value the agent places on the outcome. The utility values can be arranged in a matrix, in which the columns represent the possible states of reality, and the rows represent the various decisions the agent can make. The agent making the decision may not know the current state of reality, but can assign a probability value to each state. Then the expected utility of each decision summarizes its merit, and the decision with the highest expected utility is considered to be the best (most rationale) decision, given these probabilities and utilities.

In our formulation, the five possible outcomes are the five possible conclusions of the trial (declare harm, etc.), and the possible decisions of the agent are to initiate a Phase 2b trial with *n* total events (we consider *n* between 45 and 150) or a Phase 3 trial with 300 total events. We assign a utility to each cell of the matrix by factoring in four factors: (1) Correct vs incorrect decision-making; (2) Speed to reach a decision; (3) Precision for estimating VE; and (4) Resources. While a relatively complete decision analysis would entail considerable background work and consensus building for choosing utilities in each dimension (as well as for choosing prior probabilities for VE), due to space and time constraints we focus on illustrating one particular choice of utilities, followed by a brief sensitivity analysis.

We perform the decision analysis for each of five underlying true values of VE (−.5, 0, .3, .5, .7); this allows demonstration of comparative utility for vaccine candidates with various pre-test beliefs about their plausibility for efficacy. For factor (1), we assign a positive utility of 1.0 for a correct decision, and assign negative utilities for incorrect decisions (Table 4a). The greater magnitude of negative utility reflects a worse outcome; for example if $VE = -.5$ but 2 Phase 3-level evidence for efficacy is declared, the outcome is very bad indeed as reflected by utility −5. Alternative utilities are considered in Section 4.4; in practice the decision-analysis would be repeated for a range of utilities reflecting the views of all stakeholders. For factor (2), faster correct decisions are more valuable than slow ones, and we suppose the ratio of utilities of speed (for a 2b trial with *n* events versus a Phase 3 trial with 300 events) equals the reciprocal of the ratio of the trial durations. We suppose that a 45 (150) endpoint 2b trial takes 2 (3) years and a 300 endpoint phase 3 trial takes 4.5 years, and use linear interpolation to set the trial duration of 2b trials with between 45 and 150 endpoints. For factor (3), more precise estimates of VE are more valuable, and we suppose the ratio of utilities for precision (for a 2b trial with *n* events versus a Phase 3 trial with 300 events) equals the ratio of 95% confidence interval widths for the log HR, which equates to $\sqrt{n/300}$. For factor (4), the relative amount of resources expended to conduct a trial depends on many inputs including the ratio of the number of infection events, the percentage of total trial costs in each of several cost-categories, and the “scale-up” elasticity factor for each cost category, which reflects the percentage cost increase for each additional infection event. Based on meetings at the Vaccine Infectious Disease Institute with experts in all areas of HIV vaccine efficacy trials that our group conducts, we considered nine cost categories defined by the cross-classification of trial period (pre follow-up, follow-up, post follow-up) with activity type (sites and operations, statistical data management center, lab). Across the three activity types, respectively, the estimated cost fractions were .12, .03, .04 for pre; .47, .10, .16 for during; and .05, .01, .02 for post follow-up. Likewise the estimated percentage increases in cost per additional event were .30, .20, .20 for pre; 1.0, .50, 1.0 for during; and .30, .15, .75 for post follow-up. From these inputs we can compute the total cost ratio *Cost.ratio*(2*b n*/3 300) of a 2b trial with *n* events versus a Phase 3 trial with 300 events.

To compute an overall utility for each cell in the decision matrix, we first set each utility of the Phase 3 design to be the decision-making utility. This sets a benchmark with maximum possible utility of 1.0, which is achieved if the probability of a correct decision is 1.0. Next, the “precision-resource ratio” of the 2b trial with *n* events versus the Phase 3 trial, which heuristically reflects the “amount of science gleaned per time-and-money resource expenditure,” is the precision utility ratio divided by the product of the speed utility ratio and the monetary cost utility ratio, equal to

$$PR(n)=\frac{\sqrt{n/300}}{\frac{[2+(n-45)/(150-45)]}{4.5}\times \mathrm{Cost.ratio}(2\mathrm{b}\, n/3\ 300)}.$$

Then we set the utility of the 2b trial with *n* events equal to

$$u^{2b}(n)=u^{dec}\left\{PR(n)I(u^{dec}\geq 0)+\frac{1}{PR(n)}I(u^{dec}<0)\right\},$$

where u^{dec} is the decision-making utility.

To make the overall utility reasonable and interpretable, care is needed in the choice of scales for the component utilities, and in the method for combining them, which encodes a relative valuation of the components. The above formulation combines the component utilities multiplicatively, such that doubling one has the same effect on the overall utility as doubling any of the others. For example, it encodes the views that halving the trial duration is equally valuable as halving the the total monetary cost, and halving the confidence interval width is equally valuable as halving either the trial duration or the total monetary cost. In addition, the scale of the decision-making utility u^{dec} is interpreted relative to $PR(n)$ (which multiplicatively combines component utilities (2), (3), and (4)). Thus, for example, an incorrect decision with $u^{dec} = -1$ compared to an incorrect decision with $u^{dec} = -2$ encodes the view that it is worth twice as much precision-resources $PR(n)$ to arrive at the less incorrect decision.

4.3 Results of Comparative Expected Utility Analysis

Figure 2 shows the expected utilities for each assumed value of true VE. Figure 2b shows that for a useless vaccine ($VE = 0$), it is a more rationale decision to use a Phase 2b trial with any $n \in [45, 150]$ than a Phase 3 trial. The smallest 2b trial is best, about 3-fold better than the Phase 3, whereas the largest 2b is about 2-fold better. If the vaccine has low efficacy ($VE = .3$), then again the 2b trial is always superior, although now the largest 2b is best, about 3-fold better than the Phase 3, while the smallest 2b trial is about 60% better. If $VE = .5$, then Phase 2b trials with more than 115 events are superior to a Phase 3 but inferior otherwise, and small 2b designs perform poorly. The poor performance results in part because small 2b designs have relatively unstable probabilities of correct decisions. If the vaccine has substantial efficacy ($VE = .7$), then the Phase 3 trial has much greater utility than all of the 2b designs. These results provide some quantification of the intuitive notion that Phase 2b trials are superior for candidate vaccines with low pre-test plausibility of efficacy and Phase 3 trials are superior for candidate vaccines with high pre-test plausibility of efficacy.

Albeit a 2b trial of a candidate vaccine would not be conducted if the prevailing belief is that the vaccine will increase susceptibility to HIV infection, Figure 2a shows comparative expected utility if $VE = -.5$, which corresponds to a 50% higher rate of infection in the vaccine group. It shows that in this case the smallest Phase 2b trial has 1.2-fold less utility than a Phase 3 trial, and at least 52 events in a Phase 2b trial are needed to match the utility of the Phase 3. This suggests that a small Phase 2b trial is not well-suited for reliably detecting a harmful vaccine. Whether this is a problem with the design depends on the objective of the trial and its context in the whole vaccine field, both for the pathogen under consideration and for other pathogens. For the narrow objective of deciding whether to permanently screen out the candidate product, it may not be a serious diasadvantage, as with high likelihood the design will declare the product at best useless, leading to its discontinuance, and failure to detect the harmful effect will not lead to the exposure of future study volunteers. But if components of the product (e.g., vaccine vector or adjuvant) may be used in other future candidate vaccines against any pathogen, then the high risk of failure to detect vaccine-harm may be unacceptable.

Figure 3a shows comparative expected utilities for three different prior distributions on the five values of true VE. Each expected utility is computed as the weighted sum of VE-specific expected utilities shown in the five panels of Figure 2, with weights equal to the prior probabilities that $VE = -.5, 0, .3, .5$, or $.7$. The skeptic/pessimist has no hope that the vaccine can confer protection, and places prior probabilities of .2, .8, 0, 0, 0 on the five VE values, respectively. The believer/optimist in contrast places prior probabilities of 0, .2, .6, .

2, 0; thus s/he believes there is 80% chance of $VE \geq .3$. The centrist has prior distribution equal to the average of the two: .1, .5, .3, .1, 0. The results show that under all three prior distributions all of the Phase 2b designs have greater utility than the Phase 3 design, and that pessimists prefer small Phase 2b trials whereas optimists prefer large ones.

4.4 Results of Comparative Expected Utility Analysis: Sensitivity Analysis

Figures 3b and 3c repeat the decision analysis reported in Figure 3a using different decision-making utilities u^{dec}, which are specified in the middle and bottom sub-tables of Table 4. The “safety-first” individual is primarily concerned with protecting study participants, believing that the worst mistake is missing a harmful effect of the vaccine, and therefore assigns a particularly large negative utility to an incorrect decision that misses a harmful vaccine ($VE = -.5$). S/he is also less concerned with incorrectly declaring harm and with incorrectly missing an efficacious vaccine. S/he assigns decision-making utilties equal to the original u^{dec}’s except s/he doubles the negative values in the $VE = -.5$ row and halves the negative values in the “declare harm” column and in the $VE = .5$ and $VE = .7$ rows (Table 4b). In contrast, the “efficacy first” individual believes that the worst mistake is passing over a partially efficacious vaccine. His/her decision-making utilties modify the original utilities in the opposite way as the safety-first individual, by halving negative values in the $VE = -.5$ row and doubling negative values in the “declare harm” column and in the $VE = .5$ and $VE = .7$ rows (Table 4c).

Figure 3 shows that, under all scenarios, all of the Phase 2b designs are superior to the Phase 3 design. Pessimists tend to favor small 2b trials while optimists favor large ones. One’s disposition for safety-first versus efficacy-first has minimal influence on decision-making for the pessimist or the centrist, but has significant influence for the optimist, with the efficacy-first optimist preferring larger 2b trials.

Following a helpful suggestion from a referee, we also briefly evaluated the influence of each of the four component utilities on the relative merit of trial designs. We considered fixed levels of the component utilities as follows: (1) u^{dec} fixed at the original utilities; (2) Precision-ratio fixed at the level n midway between 45 and 150 (n = 98), $\sqrt{98/300}=0.572$; (3) Trial duration utility fixed at the level n = 98, [2 + (98 − 45)/(150 − 45)]/4.5 = 0.557; (4) Total monetary cost utility fixed at the level n = 98, Cost.ratio(2b 98/3 300) = 0.393. Under the centrist prior considered above, Figures 4(a)–(d) show the expected utilities holding all component utilities fixed except component (1)–(4), respectively; in (a) u^{dec} was varied over the levels safety-first, original, and efficacy-first; whereas in (b), (c), and (d) the utilities were varied as functions of n in the same way as done for Figures 2 and 3. The results show that the trial duration and monetary cost utilities have about the same influence, and that 2b designs always have more utility than a Phase 3, slightly moreso for smaller 2bs. In addition, the precision utility has the greatest influence, with utility of the 2b design sharply increasing with n.

In conclusion, the decision analysis shows that under a prior belief that the vaccine will have low efficacy at best (VE between 0 and 30%, say), the 2b trial is clearly superior to the Phase 3 trial. The prior probabilities for zero efficacy versus for low-to-moderate efficacy impact the optimal size of a Phase 2b trial, where the smallest design is favored under a belief of zero efficacy, and increasingly larger designs are favored as the believed level of efficacy increases. Furthermore a belief in high enough efficacy makes the Phase 3 design optimal. Only if there are good reasons to believe that the efficacy is fairly high, or if the utilities reflect much greater value in correctly detecting useful vaccines than in correctly screening out useless ones (more extreme than we have considered), is it rational to skip the Phase 2b design and go straight to Phase 3.

5. DISCUSSION

As intermediate-sized Phase 2b efficacy trials for vaccines and other prevention interventions are increasingly used, trial designers face many questions about their optimal design. For two-group placebo-controlled prevention trials with a single time-to-event primary endpoint such as HIV infection, we have discussed several of these questions, including what is the trial objective (to potentially support licensure or to test a concept), what are the appropriate study endpoints and controls placed on false positive and false negative error rates, and what is the optimal number of primary endpoint events n. We have addressed these questions with a simple illustrative decision analysis that compares the expected utiilities of different product Phase 2b and 3 designs indexed by n; such a decision analysis provides more complete guidance than standard power calculations. The comparison of expected utilities for a range of candidate designs, where, for each design, a range of utility values for the possible trial outcomes and a range of probabilities of these outcomes are considered, provides a framework for the trial design team to understand the comparative strengths, weaknesses, and assumptions of the various candidate designs, and ultimately to decide on the final design.

A challenge posed to selecting the final design is that the trial design team– which may consist of diverse stake-holders spanning product developers to regulatory authorities– must come to rough agreement on what are the appropriate utilities, or, at least, an appropriate range of utilities. As we have discussed, these utilities depend on many partly subjective factors, including (1) Views about “how bad” is each possible incorrect decision/inference, which depend on many attributes of an individual’s viewpoint including his/her weighing of individual-versus-societal ethics; (2) Views about the relative importance of different trial attributes such as risks of incorrect decisions, timeliness of achieving results, precision for estimating the intervention effect, and resources expended; and (3) Views on how to measure and integrate the many trial attributes in (2) into an overall utility. In addition, the trial design team must come to rough agreement on the pre-test distribution of the true level of intervention efficacy. While achieving rough agreement on all these factors and others may be daunting, the alternative is worse– foregoing an open discussion and winding up with a design with a weak articulation of why it is appropriate. Any selected design will be near-optimal for some sets of utilities and pre-test efficacy distributions and nowhere near-optimal for others; therefore picking a design without an explicit discussion amounts to a hidden viewpoint deciding the design. One can imagine an unscrupulous statistician secretly performing a decision-analysis, deriving the optimal design under his/her subjective utilities and pre-test efficacy distribution, and suggesting it to the trial design team without showing the decision-analysis. Better that the whole team engage in a thorough discussion so that the values, philosophies, and assumptions underlying the different designs are transparent.

In our illustrative example (the HIV vaccine field), views on the pre-test distribution of true vaccine efficacy has seemed to play a particularly strong role in decision-making. In the workshop where this work was presented Don Berry reminded us that power calculations often ignore these pre-test efficacy distributions, in which case key probabilities of interest– for declaring efficacy under various alternative hypotheses– are not stated. VaxGen’s decision to spend more than $100 million on its two Phase 3 trials is only rational under a pre-test high probability that the vaccine would be quite effective or under the utilities of a gambler who is willing to take a great risk for a long-shot at efficacy. In contrast, NIAID’s 1994 decision to not publicly fund VaxGen’s trials can be understood in terms of their advisory committees’ assignment of low probability that the vaccine would have any efficacy and its utilities that are more cautious than those of a bold gambler.

We have emphasized that the context of a particular trial impacts design decisions. For example, requiring stringent false positive error control seems generally important for licensure trials, but for Phase 2b product trials, and even moreso for Phase 2b test-of-concept trials, there are occasions where false negative error control is more important. In particular, if the pipeline of candidate interventions is thin, and the objective of the Phase 2b trial is to screen out a useless product or concept, then high power is critical even at a higher risk of a false positive result, to avoid abandoning a rare auspicious approach; whereas with a thick pipeline a false negative result would be less damaging. In addition, the Step HIV vaccine trial illustrates that results of a Phase 2b trial in one field can heavily influence other fields, implying that utilities for incorrect decisions should account for the whole context. In particular, a lesson learned from Step is that, for a trial designer that values protecting the vaccine field, for future vaccine efficacy trials it is appropriate to assign a large negative utility to an incorrect decision that the vaccine harms people. While I am not suggesting the Step vaccine did not harm people (I simply do not know– analyses of mechanisms of potential harm are ongoing and are currently inconclusive), this possibility remains plausible, and if true, this error has exacted a high toll on many vaccine development programs. This lesson suggests caution is warranted before selecting quite small Phase 2b HIV vaccine efficacy trials, e.g., with 30–45 events, which run a relatively high risk of a spurious trend of an increased infection rate in the vaccine group. While trial designers may relegate the infection endpoint to secondary status to mitigate this problem, in reality some people will nevertheless over-interpret negative trends, failing to appropriately account for the secondary status and wide confidence intervals. Therefore too-small efficacy trials may run an unacceptably high risk of unintended adverse consequences for both the pathogen-specific field and for other fields.

We have not addressed many remaining issues in Phase 2b product and test-of-concept prevention intervention efficacy trials. For example, it is of interest to conduct comparative expected utility analyses that account for sequential monitoring, for multiple endpoints, and for knowledge about the level of validation of any surrogate endpoints that are used. Accounting for the early stopping guidelines may be of particular importance, as the guidelines may strongly influence the decision probabilities and utilities. For the HIV vaccine field, it is also of particular interest to conduct decision analyses of Phase 2b trials with two primary endpoints (HIV infection and viral load [16]) and with sole primary endpoint viral load [17].

Acknowledgments

The authors thank Drienna Holman, John Hural, Bonnie Kroon, Ben Robinson, and Larry Smith for input on the costs of HIV vaccine efficacy trials, the anonymous reviewers for helpful comments, and Jonas and Susan Ellenberg for organizing the University of Pennsylvania 2008 Workshop. This work was supported by NIH grant 2 R01 AI54165-04.

References

1. Fleming TR, Richardson BA. Some design issues in trials of microbicides for the prevention of HIV infection. The Journal of Infectious Diseases. 2004; 190:666–674.10.1086/422603 [PubMed: 15272392]
2. Moore JP, Ho DD. HIV-1 neutralization: the consequences of viral adaptation to growth on transformed T cells. AIDS. 1995; 9 (Suppl A):S117–S136. [PubMed: 8819579]
3. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF. the rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. The Journal of Infectious Diseases. 2005; 191:654–665.10.1086/428404 [PubMed: 15688278]

4. Pitisuttithum P, Gilbert PB, Gurwith M, Heyward W, Martin M, van Griensven F, Hu D, Tappero JW, Choopanya K. the Bangkok Vaccine Evaluation Group. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. The Journal of Infectious Diseases. 2006; 194:1661–1671.10.1086/508748 [PubMed: 17109337]

5. Anderson RM, Garnett GP. Low-efficacy HIV vaccines: potential for community-based intervention programs. Lancet. 1996; 348:1010–1013.10.1016/S0140-6736(96)07100-0 [PubMed: 8855867]

6. Self, SG. Issues in the design of HIV vaccine efficacy trials. In: Kahn, Patricia; Gust, Ian; Ko3, Wayne, editors. Accelerating AIDS Vaccine Development: Challenges and Opportunities. Horizon Scientific Press, Norfolk; United Kingdom: 2006.

7. Rida W, Fast P, Ho3 R, Fleming TR. Intermediate-size trials for the evaluation of an HIV vaccine candidate: a workshop summary. Journal of the Acquired Immune Deficiency Syndrome and Human Retrovirology. 1997; 16:195–203.

8. Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, Gilbert PB, Lama JR, Marmor M, del Rio C, McElrath MJ, Casimiro DR, Gottesdiener KM, Chodakewitz JA, Corey L, Robertson MN. Step Study Protocol Team. 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; 372:1881–1893.10.1016/S0140-6736(08)61591-3 [PubMed: 19012954]

9. Karim, SA. Safety and effectiveness of vaginal microbicides BufferGel and 0.5% PRO 20005 Gel for the prevention of HIV infection in women: Results of the HPTN 035 trial. Program and Abstracts of the 16th Conference on Retroviruses and Opportunistic Infections; Montreal. 2009; Abstract #48LB

10. Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. John Wiley & Sons; New York: 1991.

11. Berry, DA. Statistics: A Bayesian Perspective. Duxbury Press; Belmont, California: 1996.

12. Girón, FJ. Applied Decision Analysis. Springer; New York: 1998.

13. Petitti, DB. Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine. Oxford University Press; London: 2000.

14. Pascal, B. The Pensées. Krailsheimer, AJ., translator. Penguin Books; London: 1966. 1670

15. Hacking, I. The Emergence of Probability. Cambridge University Press; London: 1975.

16. Mehrotra DV, Li X, Gilbert PB. A comparison of eight methods for the dual-endpoint evaluation of efficacy in a proof-of-concept HIV vaccine trial. Biometrics. 2006; 62:893–900.10.1111/j.1541-0420.2005.00516.x [PubMed: 16984333]

17. Excler J-L, Rida W, Priddy F, Fast P, Ko3 W. A strategy for accelerating the development of preventive AIDS vaccines. AIDS. 2007; 21:2259–2263.10.1097/QAD.0b013e3282eee70c [PubMed: 18090273]

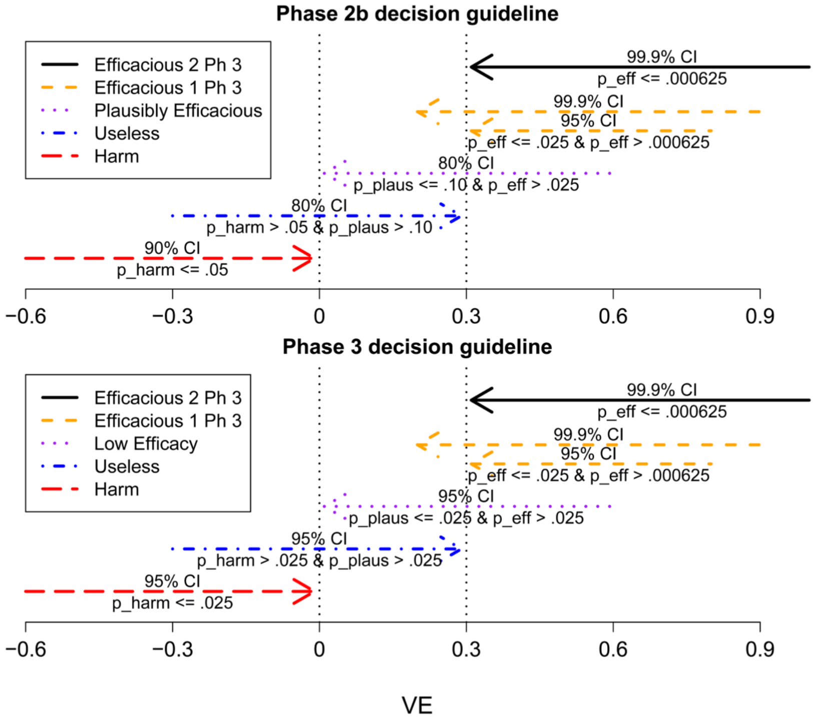

Figure 1.

Confidence Interval Formulation of Phase 2b and 3 Decision Guidelines

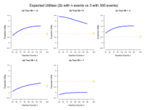

Figure 2.

Comparative Expected Utility for Phase 2b and 3 Efficacy Trials

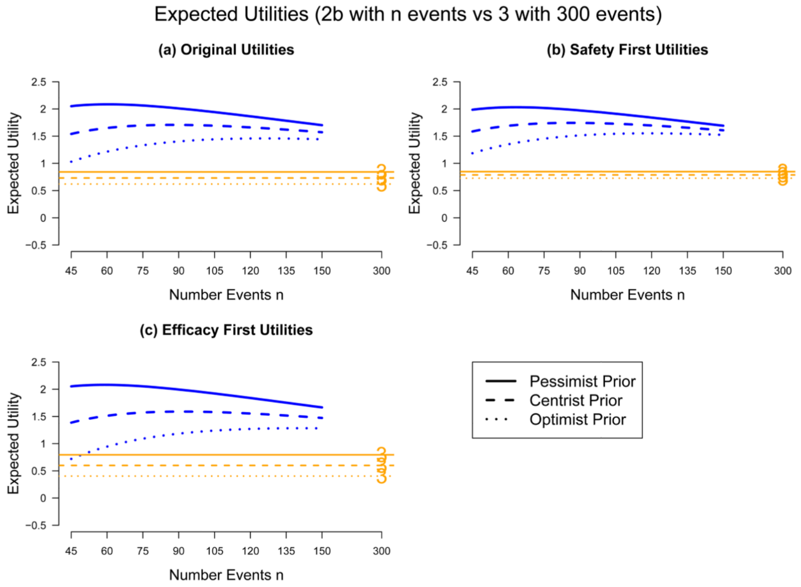

Figure 3.

Sensitivity Analysis: Comparative Expected Utility for Phase 2b and 3 Efficacy Trials

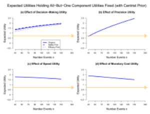

Figure 4.
Sensitivity Analysis: Comparative Expected Utility for Phase 2b and 3 Efficacy Trials, Holding Fixed Three Component Utilities and Varying the Fourth Utility (a) Decision-Making, (b) Precision, (c) Speed, (d) Cost

Table 1

History of HIV vaccine efficacy trials

Efficacy Trial	Phase	Time Period	Vaccine	Study Population	Primary Endpoint(s)	Sample Size	No. Events	Outcome
VaxGen 004 [3]	3	1998–2003	gp120 protein (antibodies)	North America Men sex w/men (MSM)	Infection	5403 2:1 V:P*	**368** 241:127	$\widehat{RR}=.94$ 95% CI .76–1.17 *p* = .59
VaxGen 003 [4]	3	1999–2003	gp120 protein	Thailand IDU	Infection	2527 1:1 V:P	**211** 106:105	$\widehat{RR}=1.0$ 95% CI .76–1.31 *p* = .99
RV 144	2b	2004–2009	ALVAC prime: gp120 boost	Thailand General Pop^n	Infection	16,000 1:1 V:P	**Expect ≈ 90**	Results Due in 2009
Step HVTN 502 [8]	2b	2004–2008	Ad5 vector	Americas MSM + Women	Infection; Viral Load	3000 1:1 V:P	**82**** 49:33	$\widehat{RR}=1.5$ 95% CI .95–2.41 *p* = .07
Phambili HVTN 503	2b	2006–2008	Ad5 vector	South Africa Heterosexual Men + Women	Infection; Viral Load	3000 1:1 V:P	**11**	Trial unblinded after public release of Step results

[*] V:P denotes Vaccine:Placebo.

[**] Of 83 total infections, 82 were in men; the analysis focused on the 82 men.

Table 2

Advancement probabilities for a 90 endpoint product Phase 2b trial*

Move ahead if $p_{plaus} \le$	$\boldsymbol{VE = .3}$	$\boldsymbol{VE = .4}$	$\boldsymbol{VE = .5}$
.025 (traditional)	.38	.66	.90
.05	.51	.77	.94
.10 (2b design)	**.65**	**.87**	**.98**
.15	.74	.91	.99
.20	.80	.94	>.99

*The candidate vaccine is moved ahead to further clinical testing if the 1-sided p-value p_{plaus} for testing $H_0 : VE \le 0$ is less than or equal to the cut-point in the first column.

Table 3

Operating characteristics of a 90 endpoint product Phase 2b trial vs a 300 endpoint Phase 3 trial

(a) Decision guideline for Phase 2b design (*n* = 90 infections)

	Harm	**Useless**	**Plaus Eff**	**Eff 1 Ph 3**	**Eff 2 Ph 3**
			Probability of obtaining an estimated *V E*		
True *V E*	**≤ −.42**	**> −.42 and ≤ .24**	**> .24 and ≤ .55**	**> .55 and ≤ 1.0**	**> 1.0**
−.5	**.60**	.40	<.001	<.001	<.001
0	.05	**.85**	.10	<.001	<.001
.3	<.001	.35	**.63**	.02	<.001
.5	<.001	.02	**.66**	**.32**	<**.001**
.7	<.001	<.001	.04	**.96**	<**.001**

(b) Decision guideline for Phase 3 design (*n* = 300 **infections)**

	Harm	Useless	Plaus Eff	Eff 1 Ph 3	Eff 2 Ph 3
			Probability of obtaining an estimated *V E*		
True *V E*	≤ −.26	> −.26 and ≤ .20	> .20 and ≤ .45	> .45 and ≤ .53	> .53
−.5	**.94**	.06	<.001	<.001	<.001
0	.02	**.95**	.02	<.001	<.001
.3	.<.001	.13	**.84**	.02	<.001
.5	<.001	<.001	**.20**	**.49**	**.31**
.7	<.001	<.001	<.001	<**.001**	**1.0**

[*] Bolded entries are **correct decisions**

Table 4

Utilities u^{dec} for correct or incorrect decisions for different true V E[*]

(a) Baseline utilities

True *V E*	**Harm**	**Useless**	**Plaus Eff**	**Eff 1 Ph 3**	**Eff 2 Ph 3**
−.5	**1**	−2	−3	−4	−5
0	−3	**1**	−1	−2	−3
.3	−3	−2	**1**	−1	−2
.5	−4	−3	−1	**1**	**1**
.7	−5	−4	−2	0	**1**

(b) Alternative utilities u^{dec} “safety first”

True *V E*	Harm	Useless	Plaus Eff	Eff 1 Ph 3	Eff 2 Ph 3
−.5	**1**	−4	−6	−8	−10
0	−1.5	**1**	−1	−2	−3
.3	−1.5	−1	**1**	−1	−2
.5	−2	−1.5	−.5	**1**	**1**
.7	−2.5	−2	−1	0	**1**

(c) Alternative utilities u^{dec} “efficacy first”

True *V E*	Harm	Useless	Plaus Eff	Eff 1 Ph 3	Eff 2 Ph 3
−.5	**1**	−1	−1.5	−2	−2.5
0	−6	**1**	−1	−2	−3
.3	−6	−4	**1**	−1	−2
.5	−8	−6	−2	**1**	**1**
.7	−10	−8	−4	0	**1**

[*] Bolded entries are **correct decisions**