An adaptive design to bridge the gap between Phase 2b/3 microbicide effectiveness trials and evidence required for licensure

Douglas J Taylor, Anneke Grobler and Salim S Abdool Karim

Clin Trials 2012 9: 377 originally published online 18 May 2012
DOI: 10.1177/1740774512445512

CLINICAL TRIALS *ARTICLE* *Clinical Trials* 2012; **9**: 377–384

An adaptive design to bridge the gap between Phase 2b/3 microbicide effectiveness trials and evidence required for licensure

Douglas J Taylor[a], Anneke Grobler[b] and Salim S Abdool Karim[b,c]

Background Vaginally and rectally applied microbicides are being developed to help prevent sexual acquisition of HIV. Due to the lack of surrogate outcomes, the path toward licensure typically moves directly from expanded safety studies to expensive Phase 2b/3 trials with rare incident infection outcomes. The need to confirm an initial trial's significant finding can lead to serious delays in implementing essential programs to reduce the spread of HIV.

Purpose To propose an adaptive design where a Phase 2b/3 study powered to detect a clinically meaningful effect with evidence of one trial (observing one-sided $p < 0.025$) is allowed to expand by a prespecified, feasible amount if interim data suggest the chance of further achieving a more robust evidence threshold ($p < 0.001$, potentially sufficient for licensure from a single trial) is promising.

Methods As an example, prespecified conditional power criteria are used to determine whether a 90-event trial with 90% power to detect a 50% reduction in risk should be expanded to 130 events. Asymptotic results and simulations are used to assess false-positive error rates and other operating characteristics of the design.

Results False-positive error rates can be controlled at the desired 0.025 and 0.001 levels with appropriate choice of critical values or expansion criteria. The chance of achieving robust evidence can approach that of a 130-event trial with traditional stopping boundaries (controlling $\alpha = 0.001$) but with substantially lower expected size for plausible effectiveness levels.

Limitations Conditional power calculations assume the interim estimate of effect is an unbiased estimate for the remainder of the trial, an assumption which may not hold if product adherence varies over time. Observing a measure of effect with $p < 0.001$ may not be sufficient for licensure. A decision to expand the trial would be informative to investigators regarding the interim effect size.

Conclusions A moderate increase in trial size can make the difference between a study with good power to detect a clinically meaningful effect and one which may reasonably obtain the robust evidence required for regulatory bodies and public health programs to consider making a new microbicide available to persons at risk of HIV infection. The proposed design allows for this possibility while not requiring investigators to make an up-front commitment to a prohibitively large trial. *Clinical Trials* 2012; **9**: 377–384. http://ctj.sagepub.com

[a]Family Health International (FHI 360), Durham, NC, USA, [b]Centre for the AIDS Program of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa, [c]Department of Epidemiology, Columbia University, New York, USA
Author for correspondence: Douglas J Taylor, Family Health International (FHI 360), 2224 E NC 54, Durham, NC 27713, USA.
Email: dtaylor@fhi360.org

10.1177/1740774512445512

Introduction

An estimated 33.4 million people were living with HIV in 2008 [1], with the majority of infections acquired through sexual intercourse. Microbicides are vaginally or rectally administered products such as antiretroviral-based gels and rings being developed to protect individuals from infection. A major challenge facing microbicide researchers is the absence of surrogate endpoints to inform effectiveness in modest-sized Phase 2 trials. As a result, product development typically moves from safety studies to large and expensive Phase 2b/3 trials with rare incident infection outcomes. Related challenges include the possibility that participants are exposed to HIV through unprotected routes (e.g., rectal exposure in a vaginal gel study [2,3]) and the dependence of effectiveness on unobserved adherence [4]. As a consequence, determining plausible effectiveness levels to inform the size of Phase 2b/3 trials is extremely challenging.

Investigators typically specify a clinically or programmatically meaningful effect for use in power calculations, such as a 50% reduction in risk. Randomizing participants to test or placebo in equal numbers and following them until 90 infections have been observed provides approximately 90% power to detect such an effect at the one-sided $\alpha = 0.025$ level of significance under standard proportional hazards assumptions. The chance of failing to achieve a statistically significant result will be greater than 10% if the product is less than 50% effective, but this is implicitly acceptable on the grounds that smaller effect sizes are not necessarily meaningful for public health program implementation. If the microbicide were actually 65% effective, then the same trial would have greater than 99% power to achieve a statistically significant result, potentially representing an excessive use of limited resources. Incorporating appropriate interim monitoring (IM) procedures provides protection against this scenario while controlling Type I error at the desired level [5].

Standard practice dictates that a significant finding in a Phase 3 study be confirmed in a second trial before a new drug can be licensed. Fleming and Richardson [6] noted, however, that it could be difficult to justify such replication for preventing an outcome associated with high morbidity or mortality like HIV. Furthermore, the US Food and Drug Administration has expressed their willingness to consider the results of a single, well-conducted trial for licensure of a microbicide if the observed strength of evidence is comparable to that obtained from two independent trials, each with Type I error $\alpha = 0.025$ [7]. This is generally interpreted to mean observing a one-sided p-value less than about 0.001 (recognizing that the chance of making false-positive errors in two traditional Phase 3 trials is $0.025 \times 0.025 = 0.000625$). With this in mind, investigators may consider implementing a large Phase 3 trial with $\alpha = 0.001$ in hopes of accelerating the development process. Returning to the previous example, a study with 170 events would be required to achieve 90% power to detect a 50% reduction in risk with $\alpha = 0.001$. Given that obtaining 170 events would necessitate planning for nearly 8000 person-years of intense follow-up if the incidence rate in the control group is 0.03 (not an atypical assumption), undertaking such an expensive study would be imprudent without preliminary evidence of effectiveness.

Fleming and Richardson [6] proposed a Phase 2b screening trial for this setting, whereby a study with a third to a fourth as many events as required for the large Phase 3 is initially conducted. The results of the screening trial inform whether further testing of the product should be undertaken and if so, how large the subsequent trial should be to achieve the robust strength of evidence required for licensure. In our example, a Phase 2b screening trial would accumulate 45–55 HIV events. This smaller trial still has 90% chance of obtaining robust evidence ($p < 0.001$) if the product is actually 75% effective. On the other hand, the screening trial is underpowered (60%–70%) to detect the meaningful 50% effect with evidence of one trial ($p < 0.025$), and observing a promising yet nonsignificant result could still require that a large Phase 3 trial be conducted. Gilbert [8] contrasted the Phase 2b screening trial design with traditional Phase 3 designs in the context of HIV vaccine trials by comparing their expected utility.

The use of p-values and tests of significance to infer strength of evidence is the subject of longstanding debate [9]. Nonetheless, they remain integral to the processes of drug development and public health policy decision making. With this in mind, we consider another alternative to large Phase 3 trials, one that is also grounded in the use of p-values and significance tests to infer strength of evidence. We describe an adaptive design where a study with high power to detect a prespecified and meaningful effect size with evidence of one trial (observing $p < 0.025$) expands by a prespecified amount if interim data suggest that there is a reasonable chance of achieving a robust evidence threshold ($p < 0.001$; potentially sufficient for licensure). Such a situation could occur if the true effectiveness of the microbicide is modestly greater than the meaningful effect size specified when designing the study.

Using the proposed design, a trial could be declared a success at one of two different strengths of evidence; $p < 0.025$ or $p < 0.001$. This is conceptually similar to studies designed to achieve either a claim of superiority or noninferiority, depending, for example, on whether a confidence interval

(CI) for the hazard ratio excludes 1 or a larger but prespecified noninferiority margin. The approach has parallels to adaptive monitoring and other flexible sample size procedures in which an interim estimate of effect is used to determine final sample size when no clinically meaningful effect is specified in advance [10,11]; even without adjustment to the critical value, expanding a trial when the interim data are sufficiently promising will not inflate the Type I error for a fixed level-α test [12]. Although similar in idea, the proposed design differs from these methods in that the purpose for adapting the trial is to improve the chance of achieving a second evidence threshold, rather than to achieve a target power for a fixed level-α test; the concept of a clinically meaningful effect is maintained; and only one formal test of effectiveness must take place, the timing of which is determined by interim data. The fact that repeat testing is not required with the proposed design may alleviate efficiency concerns raised for some adaptive monitoring procedures [13,14]. Finally, the approach is similar to a Phase 2 trial that rolls directly into a large Phase 3 with $\alpha = 0.001$ [6]. Here, however, there is no consideration to implement such a large study, and achieving evidence of one trial ($p < 0.025$) remains key.

A hypothetical example is used to motivate the proposed trial design. Methods of assessing the operating characteristics of the design (e.g., false-positive error rates, chance of achieving various strengths of evidence, and bias of treatment effect estimators) and a comparison of its performance to that of more traditional designs are explored using asymptotic methods and simulation studies. This is followed by an example where the adaptive design was considered when planning a microbicide trial, and a discussion of limitations and conclusions.

Trial design

The proposed adaptive design requires specifying the numbers of events at which interim and final analyses do, or potentially could, take place, as follows:

N_{IA}: the number of events that triggers an interim assessment of effectiveness;

N_1: the number of events needed to achieve a desired power to detect a meaningful effect with strength of evidence of one trial ($p < 0.025$);

N^*: the largest number of events that the investigator deems prudent and feasible; and

N_2: the number of events needed to achieve a desired power to detect a meaningful effect with strength of evidence of two trials ($p < 0.001$).

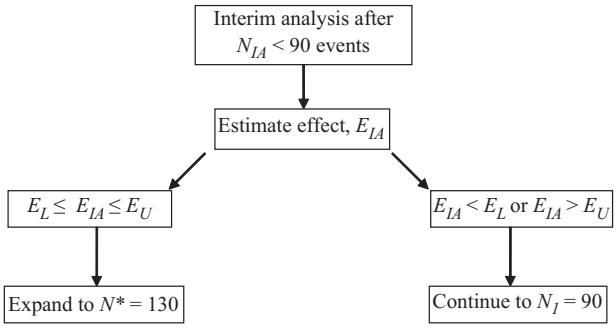

Figure 1. Illustrative adaptive design.

Only scenarios where $N^* < N_2$ are considered in this article. Besides these event counts, the criteria for expanding from N_1 to N^* must be prespecified. This is done recognizing that if the interim estimate of effectiveness (E_{IA}) is too small, then the conditional power to achieve robust evidence will be insufficient to warrant expansion. Likewise, if E_{IA} is large, then the chance of achieving robust evidence will be high without the need for expansion. The range for E_{IA} which triggers expansion is denoted (E_L, E_U). The wider the range, the more likely expansion is to take place and the more the adaptive design will behave like a fixed event trial of size N^*. Thus, the probability of trial expansion is a critical operating characteristic.

To place the design in context, consider the example depicted in Figure 1 where investigators determine that a 50% risk reduction is clinically and programmatically meaningful. Here, $N_1 = 90$ events and $N_2 = 170$ events provide a 90% chance of observing $p < 0.025$ and $p < 0.001$, respectively, but launching into the larger trial is not warranted without preliminary evidence of effectiveness. However, 130 events provide a 90% chance of observing $p < 0.001$ if the microbicide is 55% effective, just 5% greater than the 50% meaningful effect. The investigators determine that $N^* = 130$ is feasible and are open to trial expansion if the interim data are persuasive. Values for N_{IA} and (E_L, E_U) are then prespecified based on false-positive error rate, conditional power, and operational considerations discussed in the following.

Operating characteristics of the proposed adaptive design

False-positive error rates

We assess false-positive error rates by considering the following test statistic for comparing survival curves between two exposure groups

$$Z_j = \frac{\ln\{(HR_j)\}}{\sqrt{4/j}} \quad (1)$$

Here, HR_j is the estimated hazard ratio (test vs. control) obtained from a Cox's proportional hazards regression model after j events have been observed. Asymptotically, Z_j is distributed normal with mean zero and variance one under the null hypothesis of no treatment effect if standard assumptions (e.g., no informative censoring) are met. Normality does not hold, however, when the final event size is chosen conditional on an interim estimate of effectiveness. In particular, the proposed design dictates that the trial expand from $N = N_1$ to $N = N^*$ events if ($E_L \le E_{IA} \le E_U$), where $E_{IA} = 1 - HR_{IA}$. Large sample false-positive error rates in this scenario equal

$$\begin{aligned}\Pr\{Z_N < z_p \mid HR = 1\} = &\int_{-\infty}^{t_u} \Pr\{Z_{N_1} < z_p \mid Z_{N_{IA}}, HR = 1\}\phi(Z_{N_{IA}})dZ_{N_{IA}} + \\ &\int_{t_u}^{t_l} \Pr\{Z_{N^*} < z_p \mid Z_{N_{IA}}, HR = 1\}\phi(Z_{N_{IA}})dZ_{N_{IA}} + \\ &\int_{t_l}^{\infty} \Pr\{Z_{N_1} < z_p \mid Z_{N_{IA}}, HR = 1\}\phi(Z_{N_{IA}})dZ_{N_{IA}}\end{aligned} \quad (2)$$

where $\phi(\cdot)$ is the standard normal density, $t_i = \ln(1 - E_i)\sqrt{N_{IA}/4}; i \in (L, U)$, and $\Pr\{Z_j < z_p \mid Z_{N_{IA}}\}$ is the conditional probability that the final test statistic falls below the pth percentile of a standard normal variable, z_p (see Appendix 1 for an asymptotic expression of this probability).

It is apparent from Equation (2) that choosing $z_p = z_\alpha$ will not in general provide a size α test; false-positive error rates can be inflated or deflated, depending on the choice of N_{IA} and (E_L, E_U). Since Equation (2) is an increasing function of z_p, however, an asymptotically unbiased test can be obtained by solving for p such that the false-positive error rate equals α. In order to understand the potential magnitude of bias when letting $z_p = z_\alpha$, false-positive error rates for the motivating example in Figure 1 were computed for nominal $\alpha \in \{0.025, 0.001\}$, all $N_{IA} < N_1$ and $(E_L, E_U) \in \{(0.01, 0.02), (0.01, 0.03), \ldots, (0.01, 0.99), \ldots, (0.98, 0.99)\}$. Equation (2) was solved using the QUAD function in SAS/IML® [15] to perform the required numerical integration or a simple trapezoidal rule if the function failed to converge.

Some of the evaluated N_{IA} and (E_L, E_U) make no practical sense (e.g., $N_{IA} = 1$) and others would represent obvious attempts to inflate the false-positive error rate. Nonetheless, the bias was generally modest over the comprehensive range of assessed conditions, with realized error rates between 0.014 and 0.036 for a nominal $\alpha = 0.025$ level test and between 0.0004 and 0.0016 for a nominal $\alpha = 0.001$ level test. Maximum and minimum error rates occur when $N_{IA} = 89$ and information regarding the likely outcome after $N_1 = 90$ events is greatest. As extreme examples, choosing (E_L, E_U) such that expansion must occur if the conditional probability of rejection is essentially zero without expansion will inflate the error rates. Likewise, choosing (E_L, E_U) such that expansion must occur if the chance of rejection is already 100% will deflate the error rates. As exemplified below, however, the absolute bias may be negligible for reasonable choice of expansion criteria (akin to the fact that the false-positive error rate for a fixed level-α test need not be inflated when sample size is increased based on interim data [12]).

Power considerations

Building on the asymptotic results in the previous section, simulation studies were used to assess the chance of achieving the desired strengths of evidence under null and alternative hypotheses in finite samples. There are infinitely many combinations of N_{IA} and (E_L, E_U) which could be considered for the example in Figure 1, and any attempt to optimize their selection needs to balance power and the chance of trial expansion for a range of plausible effect sizes against the willingness of investigators to invest in a larger trial under those scenarios. Recognizing the inherent subjectivity of such an exercise, data were simulated under the assumption that the interim analysis takes place after either $N_{IA} = 60$ or 89 events; $N_{IA} = 60$ representing a practical choice in that investigators might want appreciable time between the interim analysis and the expected date of the 90th event to plan for expansion, and $N_{IA} = 89$ representing the maximum possible information to inform the decision to expand. Next, values of (E_L, E_U) were chosen such that the trial would expand from $N_1 = 90$ events to $N^* = 130$ events if the conditional power to achieve robust evidence ($p < 0.001$) under expansion is at least 50% and the conditional power in the absence of expansion is less than 95%. Using formulae presented in Appendix 1, and assuming the true hazard ratio equals HR_{IA}, we obtain (E_L, E_U) = (0.418, 0.573) when $N_{IA} = 60$ and (E_L, E_U) = (0.418, 0.497) when $N_{IA} = 89$. Although not optimized in any formal sense, these criteria for expansion are consistent with the motivation for the proposed design; expand if there is a reasonable chance of obtaining robust evidence of effect, unless the chance is already very high without expansion.

Simulated data were generated under an exponential failure time distribution for a range of true hazard ratios between 0.3 and 0.9 (i.e., between 70% and 10% effectiveness), as well as the null case. For each simulated condition, 10,000 replicates of data (50,000 for the null case) were generated and

analyzed using proportional hazards regression in SAS®, assuming 90% censoring at 1 year in the control group. Probabilities of achieving $p < 0.025$ and $p < 0.001$, respectively, were estimated based on the proportion of replicates for which upper one-sided 97.5% and 99.9% confidence bounds for the hazard ratio excluded 1.0 (nominal coverage rates were used since numerical evaluation of equation (2) determined that false-positive error rates are well controlled in these scenarios: 0.0220 and 0.00084, respectively, for $\alpha = 0.025$ and $\alpha = 0.001$ level tests when $N_{IA} = 60$; 0.0237 and 0.0010 when $N_{IA} = 89$). The costs and benefits associated with the design are summarized in terms of the probability that trial expansion occurs as a function of effectiveness, as well as the chance of observing $p < 0.001$ given that expansion takes place.

In order to compare the operating characteristics of the adaptive trial to more traditional designs, false-positive error rates and power of fixed N = 90-event and N = 130-event trials were computed. Likewise, power was computed for a trial with IM based on the Lan–DeMets spending function [16] with O'Brien–Fleming-type stopping boundaries [17] (computed in East® [18]) to control the false-positive error rate at $\alpha = 0.001$, assuming interim analyses occur after 60 and 90 events and maximum trial information of $N_{max} = 130$.

Bias of estimated treatment effects

An additional concern when implementing a trial which allows for early stopping is the potential bias of estimated treatment effects, since this can impact the interpretation of study results. Unconditional bias for the adaptive design in Figure 1 (with N_{IA} = 60, $N_1 = 90$, and $N^* = 130$) was estimated based on the mean of the simulated HR estimates, regardless of whether or not expansion took place. For comparison, bias was also summarized for a trial that incorporates O'Brien–Fleming-type stopping boundaries with $N_{max} = 130$ and interim analyses after 60 and 90 events.

Simulation results

False-positive error rates and power

As expected, the estimated false-positive error rates are near, but do not exceed, the nominal 0.025 and 0.001 levels for the example adaptive design (Table 1), emphasizing that test size may be well controlled when using unadjusted critical values. The chance that a traditional trial using O'Brien–Fleming-type stopping boundaries achieves $p < 0.025$ is essentially identical to that of a fixed N = 130-event trial for all evaluated effect sizes (results not shown). In contrast, power of the adaptive design to observe $p < 0.025$ is essentially the same as a fixed 90-event trial, regardless of whether the interim analysis takes place after N_{IA} = 60 or 89 events (Figure 2). Given that the purpose of the adaptive design is to increase the chance of achieving robust evidence of effect, however, the chance of observing $p < 0.001$ is of greater interest.

For a trial using O'Brien–Fleming-type boundaries, power to achieve $p < 0.001$ is again very close to that of a fixed N = 130-event trial (results not shown). Power of the adaptive design was always within 12% and 9% of a fixed 130-event trial when $N_{IA} = 60$ and $N_{IA} = 89$, respectively (Figure 3). For example, for a 55% effective product, the estimated chance of observing $p < 0.001$ was 81% when $N_{IA} = 60$ and 84%

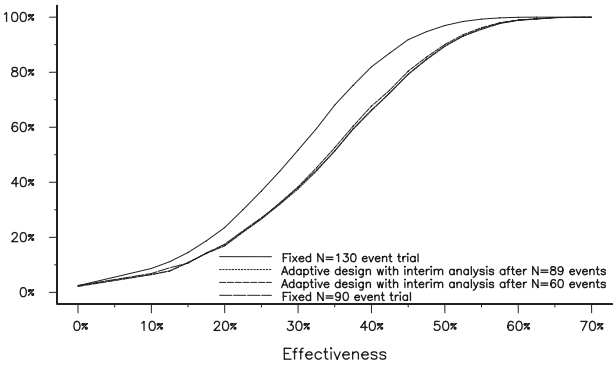

Figure 2. Simulated chance of observing $p < 0.025$ for fixed event and adaptive design trials.

Table 1. Simulated false-positive error rates for adaptive[a], fixed event total, and traditional IM[b] design trials

Evidence level	Trial design		False-positive error (95% CI)
$p < 0.025$	Adaptive	$N_{IA} = 60$	0.0213 (0.0200, 0.0226)
		$N_{IA} = 89$	0.0233 (0.0220, 0.0246)
	Fixed event	$N = 90$	0.0248 (0.0235, 0.0262)
		$N = 130$	0.0242 (0.0229, 0.0256)
	Traditional IM	$N_{max} = 130$	0.0242 (0.0229, 0.0256)
$p < 0.001$	Adaptive	$N_{IA} = 60$	0.0008 (0.0006, 0.0011)
		$N_{IA} = 89$	0.0010 (0.0007, 0.0012)
	Fixed event	$N = 90$	0.0010 (0.0008, 0.0013)
		$N = 130$	0.0011 (0.0008, 0.0014)
	Traditional IM	$N_{max} = 130$	0.0011 (0.0008, 0.0014)

CI: confidence interval; IM: interim monitoring.
[a]Interim analysis takes place after N_{IA} events. Trial either stops after 90 events or is expanded to 130 events if (1) conditional power to achieve $p < 0.001$ is at least 50% with expansion and (2) conditional power is no more than 95% without expansion.
[b]IM using Lan–DeMets spending function with O'Brien–Fleming-type boundaries to control false-positive error rate at the one-sided 0.001 level, with interim analyses after 60 and 90 events and maximum number of events $N_{max} = 130$.

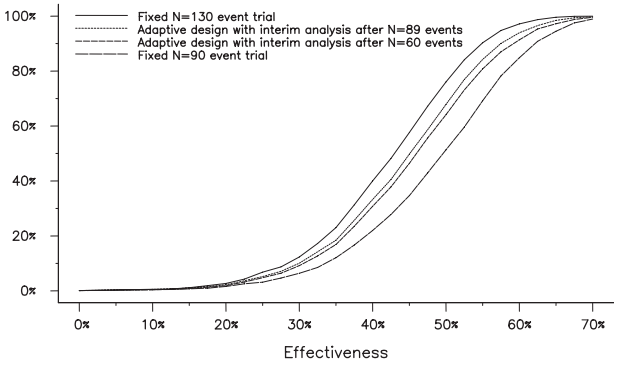

Figure 3. Simulated chance of observing $p < 0.001$ for fixed event and adaptive design trials.

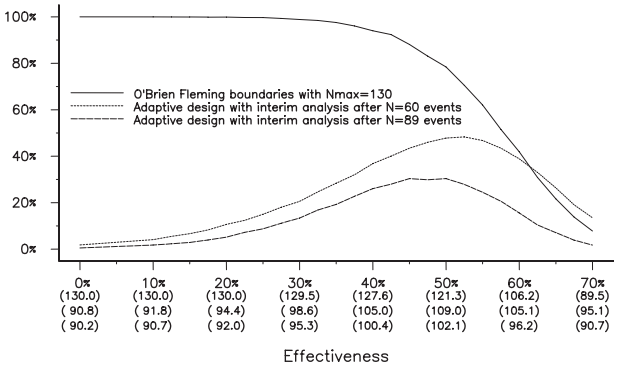

Figure 4. Simulated chance that traditional interim monitoring and adaptive design trials continue to 130 events (values in parentheses are expected event totals for traditional interim monitoring trial, adaptive design with $N_{IA} = 60$ and adaptive design with $N_{IA} = 89$, respectively).

when $N_{IA} = 89$, compared to 90% for a fixed $N = 130$-event trial. The chance that the adaptive trial expands from $N_1 = 90$ to $N^* = 130$ is always less than 50% when $N_{IA} = 60$ and always less than 30% when $N_{IA} = 89$ (Figure 4). In contrast, the chance of continuing to $N_{max} = 130$ for a trial using O'Brien–Fleming-type boundaries increases as effectiveness decreases and is essentially 1.0 for all effectiveness levels below 35% when not incorporating stopping bounds for futility. Although the latter design is more powerful, the expected trial size is larger for all evaluated effectiveness levels below 70% when $N_{IA} = 89$ and larger for all effectiveness levels below 60% when $N_{IA} = 60$.

By comparing Figure 3 with Figure 4, it is apparent that the chance the adaptive trial expands can be nontrivial in scenarios where the overall probability of observing $p < 0.001$ is quite low. For example, there is nearly a 30% chance that the adaptive trial expands when true effectiveness is 35%, even though the overall chance of observing $p < 0.001$ is less than

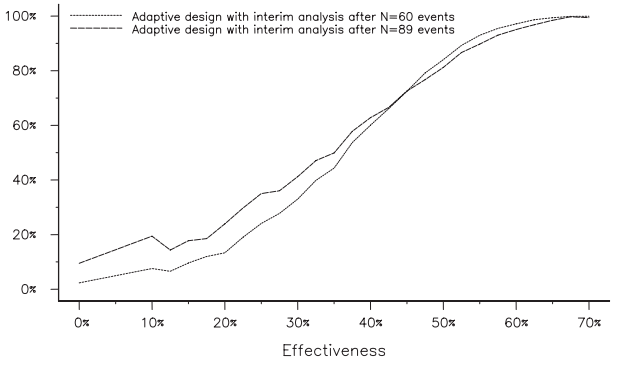

Figure 5. Simulated chance of observing $p < 0.001$ given that adaptive design trial expands from $N = 90$ events to $N = 130$ events (owing to the small chance of expansion, fewer than 500 replicates contribute to analysis for effectiveness levels below 10%).

20%. Given that expansion occurs, however, the chance of observing $p < 0.001$ is about 50% or greater so long as true effectiveness is at least 35% (Figure 5).

Bias of estimated treatment effects

The estimated bias of the *HR* estimator was modest for the proposed adaptive design, ranging from 0.025 for true hazard ratios near unity to 0.000 for a true hazard ratio of 0.30. Estimated bias for the traditional trial with O'Brien–Fleming-type boundaries was also small, ranging from 0.015 for true hazard ratios near unity to –0.005 for a true hazard ratio of 0.30 (results not shown).

Example

The CAPRISA 004 study, conducted among 889 HIV-negative women in KwaZulu-Natal, South Africa, was the first microbicide trial to demonstrate a significant protective effect [19]; randomization to a regimen of 1% tenofovir gel was associated with an estimated 39% reduction in HIV risk compared to an identical placebo regimen (one-sided $p = 0.009$; 95% CI: 6%–60%).

Initial design considerations revolved around a 66-event trial with 80% power to detect a 50% reduction in risk ($\alpha = 0.025$), incorporating interim effectiveness analyses after 22 and 44 events. During the planning stages, however, the principal investigator raised the question which motivated the adaptive design considered here: could a trial which looked promising near its planned end be expanded in hopes of further achieving strength of evidence of two trials. As part of these discussions, the investigator concluded that a trial much larger than 88 events was not feasible. This led the statisticians to consider the

operating characteristics of an adaptive design with $N_{IA} = 44$, $N_1 = 66$, $N^* = 88$, and $(E_L, E_U) = (40\%, 70\%)$.

Based on these assumptions, the adaptive trial could achieve power near that of a fixed $N = 88$-event trial while controlling false-positive error rates near 0.025 and 0.001 (0.021 and 0.0009, respectively, using unadjusted critical values). For example, if the microbicide were truly 60% effective, then the adaptive design has a simulated 80% chance of observing $p < 0.001$, compared to 83% for a fixed 88-event trial. The chance that the trial would expand exceeds 70% for many plausible effect sizes, however, diminishing the benefits of the adaptive trial compared to an $N = 88$ event study with standard stopping rules. Ultimately, a 92-event trial with 90% power to detect a 50% reduction in risk with evidence of one trial was implemented.

In hindsight, $(E_L, E_U) = (40\%, 70\%)$ was too wide an interval, requiring expansion when the conditional chance of achieving robust evidence was as little as 16%. Nonetheless, the CAPRISA 004 data can be analyzed as if the adaptive design had been employed. After 44 events, the interim estimate of effectiveness was 29.8%, which would not have triggered trial expansion. After 66 events, the interim estimate of effectiveness was 46.2%, with one-sided $p = 0.007$. Although highly speculative, this suggests that a comparable strength of evidence (significant but not sufficient for licensure) might have been achieved in a smaller study had the adaptive trial been implemented.

Discussion

An adaptive design was proposed in which a microbicide study with 90% power to detect a 50% reduction in risk of HIV with evidence of one trial (observing one-sided $p < 0.025$) is expanded to a moderately larger number of events, N^*, if the interim conditional power to achieve a more robust evidence threshold ($p < 0.001$, potentially sufficient for licensure) appears promising. Such may be the case if the microbicide's actual effectiveness is modestly higher than 50%.

Asymptotic and simulation results demonstrate that, through appropriate choice of expansion criteria, the chance that the adaptive design achieves robust evidence of effect can approach that of a fixed $\alpha = 0.001$ level trial of size N^*, but with the potential for much smaller expected study size and while controlling false-positive error rates at the desired 0.025 and 0.001 levels. Group sequential trials using traditional stopping boundaries to control the Type I error at $\alpha = 0.001$ should provide greater power but at a cost of potentially much larger study sizes, the avoidance of which was the motivation to consider alternative designs.

Numerous related designs could be proposed. For example, a trial of size N^* could be implemented which incorporates both interim effectiveness boundaries (controlling $\alpha = 0.001$) and two-stage futility boundaries: one stage would trigger a reduction in size to $N_1 < N^*$ events if achieving robust evidence appears unlikely and a second stage would allow the trial to stop entirely if there was little hope of achieving even evidence of one trial ($p < 0.025$). Prespecified effectiveness boundaries which control the applicable false-positive error rates could likely be developed using the types of asymptotic and simulation results presented here.

There are a number of important challenges and limitations to consider. First, the benefits of the proposed design in terms of expected study size become more prominent as the period between when the interim analysis which triggers expansion occurs and when the trial would otherwise stop gets shorter, whereas implementing the trial would get operationally more difficult as this period gets shorter. Second, control of false-positive error rates may not be maintained using unadjusted tests for studies with different design parameters than those considered here. Hence, it is essential to assess error rates using simulations or asymptotic results and adjust critical values using the formula provided as necessary, before implementing such a trial. Third, anyone with knowledge of the criteria for expansion would gain substantial insight into the effectiveness of the product if the trial expands; withholding the cumulative number of events from all but the Data and Safety Monitoring Board (DSMB) and a select few individuals could help avoid unintended consequences of such knowledge. Fourth, the rules for trial expansion are binding once the interim data have been assessed. Since the DSMB might choose to stop the trial at the interim review due to overwhelming evidence of effectiveness, incorporating a highly conservative Haybittle–Peto-type boundary at the interim analysis may be recommended. Fifth, conditional power considerations presume that effectiveness is constant over time. For microbicides, effectiveness is driven by participant adherence to the regimen, and a temporal shift in adherence could undermine this assumption. In CAPRISA 004 [19], for example, there was evidence to suggest that effectiveness diminished after 18 months of product use, possibly due to participant fatigue and reduced adherence to the treatment regimen. Sixth, since the trial can stop after either N_1 or N^* events, the final p-value cannot be interpreted as the probability of obtaining a more extreme result than the one observed. Rather, the unadjusted p-value is only used to determine whether a particular evidence threshold has been reached. Similar limitations hold for CIs; a naïve interval computed on the final study data may not have the desired coverage. Developing methods of estimating p-values and CIs which maintain the desired frequentist interpretations is an area of future research.

Finally, the assumption that observing $p < 0.001$ would lead regulators to license a product without a confirmatory trial may not be justifiable in settings that do not involve outcomes associated with high morbidity or mortality like HIV. Even here, regulators or public health programs may require additional studies to gather sufficient safety data.

Despite the challenges, adaptive designs like the one proposed should be considered when there is little preliminary evidence of effectiveness to inform the size of a Phase 3 trial. Such evidence is generally not available to microbicide researchers owing to the absence of surrogate outcomes to inform product effectiveness in modest-sized Phase 2 trials.

Funding

This work was supported in part by the United States Agency for International Development (USAID) through Cooperative Agreement GPO-A-OO-05-00022-00

Conflict of interest

The authors report no conflicts of interest.

References

1. **World Health Organization, UNAIDS, UNICEF.** Towards universal access: scaling up priority HIV/AIDS interventions in the health sector. Progress Report 2010. Available at: http://www.who.int/hiv/pub/2010progressreport/full_report_en.pdf (accessed 8 October 2010).
2. **McGowan I, Taylor D.** Heterosexual anal intercourse has the potential to cause a significant loss of power in vaginal microbicide effectiveness studies. *Sex Transm Dis* 2010; **37**: 361–64.
3. **Mâsse BR, Boily M-C, Dimitrov D, Desai K.** Efficacy dilution in randomized placebo-controlled vaginal microbicide trials. *Emerg Themes Epidemiol* 2008; **6**: 5.
4. **Weiss HA, Wasserheit JN, Barnabas RV, Hayes RA, Abu-Raddad LJ.** Persisting with prevention: The importance of adherence for HIV prevention. *Emerg Themes Epidemiol* 2008; **5**: 8.
5. **Jennison C, Turnbull BW.** *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, Boca Raton, FL, 2000.
6. **Fleming TA, Richardson BA.** Some design issues in trials of microbicides for the prevention of HIV infection. *J Infect Dis* 2004; **190**: 666–74.
7. In *Transcript of the Food and Drug Administration (FDA) Antiviral Drugs Advisory Committee Meeting*. Bethesda, MD, 20 August 2003. http://www.fda.gov/ohrms/dockets/ac/03/transcripts/3970T1.htm (accessed 25 April 2012)
8. **Gilbert P.** Some design issues in phase 2B vs phase 3 prevention trials for testing efficacy of products or concepts. *Stat Med* 2010; **29**(10): 1061–71.
9. **Goodman SN.** Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med* 1999; **130**: 995–1004.
10. **Proschan MA, Hunsberger SA.** Designed extension of studies based on conditional power. *Biometrics* 1995; **51**: 1315–24.
11. **Mehta CR, Patel NR.** Adaptive, group sequential and decision theoretic approaches to sample size determination. *Stat Med* 2006; **25**: 3250–69.
12. **Chen YHJ, DeMets DL, Lan KKG.** Increasing the sample size when the unblinded interim result is promising. *Stat Med* 2004; **23**: 1023–38.
13. **Jennison C, Turnbull BW.** Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat Med* 2003; **22**: 971–93.
14. **Fleming TR.** Standard versus adaptive monitoring procedures: A commentary. *Stat Med* 2006; **25**: 3305–12.
15. **SAS Institute Inc.** *SAS/STAT® 9.2 User's Guide*. SAS Institute Inc., Cary, NC, 2008.
16. **Lan KKG, DeMets DL.** Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–63.
17. **O'Brien PC, Fleming TR.** A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–56.
18. **Cytel Statistical Software and Services.** *East: Software for Advanced Clinical Trial Design, Simulation, and Monitoring, Version 5.2*. Cytel Statistical Software and Services, Cambridge, MA, 2008.
19. **Abdool Karim Q, Abdool Karim SS, Frohlich JA, et al.** Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science* 2010; **329**: 1168–74.

Appendix 1

The conditional power to achieve $p < \alpha$ if the trial is expanded to $N^* > N_1$ events was computed as

$$1 - \Phi\left\{\frac{\left((\ln(HR_{IA})/SE)\sqrt{N_{IA}/4} - z_\alpha\sqrt{N^*/4} + ((N^* - N_{IA})/4)\ln(HR_{alt})\right)}{\sqrt{(N^* - N_{IA})/4}}\right\} \quad (3)$$

where SE is the standard error of the logged interim HR estimate (HR_{IA}) from a proportional hazards regression model, N_{IA} is the number of events at the time of the interim analysis, N^* is the number of events if the trial is allowed to expand, and HR_{alt} is the hazard ratio under the alternative hypothesis ($HR_{alt} = HR_{IA}$ in power simulations and $HR_{alt} = 1$ when assessing false-positive error rates). Likewise, the conditional power to achieve evidence of effect if the trial is not expanded is computed as

$$1 - \Phi\left\{\frac{\left((\ln(HR_{IA})/SE)\sqrt{N_{IA}/4} - z_\alpha\sqrt{N_1/4} + ((N_1 - N_{IA})/4)\ln(HR_{alt})\right)}{\sqrt{(N_1 - N_{IA})/4}}\right\} \quad (4)$$

The lower and upper bounds on the interim estimate of effectiveness which trigger expansion of the study (E_L, E_U) may be computed by iteratively solving Equations (3) and (4) for any chosen conditional power values, assuming $SE = \sqrt{4/N_{IA}}$ and effectiveness = $1 - HR_{IA}$.