AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 29, Number 4, 2013

DOI: 10.1089/aid.2012.0180

The Design of Single-Arm Clinical Trials of Combination Antiretroviral Regimens for Treatment-Naive HIV-Infected Patients

Lu Zheng,[1] Susan L. Rosenkranz,[1] Babafemi Taiwo,[2] Michael F. Para,[3] Joseph J. Eron, Jr.,[4] and Michael D. Hughes[1]

Abstract

Single-arm clinical trials are useful to evaluate antiretroviral regimens in certain populations of HIV-infected treatment-naive patients for whom a randomized controlled trial is not feasible or desirable. They can also be useful to establish initial estimates of efficacy and safety/tolerability of novel regimens to inform the design of large phase III trials. In this article, we discuss key design considerations for such single-arm studies.

Introduction

The randomized controlled trial (RCT) is the gold standard for evaluating clinical interventions, including antiretroviral (ARV) therapy in treatment-naive HIV-infected patients. There are, however, situations in which single-arm clinical trials may be valuable. One such example concerns the evaluation of a treatment in populations that may be excluded from large phase III RCTs yet are not common enough to populate a well-powered RCT specifically targeting that population. This scenario is illustrated by HIV-infected patients with transmitted resistance to one or more ARVs who are commonly excluded from phase III RCTs in ARV treatment-naive patients. Some patients may also be inappropriate for inclusion in an RCT due to preexisting conditions that may compromise safety such as chronic renal failure, or skew outcomes such as substance abuse. For many of these populations, the enrollment of the large number of patients necessary for a separate RCT that includes regimens the patients are able to take may not be viable, making a single-arm trial a potential alternative.

Single-arm trials are also useful in assessing the efficacy and safety/tolerability of novel regimens comprising one or more already approved ARVs, as an initial assessment before proceeding to larger scale evaluation in an RCT. When the standard of care is already well defined from previous RCTs, the established efficacy allows rational estimation of acceptable efficacy/safety of the experimental intervention. In this article, we discuss issues that are central to the design of single-arm clinical trials in treatment-naive HIV-infected patients and illustrate them with the design of the AIDS Clinical Trials Group (ACTG) study A5262.

Design Issues

The primary objective of a single-arm study of an ARV regimen must be achievable without a concurrent comparator arm, and parameters guiding interpretation of study findings including threshold for success or failure must be prespecified and well understood. Figure 1 illustrates prespecified guidelines for interpreting a hypothetical study. The primary objective of the study is to estimate regimen efficacy in a target population as measured by the proportion of patients failing to achieve and maintain virologic suppression below a defined assay threshold. Given the lack of a concurrent comparator arm, the observed failure rate and the associated two-sided confidence interval (CI) are compared with a prespecified maximally acceptable failure rate (threshold; we discuss the choice of this threshold below). If the CI for the failure rate is entirely below the threshold (scenario A in Fig. 1), then the regimen under evaluation is considered acceptable and so may be recommended for use in the target population or, in a drug development program, for further evaluation in a larger comparative RCT. Conversely, if the CI is entirely above the threshold (scenario B), then regimen is considered unacceptable and so not recommended for use or further evaluation. If the CI includes the threshold of interest (scenario C), the study may be considered inconclusive, and

[1]Statistical Data Analysis Center, Harvard School of Public Health, Boston, Massachusetts.
[2]Division of Infectious Diseases, Northwestern University, Chicago, Illinois.
[3]Division of Infectious Diseases, Ohio State University, Columbus, Ohio.
[4]Division of Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

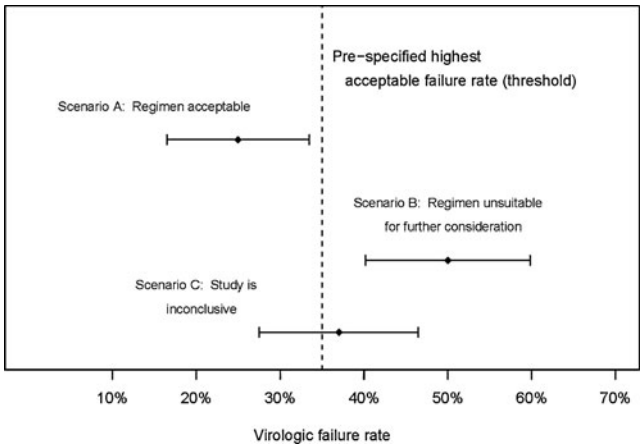

FIG. 1. Representative hypothetical confidence intervals achieved at the end of the single-arm study and how each is interpreted.

consideration for use may depend on factors such as tolerability or availability of treatment options in the target population.

In some situations, the definition of acceptable outcome may need to be more stringent with scenario C also considered unacceptable. In this case, only the upper bound of the CI is relevant to the decision-making process and so a one-sided CI might be used. However, there is a general consensus, including from regulatory agencies,[1–3] favoring the use of two-sided CIs. Two-sided CIs also facilitate interim monitoring: if the observed failure rate in the study is higher than might be considered acceptable, then the lower bound of the CI gives a sense of plausible true failure rates, information critical to decisions that may be made by study monitoring committees, such as whether to terminate a study early. In general, a standard two-sided 95% CI is recommended though a 90% CI might be appropriate particularly if further study of the regimen in the population of interest is expected (e.g., if the study is to be followed by a larger RCT). To maintain the level of evidence required before considering a regimen acceptable, when a one-sided CI is used then the confidence bound should be the same as the upper bound of the two-sided interval. For example, if 95% would have been chosen for a two-sided CI, then 97.5% should be used for the one-sided confidence bound.

Example of a Single-Arm Clinical Trial

ACTG study A5262 was a prospective, multicenter, 52-week single-arm study designed to evaluate the safety and efficacy of raltegravir (RAL) plus ritonavir-boosted darunavir (DRV/r) as initial combination ARV therapy in HIV-1-infected ARV-naive participants with or without transmitted nonnucleoside reverse transcriptase inhibitor (NNRTI) or nucleos(t)ide reverse transcriptase inhibitor (NRTI) drug-resistance mutations.[4] Participants with plasma HIV-1 RNA levels $\geq$5,000 copies/ml obtained within 90 days prior to study entry were eligible. Participants were ineligible if the screening HIV RT/protease genotype or any previous RT/protease genotype showed more than one DRV resistance-associated mutation (RAM) or L76V alone[5] or if the participants had one or more known major integrase inhibitor resistance-associated mutations.[6] There was no $CD4^+$ cell count restriction and patients with renal insufficiency not requiring dialysis were allowed.

Plasma HIV-1 RNA was evaluated at study entry, and at weeks 1, 4, 12, 24, 36, and 48 after starting the study drug; HIV-1 RNA was also evaluated at week 52 if virologic failure was suspected at week 48. The primary objective was to estimate the cumulative proportion of ARV-naive participants experiencing virologic failure at or before week 24 after initiating RAL plus DRV/r. Virologic failure was defined as follows:

Week 12:

- Confirmed plasma HIV-1 RNA $\geq$1,000 copies/ml or
- Confirmed rebound from the week 4 value by 0.5 $\log_{10}$ copies/ml (for participants with week 4 value $\leq$50 copies/ml, confirmed rebound to >50 copies/ml).

Week 24:

- Confirmed value >50 copies/ml.

The study aimed to enroll 111 participants. Primary analysis used an intent-to-treat (ITT) approach, wherein any virologic failure was included in the analysis regardless of whether the participant was on or off study treatment at the time of measurement, but follow-up was censored if a participant was lost to follow-up without previously meeting the definition of virologic failure. The intent of the study was to follow all participants for 52 weeks even if they discontinued the study regimen.

The cumulative proportion of participants experiencing virologic failure at or prior to week 24 was estimated using the method of Kaplan and Meier with the associated two-sided 95% CI.[7] It was prespecified that the regimen would be considered acceptable for further investigation if the 95% CI around the proportion of participants experiencing virologic failure was below 35%, and unacceptable if the 95% CI was above or included 35%.

Choice of Outcome Measure

The interpretation of results from a single-arm trial is facilitated by use of a primary outcome measure that is identical in definition, or nearly so, to that used in the trials to which the current trial's results will be compared. For example, for A5262 a purely virologic endpoint was chosen to match comparator ACTG studies A5095 and A5142, from which week 24 failure rate estimates were available.

The duration of time over which the primary outcome measure is evaluated also needs to be defined. For studies in ARV-naive participants, the majority of virologic failures and treatment discontinuations tend to occur early, often within 24 weeks of ARV initiation. Illustratively, among 765 participants randomized to receive one of the two efavirenz-containing regimens in ACTG A5095, 115 (15%) exhibited virologic failure by week 48.[8] Of these 115 failures, 76% failed by week 24. Thus, the week 24 measure is likely to be adequate for an initial assessment of the study regimen, whereas 2 to 3 years of follow-up would be needed to evaluate long-term virologic outcome (determined largely by virologic rebound following suppression) as distinct from early outcome (determined by lack of adequate initial suppression as well as early rebound and treatment limiting toxicity).

As well as providing an earlier assessment of a regimen, a relatively early endpoint can also provide pilot efficacy and

safety data to larger RCTs evaluating the same regimen. For example, while A5262 was ongoing, the European AIDS Treatment Network designed a large phase III RCT (NEAT 001, NCT01066962) comparing RAL + DRV/r to DRV/r+ tenofovir/emtricitabine (TDF/FTC) in 800 treatment-naive HIV-1-infected patients followed for 96 weeks. The A5262 study team was able to share primary (week 24) results with the NEAT 001 team before A5262 follow-up of 52 weeks was completed, alerting the NEAT 001 team of the potentially lower than expected efficacy of the RAL + DRV/r arm in a particular patient subgroup. Of note, a later endpoint may be desirable if the single-arm study is likely to be the only trial of the regimen in the target population, thus necessitating a more comprehensive assessment.

Selection of Threshold for Unacceptable Virologic Failure Rate

It is important to prespecify a threshold failure rate below which the chosen primary outcome measure would be considered acceptable. Ideally, the threshold would be based on results from recent trials, recognizing, however, that the populations enrolled in these trials might be different from the target population of the trial being planned. A5262 study investigators obtained the following week 24 failure rates for the efavirenz and two NRTI arms of A5095 and A5142 from the study teams: 21% (95% CI 17% to 25%[8]) and 22% (17% to 28%[9]), respectively. Dose-specific estimates ranging from 5% to 15% (with upper bounds of the 95% CIs ranging from 17% to 30%) were available for the raltegravir-based regimens of Merck 004.[10]

The design of RCTs comparing ARV regimens in treatment-naive populations provides guidance for selecting acceptable outcome thresholds for single-arm studies. In RCTs, a noninferiority bound of 10% is often used for the difference in failure rates between arms. Thus, a reasonable approach in a single-arm study in treatment-naive participants might be to use a failure threshold that is 10% higher than the typical failure rate observed in the superior (or noninferior) arms of the larger recent randomized studies. In the Merck and the two ACTG trials above, observed failures rates of 22% or lower, and CI upper bounds of 30% or lower, suggest that true failures rates might reasonably be less than 30%. Hence the A5262 team considered that a true failure rate higher than 35% would be unacceptable.

An important limitation of a single-arm study concerns the choice of threshold rate when that threshold is based on results from studies of regimens that would no longer to be considered acceptable or in populations that are quite different. For example, for some regimens, higher pre-treatment viral load may be associated with higher probability of virologic failure. Thus if the population to be enrolled in the single-arm study is likely to have a much lower distribution of viral loads than populations in the earlier trials, then observing a failure rate lower than the threshold might be more readily achieved than if a population with comparable viral loads could be enrolled.

Power and Sample Size Considerations

The power, or true acceptance probability, of a single-arm study with a sample size of N participants can be expressed as the probability of concluding, at the completion of the study, that the regimen is acceptable (i.e., the upper bound of the CI around the proportion of participants experiencing virologic failure is below the prespecified threshold), when the new regimen is truly acceptable (i.e., the true failure rate takes some value that is less than the threshold). This probability can be expressed as follows:

$$\sum_{x=0}^{n} \binom{n}{x} p_1^x (1-p_1)^{n-x}\, I(p_u < p_0)$$

where $1-\alpha$ is the desired two-sided confidence level (e.g., 95%), p_0 is the failure rate threshold, p_1 is the true (unobservable) failure rate for the study regimen, p_U is the upper bound of a $100(1-\alpha)\%$ CI for failures among the sample size of participants, and $I(\,)$ is the indicator function (takes value 1 if the enclosed statement is true and 0 otherwise).

For a given power and confidence interval coverage (e.g., 95%), the corresponding necessary sample size, N, is found recursively/iteratively by solving for N in the following equation:

$$\frac{\alpha}{2} = \sum_{k=0}^{x} \binom{n}{k} p_u^k (1-p_u)^{n-k}$$

Figure 2 illustrates the relationship between power (y-axis), using a 95% CI, and the true (unobservable) virologic failure rate (x-axis) when the threshold of interest is 35% for sample sizes, N, of 50, 100, and 150. For a given true failure rate, as the sample size is increased, the power of a study increases. For example, for a true failure rate of 20%, the power increases from 57% for a sample size of 50 to 91% and 98% for samples sizes of 100 and 150, respectively. In A5262, a sample size of 100 was selected; anticipating a 10% loss to follow-up, the accrual target was then increased to 111 participants.

Table 1 illustrates how the power of a study with a sample size of 100 patients (prior to any adjustment for loss to follow-up) varies according to the underlying true failure rate and the threshold being used. For a given true failure rate, a lower failure rate threshold gives lower power (and so the sample size might be increased from 100 to achieve good power). Conversely, for a given threshold, the power is also lower as the true underlying failure rate increases.

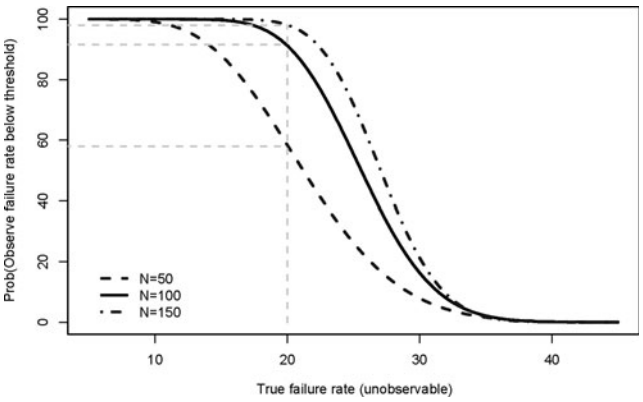

FIG. 2. Operating characteristics of two-sided 95% confidence interval for a failure rate threshold of 35% for sample sizes of 50, 100, and 150 participants.

Table 1. Power (Probability That 95% Confidence Interval Will Exclude Threshold Failure Rates) for Given True Failure Rates and a Fixed Sample Size of 100 Evaluable Subjects

Threshold (rate to exclude by 95% CI)	*True virologic failure rate*		
	15%	*20%*	*25%*
40.0%	>99%	99%	90%
37.5%	>99%	97%	72%
35.0%	>99%	91%	55%
32.5%	99%	81%	37%
30.0%	93%	56%	14%

CI, confidence interval.

Study Monitoring

Single-arm studies are usually small and—using an early outcome measure (e.g., week 24 virologic failure rate)—responses are observable in a relatively short period of time. This characteristic facilitates monitoring of the data as they accumulate. Ideally, there should be an early interim monitoring. A stopping guideline sets out the circumstances under which the study could be terminated prior to study completion by all participants; this is usually when there is strong evidence of unacceptable efficacy. On the other hand, if evidence suggests a satisfactory regimen, it is beneficial to continue the study to obtain more precise efficacy and safety information. Thus the primary purpose of interim monitoring is to evaluate virologic failure rates and safety data early enough in the study to prevent exposing current or additional participants to a regimen associated with unacceptably poor efficacy. Interim monitoring is usually performed by an entity independent of the study team. In addition to efficacy monitoring, management of individual participants for toxicities and treatment nonresponse is important, and is most often done by the study team and clinical site investigators.

The schedule of interim monitoring and the stopping guideline should be prespecified in the protocol after considering several factors including the primary endpoint week, estimated accrual rate, and duration of the trial. To minimize the risk of obtaining misleading findings from the interim review, the review should be scheduled for a time when the number of participants who have reached the primary endpoint will be sufficient to estimate the virologic failure rate with reasonable precision.

A5262 was monitored by an ACTG-appointed study monitoring committee (SMC) that was independent of the study team and clinicians enrolling patients into the study. The initial interim review by the SMC was to occur at the earlier of (1) 24 weeks after enrollment of the 40th participant (when it was expected that the primary endpoint would be available for about one-third of participants), or (2) 1 year after the enrollment of the first participant, which meets requirements of the National Institutes of Health for at least annual monitoring. The stopping guideline specified that the SMC might recommend closing or modifying the study if 19 or more of the first 40 participants to reach week 24 exhibited virologic failure. After the initial interim review, the study was to be monitored annually by the SMC. The A5262 stopping guideline was defined using operating characteristics such as those in Table 2.

A desirable operating characteristic of any stopping guideline is that it leads to early study termination with high probability when the underlying true failure rate is unacceptably high, but with low probability when the true failure rate might be considered acceptable. At an interim analysis, one approach for defining a stopping guideline is to consider stopping if a CI around the observed failure rate is entirely above the threshold failure rate (as in scenario B in Fig. 1). Given the sample size at the interim analysis (e.g., 40 in A5262), the number of failures that need to be observed in order for this to occur can then be calculated. For A5262, with 21 failures among 40 subjects, the observed proportion failing is 53% and the associated CI is 37% to 68%, so just excluding 35%. Table 2 shows the probability of meeting this criterion for various underlying true probabilities of failure.

The study team was concerned that the probability of meeting this criterion might not be sufficiently high when the true probability of failure was quite unacceptable (e.g., 68% probability of meeting this criterion when the true failure rate was 55%). They therefore considered using a slightly lower

Table 2. Several Choices of Stopping Guideline (Threshold and Number of Participants Examined at Interim Evaluation), and the Probability of Stopping Under Various True Failure Probabilities

Number of participants examined	*Interim failure rate threshold*	*Smallest number of participants where 95% CI excludes threshold*	*(Observed proportion)*	*True probability of failure*							
				0.25	*0.30*	*0.35*	*0.40*	*0.45*	*0.50*	*0.55*	*0.60*
36	0.30	17	(0.47)	<1%	2%	9%	24%	46%	69%	87%	96%
36	0.35	19	(0.53)	<1%	<1%	2%	8%	22%	43%	67%	85%
38	0.30	18	(0.47)	<1%	2%	8%	22%	45%	69%	87%	96%
38	0.35	20	(0.53)	<1%	<1%	2%	8%	22%	44%	68%	86%
40	0.30	19	(0.48)	<1%	1%	7%	21%	43%	68%	87%	96%
40	0.35	21	(0.53)	<1%	<1%	2%	7%	21%	44%	68%	87%
42	0.30	20	(0.48)	<1%	1%	6%	20%	42%	68%	87%	96%
42	0.35	22	(0.52)	<1%	<1%	2%	7%	21%	44%	69%	88%
44	0.30	20	(0.45)	<1%	2%	10%	28%	53%	77%	92%	98%
44	0.35	23	(0.52)	<1%	<1%	1%	7%	21%	44%	70%	88%

threshold for interim monitoring, specifically 30%. For this interim monitoring threshold, observing 19 failures among 40 subjects gave a higher probability of meeting the criterion when the true probability of failure was quite unacceptable (e.g., 87% probability when the true failure rate was 55%). Thus, in the event the regimen performs very poorly and 19 participants are seen to fail before the 40th participant reaches week 24, the team will notify the SMC for an earlier review. Table 2 shows related results for defining a stopping guideline for sample sizes close to 40, to allow some flexibility in practice in the timing of the interim analysis (e.g., if slightly fewer or more patients have the necessary data at the time that a monitoring committee meeting occurs). This approach can be extended to multiple interim reviews. Note that the CI is being used here as a tool for choosing an acceptable stopping guideline in the situation in which the regimen is unlikely to be useful in practice and so adjustment of the CI for repeated analyses is not a major concern.

In theory, however, if a trial involves interim analyses, the calculation of multiple CIs (at each interim analysis and at the final analysis), each with a coverage probability $100 \cdot [1 - \alpha]\%$, requires adjustment for these repeated analyses. In a study that continues to its full sample size, interpretation of the 95% CI is minimally impacted by the stopping rule and in general practice the usual calculation of the (nominal) 95% CI is reasonable. For example, for A5262 the probabilities of stopping early under true failure rates of 15%, 20%, and 25% are all less than 0.2%, not enough to change the probabilities in Table 1 appreciably. We do not advocate stopping a single-arm study early for efficacy. However, if the investigator chooses to do so, methods for calculating repeated CIs that allow for early stopping might be considered.[11]

Conclusions

We have described design considerations for single-arm studies in HIV-infected treatment-naive patients. A single-arm design may also be appropriate in studies of treatment-experienced patients who are difficult to recruit or for whom restricted viable treatment options limit the availability of a control regimen. This is illustrated by the ongoing study evaluating maraviroc plus raltegravir plus darunavir/ritonavir in exclusively CCR5-tropic, integrase-naive and darunavir-naive HIV-1-infected patients with a history of triple-class ARV treatment failure (NCT01013987). Furthermore, single-arm trials can provide supportive data for drug development.[12] An example is the single-arm study requested by the FDA to provide data on additional treatment-experienced subjects at the recommended dose of darunavir/r to refine estimated rates of failure and adverse events.[13] Lastly, we note that in a guidance about developing directing agents for the treatment of hepatitis C virus (HCV), the FDA acknowledges the value, in certain situations, of single-arm prospective trials with historical controls in evaluating drugs for the treatment of HCV infection in HIV/HCV coinfected subjects.[14]

Single-arm phase II studies have been commonly used in oncology to assess the antitumor activity of new drugs. The primary goal of oncology phase II trials is not to provide definitive evidence of drug efficacy, but to propose a promising drug for further investigations. Two-stage and multistage designs intended to control for the average number of patients required to make a correct decision have been well studied in the literature.[15–17] These designs require pauses to accrual between stages.

Key potential shortcomings of single-arm studies include limited generalizability to populations not included in the study or comparability to other studies since observed failure rates can be due to factors other than the investigational regimen. For example, a very high virologic failure rate in a single-arm study may be attributable to enrollment of a poorly adherent population and not the antiviral efficacy of the investigational regimen. Also, regimens in the RCTs considered in designing a single-arm study, hence the assumptions in designing the single-arm study, may be obsolete by study completion, although this is less likely to occur for small, short-duration single-arm studies.

Despite these limitations, single-arm studies have a unique role in clinical trials when an RCT is not feasible or desirable, and can provide critical pilot efficacy and safety data on novel antiretroviral regimens.

Acknowledgments

This research is supported by the AIDS Clinical Trials Group and funded by the National Institute of Allergy and Infectious Diseases (1U01AI068636 and 1U01AI068634).

Author Disclosure Statement

B.T. has served as an advisor and received research support and honoraria from Tibotec. J.J.E. is a consultant to Bristol Myers Squibb, GlaxoSmithKline, Merck, ViiV, Tibotec, and Gilead.

References

1. Food and Drug Administration Guidance for Industry: Antiretroviral Drugs Using Plasma HIV RNA Measurements—Clinical Considerations for Accelerated and Traditional Approval. Appendix B. Division of Antiviral Drug Products: Office of Drug Evaluation IV in the Centre for Drug Evaluation and Research (CDER), Rockville, MD, 2002.
2. Food and Drug Administration Guidance for Industry Non-Inferiority Clinical Trials: Draft March 2010. Available at www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf. Accessed February 8, 2011.
3. European Medicines Agency Committee for Medicinal Products for Human Use: EMA/CHMP Guideline on the Clinical Development of Medicinal Products for the Treatment of HIV Infection. September 2009. Available at www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003460.pdf. Accessed February 8, 2011.
4. Taiwo B, Zheng L, Gallien S, *et al.:* Efficacy of nucleoside-sparing regimen of darunavir/ritonavir plus raltegravir in treatment-naive HIV-1-infected patients (ACTG A5262). AIDS 2011;25:2113–2122.
5. Weinstock H, Respess R, Heneine W, *et al.:* Prevalence of mutations associated with reduced antiretroviral drug susceptibility among human immunodeficiency virus type 1 seroconverters in the United States, 1993–1998. J Infect Dis 2000;182:330–333.
6. Johnson VA, Brun-Vezinet F, Clotet B, *et al.:* Update of the drug resistance mutations in HIV-1: December 2008. Top HIV Med 2008;16:138–145.
7. Kalbfleisch JD and Prentice RL: *The Statistical Analysis of Failure Time Data.* John Wiley, Hoboken, NJ, 1980.

8. Gulick RM, Ribaudo HJ, Shikuma CM, *et al.:* Three- vs. four-drug antiretroviral regimen for the initial treatment of HIV-1 infection: A randomized controlled trial. JAMA 2006;296: 769–781.
9. Riddler SA, Haubrich R, DiRienzo AG, *et al.:* Class-sparing regimens for the initial treatment of HIV-1 infection. N Engl J Med 2008;358:2095–2106.
10. Markowitz M, Nguyen BY, Gotuzzo E, *et al.:* Rapid and durable antiretroviral effect of the HIV-1 integrase inhibitor raltegravir as part of combination therapy in treatment-naive patients with HIV-1 infection: Results of a 48-week controlled study. J Acquir Immune Defic Syndr 2007;46:125–133.
11. Jennison C and Turnbull BW: *Group Sequential Tests with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL, 2000.
12. Chan-Tack KM, Struble KA, Morgensztejn N, *et al.:* HIV clinical trial design for antiretroviral development: Moving forward. AIDS 2008;22:2419–2427.
13. Molina JM, Cohen C, Katlama C, *et al.:* The safety and efficacy of darunavir with low-dose ritonavir in treatment-experienced patients: 24 week results of POWER 3. J Acquir Immune Defic Syndr 2007;46:24–31.
14. Food and Drug Industry Guidance for Industry: Chronic Hepatitis C Virus Infection: Developing Direct-Acting Antiviral Agents for Treatment (DRAFT GUIDANCE). September 2010. Available at www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM225333.pdf. Accessed November 8, 2011.
15. Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 1989;10:1–10.
16. Ensign LG, Gehan EA, Kamen DS, and Thall PF: An optimal three-stage design for phase II clinical trials. Stat Med 1994;13:1727–1736.
17. Johnson VE and Cook JD: Bayesian design of single-arm phase II clinical trials with continuous monitoring. Clin Trials 2009;6:217–226.

Address correspondence to:
Lu Zheng
Statistical Data Analysis Center
Harvard School of Public Health
Boston, Massachusetts 02115

E-mail: szheng@sdac.harvard.edu