

Some Ethical Issues in Phase II Trials in Acute Leukemia

Peter F. Thall, PhD, and Elihu H. Estey, MD

Dr. Thall is a Professor in the Department of Biostatistics and Applied Mathematics and Dr. Estey is a Professor in the Department of Leukemia at the University of Texas M. D. Anderson Cancer Center in Houston.

Address correspondence to:

Peter F. Thall, PhD, Dept. of Biostatistics and Applied Mathematics, Box 447, The University of Texas, M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030; Tel: 713-794-4162; E-mail: rex@mdanderson.org.

Abstract: This paper addresses several scientific and ethical issues that arise in the design and conduct of phase II clinical trials of experimental therapies. Although we discuss chemotherapy trials in acute leukemia, the issues pertain to a much larger class of early-phase clinical trials. Our focus is on the manner in which numerical values of standard and targeted response rates of a statistical design are specified, the number of interim tests that are applied, and the effects of these design features on early stopping probabilities and the treatments that patients actually receive. These points are illustrated by numerical comparisons of alternative designs for a particular phase II trial of a new drug for relapsed or refractory acute myelogenous leukemia. We show that statistical designs that target inappropriately low response rates or that apply early stopping rules too infrequently are at odds with good statistical and medical practice and that such designs often provide less benefit to the patients in the trial than would be obtained by simply treating all patients with standard therapy. The general conclusions are that statistical designs have both scientific and ethical implications, and that science, statistics, and ethics cannot be treated as separate issues.

Three medical approaches are possible for any given illness: no treatment, treatment with standard therapy, and treatment with investigational therapy. Because few untreated patients with acute myeloid leukemia (AML) live more than 1 year, and because, in most cases, standard treatment does not improve prognosis, guidelines promulgated by academic cancer centers typically recommend investigational therapy for many such patients. These guidelines also advise that investigational treatments be administered within the context of a formal clinical trial in order to gain knowledge about their actual effects on patients. Nearly all AML patients simply accept their physician's recommendation for treatment because they rely on their physician's expertise in choosing from the wide variety of available treatments for AML. Thus, when a physician advises an AML patient to enter a trial of an investigational agent, the patient undoubtedly assumes that this will serve his or her best interests. It has become apparent to us, however, that for many patients the recommended trial, as designed, is not their best option. This gap between assumption and reality results from 2 aspects of clinical trial design. First, the ethical underpin-

Keywords

Chemotherapy, design, medical ethics, phase II clinical trial.

nings of some designs can be questioned, especially with regard to the direct consequences to a trial participant. Second, the scientific structure of these designs is frequently inadequate, limiting what may be learned about new treatments. In particular, we argue that many designs waste patient data. This wastefulness is at odds with good statistical and scientific practice, as well as patients' often-cited desire to benefit future patients as motivation for participation in a clinical trial.

Our primary focus is the dubious ethical bases for some clinical trials because we feel it is the more serious issue. We discuss the following specific points: the common practice of targeting inappropriately low response rates; the fact that many designs are insufficiently adaptive; patient heterogeneity; and the practice of focusing on the activity of a single regimen. We provide numerical illustrations of these points based on comparisons of alternative clinical trial designs. Although we focus on trials in acute leukemia, the methods that we discuss have a much wider range of application.

Low Target Response Rates

As a basis for illustration, we use a multicenter phase II trial of the new drug VNP-4010IM (Cloretazine, Vion). The trial began in 2004 and enrolled patients with relapsed or refractory AML and a first remission lasting less than 1 year or with untreated AML and aged 59 or above ("older"). The trial used an optimal 2-stage Simon design.¹ The protocol made no distinction between relapsed/refractory patients and untreated older patients. The targeted complete response (CR) rate was set at 20% ($P_1=.20$), with a null rate of 5% ($P_0=.05$), with type I and type II error probabilities both 0.10. Thus the design had a 90% probability of rejecting the drug for future study if its true response rate was "of no interest" (corresponding to the null hypothesis that $P=\text{prob}(\text{CR})=P_0=.05$), and it had a 90% probability (power) of accepting the drug for future study if its true response rate was "of interest" (with the alternative hypothesis that $P=P_1=.20$). The Simon optimal 2-stage design with these operating characteristics treats 12 patients in stage 1, stops accrual with acceptance of the null hypothesis if no responses are observed in these 12 patients, enters 25 more patients in a second stage if at least 1 of the first 12 responds, and accepts the alternative hypothesis if 4 or more responses are observed in the 37 patients accrued in both stages combined. All of these design parameters were based on the assumption that the probability of CR with VNP-4010IM would be the same in relapsed/refractory patients and untreated older patients.

Since the null CR rate ($P_0=.05$) used in applying the Simon design presumably reflected historical data,

it is instructive to examine actual CR rates following the use of standard treatment (cytarabine [ara-C]-containing regimens, specifically ara-C plus idarubicin) as given at the University of Texas M. D. Anderson Cancer Center from 1990 to 2004. The CR rate for patients with relapsed or refractory AML was 0.19 (69/356; 95% confidence interval [CI], 0.16–0.24), while that for untreated patients age 59 and above was 0.45 (46/103; 95% CI, 0.35–0.54). Thus, the targeted rate of interest specified in the protocol, $P_0=.20$, represented no improvement over the CR rate available with standard therapy for relapsed/refractory patients, while this target was well below the CR rate that could almost certainly be obtained with standard therapy in untreated older patients. This illustrates 2 obvious flaws with the design: it failed to discriminate between 2 patient subgroups having very different prognoses, and the targeted rate was inappropriate for each of the subgroups as it was too low in one and far too low in the other.

We now evaluate the consequences of using the 2-stage Simon design with a null rate of $P_0=.05$ and a desired improvement of $P_1-P_0=.15$ to achieve the targeted $P_1=.20$, which is the design used in the VNP-4010IM trial. For simplicity, we will focus on the subgroup of untreated older patients and assume a null CR rate of 0.40, which is slightly below the historical rate in these patients. Once we have explored this simple case, we will return to the issue of patient heterogeneity. We thus focus on a design with $P_0=.40$, and use the same type I and type II error rates of 0.10 and the same desired 0.15 improvement, so that the targeted alternative is $P_1=.55$. Because this design has a much larger null rate of $P_0=.40$ rather than $P_0=.05$, it requires a much larger number of patients to detect the same improvement of 0.15 with the same reliability. Specifically, for the optimal 2-stage Simon design in this case, 38 patients are enrolled in stage 1 and accrual stops with acceptance of the null if 16 or fewer responses are seen in these first 38 patients. If at least 17 responses are observed in the first stage, then another 50 patients are entered in the second stage for a maximum total sample size of 88. The alternative is accepted and the drug is considered promising if at least 41 of the 88 patients respond.

With both designs, we will also assume that if the trial is terminated after the first stage, then the remaining patients (ie, those who would have been accrued in the second stage) are given the standard therapy, recalling that under the null hypothesis, as based on historical data, this therapy produces at least a 40% response rate in untreated patients. We will then compute the expected number of responses lost (ERL) compared to the simple plan of treating all patients (37 with the design that was used, 88 with the alternative design) with standard therapy (Table 1). For example, when the true response rate with the experimental drug (here VNP-4010IM) is $P=.40$, the ERL is

Table 1. Comparison Between the Design Employed in the VNP-4010IM Trial With an Alternative Design Based on a Null Rate Equal to the Actual Historical Rate

Interim Tests	Employed Design $P_0=.05, P_1=.20$ $n_{max}=37$			Alternative Design $P_0=.40, P_1=.55$ $n_{max}=88$		
	At $n=12$			At $n=38$		
P_1	PET	ER	ERL (%)	PET	ER	ERL (%)
.05	.54	6.6	8.2 (56)	1.00	21.9	13.3 (38)
.10	.28	5.8	9.0 (61)	1.00	23.8	11.4 (32)
.20	.07	7.7	7.1 (48)	1.00	27.6	7.6 (22)
.30	.01	11.1	3.7 (25)	.96	31.2	4.0 (11)
.40	.002	14.8	0	.67	35.2	0
.50	<.001	18.5	-3.7 (-25)	.21	43.0	-7.8 (-22)
.55	<.001	20.3	-5.6 (-37)	.08	47.8	-12.6 (-36)

ER = expected number of responses; ERL = expected number of responses lost; n_{max} = maximum total sample size; PET = probability of early termination.

zero, since this is also the response rate with standard therapy. However, the employed design is very unlikely to stop the trial after the first stage even when $P=.10$ (0.30 below the standard CR rate), with a probability of early termination (PET) of 0.28. In contrast, the PET values are much higher with the alternative design for small P values, so this design has a much smaller percentage of lost responses compared to treating all patients with standard therapy. Thus with the design that was used, characterized by specifying $P_0=.05$, if the true response rate is $P=.20$ (the targeted value) then 7.7 responses would be expected as compared to 14.8 responses if all 37 patients had received standard therapy. On average, the design that was used achieves only 52% of the number of responses that would be expected by simply treating all patients with standard therapy; equivalently, 48% of the expected responses are lost. In contrast, under the same scenario, the design that uses the more appropriate $P_0=.40$, based on historical data, results in an ERL of only 22%.

The VNP-4010IM design is not exceptional; many phase II trials specify inappropriately low values for P_0 . We suspect that this may be done due to confusion between a response rate that may be of interest based on evidence that a drug is “active,” and a response rate that is

of interest based on evidence that the drug is better than standard therapy. The above illustration is motivated by the belief that the latter is of greater fundamental interest to patients and thus should be the basis of an ethical design. Regardless of the explanation, however, we hope that the above example, and the more extensive analysis in Table 1, make clear the clinical consequences of targeting an inappropriately low response rate.

A More Adaptive Design

Table 1 displays the design employed in the VNP-4010IM study as well as an alternative design, which, while preferable to the employed design, as a 2-stage design has the drawback of requiring all 38 patients in stage 1 to receive experimental therapy. Thus, even if the true CR rate is a very small value substantively inferior to standard therapy, with $P=.20$ or less, at least 38 patients still must be treated with the experimental regimen. If, hypothetically, none of the first 18 patients achieved a CR, then assuming from a Bayesian viewpoint that P followed a beta (.20,.80) prior, which has an expected value of 0.20 and effective prior sample size equal to 1 patient, the posterior probability that $P>.20$ would be .001 based on a run of 18 treatment failures. This is very strong evidence that even the 0.20 CR rate is unlikely. This is just a statistical way of quantifying how unpromising a treatment that yields 0/18 responses actually is, and it leads to the natural question of why a physician would want to treat an additional 20 patients before applying a stopping rule, or even enroll a 19th patient in the trial. Clearly, it would be very desirable to use a design that allows one to stop earlier in such cases. A more adaptive design with more and earlier decisions based on the incoming data is needed. While there are numerous “frequentist” designs based on tests of hypotheses that do this,^{2,3} for simplicity we will describe a Bayesian design that allows several interim looks. This is an example of a family of Bayesian phase II designs that are extremely flexible,^{4,5} whose use has been validated in numerous phase II trials over the past decade.

To provide more adaptive monitoring, we constructed a Bayesian design with the same 88 patients and based on the same average null and alternative values ($P_0=.40$ and $P_1=.55$) but with interim stopping rules applied after every 15 patients have been treated and evaluated. Thus, there are up to 5 interim analyses, at 15, 30, 45, 60, and 75 patients. The Bayesian design assumes that P_0 and P are random, with P_0 following a beta (400,600) prior, and P following a beta (.80,1.20) prior. Both priors have mean 0.40, but the historical standard prior is very informative while that of the experimental agent’s CR probability is very uninformative, equivalent to having information on 2 patients. The trial is stopped early if the posterior

Table 2. Contrasts Between Simon Optimal 2-Stage and Bayesian Multistage Designs, Both Based on $P_0=.40$ and $P_1=.55$ With Maximum Sample Size 88 Patients

Interim Tests	Simon Optimal 2-Stage			Bayesian Multistage		
	One at n=38			Up to 5 at n=15, 30, 45, 60, 75		
	True P	PET	ER	ERL (%)	PET	ER
.05	1.00	21.9	13.3 (38)	1.00	31.1	4.1 (12)
.10	1.00	23.8	11.4 (32)	1.00	31.5	3.7 (10)
.20	1.00	27.6	7.6 (22)	1.00	32.1	3.1 (9)
.30	.96	31.2	4.0 (11)	1.00	32.9	2.3 (6)
.40	.67	35.2	0	.85	35.2	0
.50	.21	43.0	-7.8 (-22)	.29	42.7	-7.5 (-21)
.55	.08	47.8	-12.6 (-36)	.09	47.7	12.5 (-36)

PET = probability of early termination; ER = expected number of responses; ERL = expected number of responses lost.

probability $\Pr(P_0 + .15 < P \mid \text{data})$ is less than 0.04 at any interim look, which translates to stopping if the number of CRs out of the number of patients evaluated is less than or equal to 4/15, 11/30, 18/45, 26/60, or 33/75. Thall and Sung provide details of how to construct this sort of design,⁶ and a computer program for carrying out the computations is freely available at <http://biostatistics.mdanderson.org>. Table 2 provides a comparison of the operating characteristics of this design to the Simon 2-stage design with these values of P_0 and P_1 . Because the decision that the new treatment is not superior and that the trial should be stopped early can be made much more quickly with multiple interim analyses, the multistage design is much more efficient. While the 2 designs have nearly identical values of PET for the targeted $P=.55$, the multistage design has much larger PET values for smaller values of P . The ethical consequence of this is seen in the expected numbers of responses that achieved true $P=.40$ or smaller, which are much larger for the multistage design, and so the corresponding number of responses lost is much smaller. For desirably large values of $P=.50$ or $.55$, the 2 designs have virtually identical operating characteristics, which illustrates the general fact that a properly calibrated design with a greater number of interim looks at the data is safer without sacrificing the ability to identify a true treatment advance.

It is important to emphasize that we are not arguing against the use of 2-stage designs, but rather against their misuse. Indeed, when Simon first introduced the 2-stage designs in 1989 along with good quality computer code for implementation, they provided a substantive improvement over single-arm trials conducted without any interim stopping rule at all. Our first point is that early stopping rules do not function well when the design has not been parameterized properly to reflect comparison to the actual historical rate. Such an improper parameterization is likely to render any statistical design dysfunctional. The second point is that an early stopping procedure should be constructed so that it stops the trial as soon as the interim data show that an experimental treatment is likely not to be promising. In a trial where the best available treatment has a P_0 of .05 or smaller, the goal is essentially to detect whether the new agent has any anti-disease activity, and in such cases a target of $P_1=.20$ is scientifically and ethically appropriate. In such trials, stopping early does not really protect patients, rather it clears the way for study of newer experimental agents. When there is a standard therapy with a substantively large P_0 value, however, early stopping rules have very important ethical consequences.⁷

Patient Heterogeneity

The previous example focused on a subgroup of untreated older patients. The VNP-4010IM trial, however, also included relapsed/refractory patients, and the historical CR rates in the 2 groups are very different (0.45 for older untreated patients and 0.19 for relapsed/refractory patients). Thus, for example, a new treatment achieving an actual CR rate of 0.40 in relapsed/refractory patients would be a desirable treatment advance, while this rate would not provide an improvement over the standard in untreated older patients. The point is that, because these 2 groups have very different CR rates with standard therapy, what constitutes an improvement over standard therapy also is different between the groups. Consequently, it does not make sense to conduct a trial including both groups that has a single overall targeted CR rate. There are 2 sensible alternative approaches to this problem. The first, which is simple, is to conduct separate trials in the 2 groups. For example, the trial in the relapsed/refractory group might be based on $P_0=.19$ and $P_1=.19+.15=.34$, and the trial in the untreated older patients might use $P_0=.45$ and $P_1=.45+.15=.60$, or possibly improvements other than .15 could be targeted. This design solves the problem of patient heterogeneity, but it has the undesirable property that it does not allow one to borrow strength between the 2 subgroups. That is, if an improvement over the historical rate is seen in relapsed/refractory patients, this should provide evidence that an improvement in untreated

older patients also is likely. As an alternative, a single trial including both subgroups may be conducted using a regression model to account for prognosis. This may be done in numerous ways. For example, a simple approach is to define the covariate $Z = 1$ if the patient is untreated and older and $Z = 0$ if the patient is relapsed/refractory, let P_Z denote the CR rate in subgroup Z , and assume the logistic regression model $\text{logit}(P_Z) = \log\{P_Z/(1-P_Z)\} = \alpha + \beta Z$. Then $\text{logit}(P_Z) = \alpha + \beta$ for untreated older patients and $\text{logit}(P_Z) = \alpha$ for relapsed/refractory patients, with the parameter β accounting for prognostic subgroup. Early stopping criteria could then utilize data from both subgroups. Many versions of such covariate-adjusted phase II designs are possible.⁸⁻¹⁰ What is important is to account for patient heterogeneity when it is large enough such that a design that ignores heterogeneity does not make sense.

The Fallacy of Single-Arm Phase II Trials

We have argued that it is ethically important to set standards in a phase II trial that reflect comparison to what can be achieved with standard therapy, and moreover that it may be very desirable to monitor the accruing data more intensively than taking only a single interim look. However, the cases that we have examined greatly simplify actual clinical settings, and numerous complicating issues remain. These include multiple outcomes such as adverse treatment effects, the fact that early patient outcomes such as CR may be inadequate surrogates for survival time, and settings in which multiple courses of therapy are given over an extended period of time. Although we cannot deal with all of these issues here, we will briefly address the issue of treatment-trial confounding.

The observed difference in outcome between 2 treatments administered in 2 separate single-arm trials is the sum of (a) the actual difference between the treatments, typically called the treatment effect; (b) differences due to observable patient prognostic covariates; and (c) differences due to unobserved variables in the patients and the therapeutic environments of the 2 trials. The variables causing this third class of effects are sometimes referred to as latent variables and their combined effect as trial effect.¹¹ Examples include differences in supportive care practices, the types of patients enrolled, and the skill of the physicians and nurses caring for the patients. Although differences in quantifiable covariates (eg, age, cytogenetics) can be accounted for, the same is not true of trial effects. Thus, when comparing treatments that have been evaluated in separate trials, in particular with comparison to historical data, the treatment and trial effects are completely confounded. Indeed, it has been demonstrated that differences in outcome beyond those attributable to chance

arise when the same treatment is given in 2 separate trials and that these differences persist even after accounting for the effects of known prognostic covariates.¹² This observation of course motivates the accepted use of the randomized phase III trial as the arbiter of the superiority of one treatment over another. Given this, it seems peculiar that the decision to proceed to a phase III trial of a new drug is commonly based on the performance of the drug in a single-arm phase II trial, despite the confounding between treatment and trial effects inherent in the evaluation of data from such trials. This practice can also be criticized from a patient's perspective. The most common question posed by patients to physicians involved in trials of new drugs is, "Which of the drugs that you have available is best?" This question indicates that patients view phase II trials as inherently comparative. It follows that patients' interests are poorly served by single-arm phase II trials.

A simple alternative is to conduct a randomized phase II trial employing a selection design that randomizes patients among several treatments, including 1 or more experimental regimens and possibly the standard. The objective is to select 1 experimental treatment that is best,¹³ although designs that allow more than 1 experimental treatment to be selected certainly may be used.¹⁴ This eliminates trial effects and thus ensures unbiased comparisons of the experimental therapies to each other and to the standard. Designs that do not include the standard but instead randomize patients among 2 or more experimental treatments are useful because they provide unbiased comparisons among the treatments studied, hence the selection is much more reliable than if a sequence of single-arm trials were conducted.¹⁵ In any case, the selected therapy or therapies may be studied further, for example, in comparison to the standard treatment in terms of survival or disease-free survival (DFS) time.

A selection design that aims to select the best among several experimental therapies regardless of the difference in success rates among the therapies requires fewer patients (eg, a maximum of 15–20 per treatment) than designs (eg, the Simon 2-stage) whose goal is to test the hypothesis that a given therapy is better than another by at least a given amount. Simulation studies typically indicate that the probability of selecting a truly superior therapy is 60–70%, corresponding to a power of 60–70%. Because physicians are accustomed to 80–90% power, the selection design is sometimes criticized as an "underpowered phase III trial." However such criticism ignores the fact that selection designs have the less demanding goal of choosing a best treatment rather than demonstrating a given degree of improvement, as well as the fact that any experimental therapy selected in this way still must reliably show an improvement in survival or DFS in a subsequent trial. That is, the randomized selection design

is not a substitute for a confirmatory phase III trial, but simply a much more efficient way to screen new therapies. Randomized selection designs are especially useful in settings in which there are several new agents that might be tested. Experience suggests that, at least in AML, a preclinical rationale cannot substitute for clinical data in deciding which new agent is best. Yet experience also suggests that the decision as to which new agent to test is made informally in the absence of such data. For example, if 3 new agents are available for evaluation in patients with relapsed/refractory AML, 1 must be selected for a subsequent comparison with standard therapy, and the preclinical rationale for using each seems equally compelling. It follows that the probability of selecting the best agent is 0.33. It is this figure, not 80–90%, that should be compared with the 60–70% correct selection probabilities that typify selection designs. Such designs thus should be viewed as an attempt, using relatively few patients, to substitute empiricism for informality in selecting which new drugs are worthy of further investigation.

By extension, we view the phase II/III dichotomy as artificial. In addition to the nonrandomized nature of the single-arm phase II trial, the distinction between phase II and phase III prevents use of phase II data in final evaluation of a drug because the data did not arise from a randomized trial. This is particularly unfortunate because, with respect to important endpoints such as survival, the phase II data are the most mature. Inoue, Thall, and Berry¹⁶ and Liu and Pledger¹⁷ have proposed phase II/III designs that randomize patients throughout. In the Inoue et al design, at each of several interim analyses a decision is made to stop the trial if a treatment is superior, to stop if it is implausible that any treatment will be superior (“futility”), or to expand the trial to include other centers if the accumulating data indicate that many more patients will be needed to reach the conclusion that one treatment is superior to the other. At this point, the phase III aspect of the phase II/III design is said to start. Such designs are intuitively appealing and have been demonstrated to result in substantial savings of both time and sample size.

Conclusions

We believe a cultural gap currently exists between physician and statistician, with the former often viewing the latter as divorced from clinical reality. While, unfortunately,

this viewpoint is accurate in many cases, the segment of the biostatistical community that works closely with physicians has provided a wide array of practical clinical trial designs that have ethically desirable properties. We hope that this paper will convince the reader that there is an important ethical dimension to medical statistics and that science and ethics cannot be separated when designing a clinical trial. In particular, we have argued that setting target response rates too low or looking too infrequently at available clinical trial data may result in lost responses, and that the focus on single-arm phase II clinical trials may be antithetical to the best interests of patients. We have highlighted statistical methods that we firmly believe are more relevant than conventional methodologies to the scientific and ethical imperatives of clinical research.

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10:1-10.
2. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982;38:143-151.
3. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Stat Med*. 1997;16:2701-2711.
4. Thall PF, Simon R, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med*. 1995;14:357-379.
5. Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clin Oncol*. 1996;14:296-303.
6. Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med*. 1998;17:1563-1580.
7. Thall PF. Ethical issues in oncology biostatistics. *Stat Methods Med Res*. 2002;11:429-448.
8. Thall PF, Sung H-G, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *J Am Stat Assoc*. 2002;97:29-39.
9. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LO, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat Med*. 2003;22:763-780.
10. Thall PF, Wathen JK. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Stat Med*. 2005;27:1947-1964.
11. Thall PF, Wang X. Bayesian sensitivity analyses of confounded treatment effects. In: Crowley J, Pauler D, eds. *Handbook of Statistics in Clinical Oncology: Second Edition, Revised and Expanded*, New York: Marcel-Dekker, 2005. In press.
12. Estey EH, Thall PF. New designs for phase 2 clinical trials. *Blood*. 2003;102:442-448.
13. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*. 1988;75:303-310.
14. Schaid DJ, Wieand S, Therneau, TM. Optimal two-stage screening designs for survival comparisons. *Biometrika*. 1990;77:507-513.
15. Simon R, Wittes RE, Ellenberg, SS. Randomized phase II clinical trials. *Cancer Treat Rep*. 1985;69:1375-1381.
16. Inoue LYT, Thall PF, Berry, DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*. 2002;58:823-831.
17. Liu Q, Pledger GW. Phase 2 and 3 combination designs to accelerate drug development. *J Am Stat Assoc*. 2005;100:493-502.