VOLUME 30 · NUMBER 6 · FEBRUARY 20 2012

JOURNAL OF CLINICAL ONCOLOGY

STATISTICS IN ONCOLOGY

Design Issues in Randomized Phase II/III Trials

Edward L. Korn, Boris Freidlin, Jeffrey S. Abrams, and Susan Halabi

Edward L. Korn, Boris Freidlin, and Jeffrey S. Abrams, National Cancer Institute, Bethesda, MD; and Susan Halabi, Duke University Medical Center, Durham, NC.

Submitted July 29, 2011; accepted November 22, 2011; published online ahead of print at www.jco.org on January 23, 2012.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Edward L. Korn, PhD, Biometric Research Branch, EPN-8129, National Cancer Institute, Bethesda, MD 20852; e-mail: korne@ctep.nci.nih.gov.

0732-183X/12/3006-667/$20.00

DOI: 10.1200/JCO.2011.38.5732

ABSTRACT

Phase II trials are used to show sufficient preliminary activity of a new treatment (in single-arm designs or randomized screening designs) or to select among treatments with demonstrated activity (in randomized selection designs). The treatments prioritized in a phase II trial are then tested definitively against a control treatment in a randomized phase III trial. Randomized phase II/III trials use an adaptive trial design that combines these two types of trials in one, with potential gains in time and reduced numbers of patients required to be treated. Two key considerations in designing a phase II/III trial are whether to suspend accrual while the phase II data mature and the choice of phase II target treatment effect. We discuss these phase II/III design parameters, give examples of phase II/III trials, and provide recommendations concerning efficient phase II/III trial designs.

J Clin Oncol 30:667-671. © 2012 by American Society of Clinical Oncology

INTRODUCTION

The traditional sequential approach to development of anticancer therapies can be long, with new therapies screened in phase II trials, and those showing promising activity subsequently tested in large phase III trials. Combining phase II and III trials together in a phase II/III design has become increasingly more accepted and could offer greater efficiency both in streamlining the timeline and in allowing the data from the phase II patients to be used in the phase III analysis, thereby reducing the total number of trial patients.[1-8] Just as there are a number of possible phase II designs, there are a number of possible phase II/III designs corresponding to these phase II designs. We review phase II/III trial designs and provide general recommendations for choosing the design parameters.

TRIAL DESIGNS

Stand-Alone Phase II and III Trials

Table 1 displays typical design parameters for phase III and II trials. These prespecified parameters include the primary outcome, type I error (probability of rejecting the null hypothesis when there is no treatment benefit), target alternative hypothesis (target treatment effect), and power (probability of rejecting the null hypothesis when the target effect is true). Several phase II designs are available. For single agents that are expected to shrink tumors, a small single-arm design to document responses as a measure of activity may be appropriate.[9] When the experimental treatment is not expected to have responses (eg, it is cytostatic[10]), or when it involves a combination of agents, some of which are known to have responses, a randomized phase II screening design[11] is necessary to establish reliably clinical activity, unless historical data from multiple trials have been validated to provide an appropriate benchmark for a single-arm trial.[12] A randomized phase II screening trial typically involves random assignment between the same experimental and control treatments as the future phase III trial and requires specification of the same set of design parameters, but the parameter values differ. First, unlike in a phase III trial, in which the primary outcome and its targeted difference represent direct clinical benefit to the patient, in a randomized phase II trial, the primary outcome is only required to represent activity (eg, progression-free survival [PFS] gains); the corresponding phase II–targeted alternative hypothesis is chosen to represent a reasonably high probability that the experimental treatment will also be effective in a phase III trial. Second, in phase II designs, the targeted alternative hypothesis and type I error rates are typically larger than those in a phase III design, leading to a smaller required sample size.

When one needs to select a regimen (for further testing) among several experimental regimens that already have shown preliminary evidence of activity (eg, different doses/schedules of an active agent where the between-arm treatment differences would be expected to be relatively small), a randomized selection design may be used. The selection is typically made by choosing the experimental arm that is better on some predefined measure,

Table 1. Examples of Typical Trial Design Parameters for Stand-Alone Phase II and III Trials

Trial Type	Primary End Point	One-Sided Type I Error (%)	Power (%)	Target Alternative Hypothesis	Sample Size	No. of Events
Phase III design	OS	2.5	90	9- *v* 12-month median OS; HR, 0.75	600	509
Single-arm phase II design	RR	10	90	5% *v* 20% RR	40	NA
Randomized phase II screening design	PFS	10	90	4- *v* 7-month median PFS; HR, 0.57	100	84
Randomized phase II selection design	PFS	50	90	4- *v* 5.5-month median PFS; HR, 0.73	76	65

Abbreviations: HR, hazard ratio; NA, not applicable; OS, overall survival; PFS, progression-free survival; RR, response rate.

where better means better by any amount (no control arm). The selection design has a 50% type I error and thus, unlike the randomized screening design, is not intended to show that one arm is better than another.[13]

Combining Phase II and III Designs Into One Trial

Designs that combine phase II and III functions (ie, phase II/III designs) have separate sets of design parameters that correspond to their phase II and III components. These parameters affect the overall power of the phase II/III trial (the probability of positive phase II and III analyses under their respective alternative hypotheses), which is approximately the product of the individual component powers. (The formal statistical justification of this statement relies on the fact that the correlation of the component treatment-effect estimators tends to be small.) Note that the overall type I error of phase II/III trial is equal to the type I error of its phase III component alone, because it refers to the probability of rejecting the phase III null hypothesis when there is no clinical benefit regardless of the effect the treatment has on the intermediate phase II outcome. Besides the component design parameters, the other important consideration for phase II/III designs is whether to suspend accrual after the phase II component patients are enrolled while their data are maturing.

Phase II/III Trials With Noncomparative Single-Arm Phase II Analyses

For this design, patients are randomly assigned to either the experimental or control treatment. When a sufficient number of patients have been evaluated in the experimental arm, the phase II analysis of these patients is performed. If the phase II analysis is negative, the trial stops; otherwise, the trial continues until it meets its phase III accrual goal.

There is typically little reason to suspend accrual in this design while awaiting the clinical results for the phase II analysis, because the phase II outcome is usually clinical response, which can be evaluated relatively quickly. In addition, the sample size for the phase II analysis is relatively small, and its accrual occurs at the beginning of the trial, when the accrual rate is ramping up. These factors imply that a limited number of extra patients will be accrued if accrual is not suspended.

An example is a trial by Conroy et al[14] to assess whether FOLFIRINOX (fluorouracil, leucovorin, irinotecan, and oxaliplatin) was better than gemcitabine in first-line treatment of metastatic pancreatic cancer. The phase II component involved 88 randomly assigned patients, with targeted response rates of 10% versus 24% for the 44 patients in the experimental arm; the phase III trial targeted an overall survival (OS) hazard ratio (HR; experimental over control) of 0.70 (Table 2). The trial passed its phase II required efficacy cutoff and was a positive phase III trial, with 342 randomly assigned patients demonstrating an HR of 0.57.[14]

Phase II/III Trials With Between-Arm Phase II Analyses

For this design, patients are again randomly assigned to either the experimental or control treatment. The phase II analysis involves a comparison between treatment arms that requires a relatively large sample size and typically uses a time-to-event outcome. Because of this type of outcome, the sample size can be substantially higher if accrual is not suspended while waiting for the phase II outcomes to occur.

An example of this trial design is CALGB-80802 (Cancer and Leukemia Group B 80802), which is assessing whether doxorubicin plus sorafenib is superior to sorafenib alone in treating advanced hepatocellular carcinoma. The phase II component involves an analysis of PFS based on the first 170 patients enrolled and a 6-month

Table 2. Examples of Phase II/III Trials With Design Parameters

Trial	Phase II Analysis				Phase III Analysis			
	End Point	Type I Error (%)*	Alternative	Power (%)	End Point	Type I Error (%)*	Alternative	Power (%)
Conroy et al[14]	RR	5	10% *v* 24% RR	92	OS	5	HR, 0.70	80
CALGB-80802	PFS	15	HR, 0.67	90	OS	2.5	HR, 0.73	90
CALGB-30610	Toxicity	50	NA	NA	OS	2.5	HR, 0.77	81
GOG0182/ICON5[15]	PFS	7.5	HR, 0.75	93	OS	2.5/4†	HR, 0.75	90

Abbreviations: CALGB, Cancer and Leukemia Group B; GOG, Gynecologic Oncology Group; HR, hazard ratio; ICON, International Collaborative Ovarian Neoplasm; NA, not applicable; OS, overall survival; PFS, progression-free survival; RR, response rate.

*One sided.

†Typical type I error divided by four to account for the multiple comparisons of each of the four experimental arms with the control arm (ie, Bonferroni correction).

accrual suspension; it targets a PFS HR of 0.67 with 90% power. The phase III trial targets an OS HR of 0.73 with 480 patients (Table 2).

Phase II/III Trials With Multiple Experimental Arms

In a phase II/III trial with multiple experimental arms, one possibility is that the phase II component is a selection design. As noted earlier, the selection need not involve a measure of efficacy. For example, CALGB-30610 is randomly assigning patients with small-cell lung cancer among three treatment arms. Although all arms will receive chemotherapy, the control arm will receive standard radiotherapy, whereas the two experimental arms will receive increased doses of radiation with two different radiotherapy regimens. The phase II analysis will involve an analysis of toxicity scores calculated for up to 70 patients treated in each of the experimental arms. The experimental arm with higher average toxicity score will be dropped (up to 210 randomly assigned patients), with the random assignment continuing between the other experimental arm and the control arm. With 606 patients randomly assigned to the control and continuing experimental arms, the trial is designed to detect an OS HR of 0.77 (Table 2). At this time, the trial is still randomly assigning patients among the three treatment arms.

Another possibility with a phase II/III trial with multiple experimental arms is to perform a phase II analysis comparing each experimental arm with the control arm. For example, GOG0182/ICON5 (Gynecologic Oncology Group 0182/International Collaborative Ovarian Neoplasm 5) had four experimental chemotherapy arms to be compared with a control chemotherapy arm for advanced-stage ovarian cancer.[15] The phase II part of the trial involved comparing PFS for each experimental arm versus the control arm after 240 PFS events were observed in the control arm, with accrual being capped at 4,000 if the PFS data were not sufficiently mature after 4,000 patients had been accrued. The phase II design targeted a PFS HR of 0.75 for each pairwise comparison. (This corresponds to an observed PFS HR of 0.87 for continuing each experimental arm, although it seems the data monitoring committee also considered a phase II PFS HR cutoff of 0.94.[6]) The phase III analysis of the remaining experimental arms versus the control arm targeted an OS HR of 0.75 (Table 2). Accrual was rapid, and 4,312 patients were accrued with no PFS or OS advantage observed in any of the experimental arms.[15] If this trial had been designed with a 15-month accrual suspension (the projected median PFS expected for the control arm), the total sample size would have been 1,740 instead of 4,312.

When a measure of efficacy is used to decide which phase II trial arms will continue on to phase III, there may be a concern that the type I errors of the phase III analyses are inflated because of the multiple comparisons of experimental arms with the control arm. This same issue arises in multiarm stand-alone phase III trials, in which the need for multiple-comparison adjustment to the type I errors depends on the nature of the experimental treatment arms.[16] In particular, if experimental treatments include different drugs that might have reasonably been tested in separate phase III trials against the control treatment, a multiple-comparison adjustment is not required. On the other hand, if experimental treatments are related (eg, different doses/schedules of the same agent or modifications of a backbone therapy), a multiple-comparison adjustment should be applied. It is also possible to have multiple stages in phase II/III trials, in which the decision to drop experimental arms occurs sequentially as the data accrue.[17]

Phase III Interim Inefficacy Analyses Versus Phase II/III Designs

Almost all phase III trials include prespecified inefficacy/futility interim monitoring rules to stop the trial early if the interim results strongly suggest that the experimental treatment has no benefit over control.[18] A typical inefficacy/futility rule might stop the trial if, at approximately 50% of the required events, the experimental arm is not better than the control treatment.[19] Because inefficacy/futility interim analyses of phase III trials and phase II/III trial designs stop experimental arms that are not doing well, it is important to highlight the difference in the aims of the two strategies. An interim inefficacy/futility analysis in a phase III trial does not require much evidence to keep a trial going, and the trial will typically continue unless the experimental treatment seems worse than the control treatment (ie, continue if the observed HR ≤ 1); this would be expected to happen approximately 50% of the time when the treatments are equally efficacious. In contrast, a phase II analysis in a phase II/III trial requires more evidence that the experimental treatment works better than

Table 3. Features and Operational Characteristics of Typical Phase II, III, and II/III Trial Designs

Characteristic	Phase III*	Phase II Followed by Phase III*†	Phase II/III*†	
			Without Accrual Suspension	With Accrual Suspension
Earliest stopping option‡				
Average sample size	429	100§	156	100§
Average time, months	34.3	14.1	12.5	14.1
Observed HR (cutoff) required for continuing	1	0.756	0.756	0.756
Maximum sample size	600	700‖	600	600
Average sample size	514	151‖	192	145
Average duration, months	45.8	19.3¶	15.9	18.3

NOTE. Under no treatment effect.
*Trial uses design parameters specified in Table 1 and assumes an average accrual rate of 12.5 patients per month. All designs allow for a 2-month delay in collecting follow-up data.
†A phase III inefficacy/futility analysis is also included in these designs.
‡Interim futility look for the phase III design; phase II analysis for phase II followed by phase III and phase II/III designs.
§Sample size is always 100 (not an average).
‖No. includes the 100 patients enrolled in phase II.
¶Time from start of phase II (allowing 6 months to commence phase III after completion of phase II).

the control before continuing on to phase III (eg, observing an HR < 0.756; Table 3); this would be expected to happen approximately 10% of the time when the treatments are equally efficacious.

Some trial designs cannot be easily categorized, because they could be considered phase II/III or phase III with an aggressive inefficacy/futility analysis. For example, consider CALGB-90802, which is assessing whether bevacizumab plus everolimus is better than everolimus for previously treated renal cell carcinoma.[20] The primary end point is OS targeting an HR of 0.77 with 700 patients. After 100 patients have been observed for 4 months, the trial will continue only if the 4-month PFS rate is at least 6% higher in the experimental arm than the control arm; this stopping rule approximately corresponds to continuing the trial only if the observed PFS HR is 0.84 or smaller (an observed increase in median PFS from 4 to 4.8 months). The trial will continue to accrue while waiting for the 4-month assessments. On the one hand, the PFS HR cutoff for continuing is less than 1, suggesting this is a phase II/III design. On the other hand, the cutoff is relatively modest (an improvement corresponding to 0.8 months in median PFS) and is presumably lower than one would have used in a standalone phase II trial (Table 3). This suggests a phase III trial with an aggressive interim inefficacy/futility analysis. (Although it should be noted that interim analyses are most typically based on the primary phase III end point, an intermediate end point as in this trial can sometimes be efficiently used in an inefficacy/futility analysis.[21])

Phase II/III Designs With OS As the Phase II and III End Point

The efficiency gain in using phase II screening is the result of the availability of an intermediate phase II end point, which, if negative, is highly predictive of a negative result for the definitive end point. When there is no reliable intermediate end point for OS (eg, glioblastoma), OS can be used in both the phase II and III components of a phase II/III trial provided that accrual is suspended while awaiting the phase II results. However, the gains over just performing a phase III trial (with interim monitoring) would be considerably reduced. For example, a phase II/III design with an intermediate PFS end point would allow a decision to be made with 145 patients (on average), considerably fewer than the 514 patients (on average) needed in a phase III trial with interim monitoring (Table 3). If one used OS for the phase II analysis (HR, 0.75) and kept the phase II operating characteristics the same, the average sample size would be 420. If one relaxed the type I error of the phase II decision from 10% to 20%, the average sample size would be 360. Another point to consider in using OS in the phase II analysis is that it indirectly reveals interim results about the primary outcome.

DISCUSSION

Phase II/III trials are an adaptive approach to decreasing the time and numbers of patients required in moving an experimental treatment from its development to a definitive assessment of its benefit (Table 3). This is true if the phase II component is used to provide preliminary evidence of the efficacy of the experimental treatment (when such evidence does not exist) or to select among experimental treatments to move forward when such evidence already exists. The overall power of the phase II/III design is approximately the product of the powers of its individual components. For example, if the powers of the phase II and III analyses were both 90% (80%), the overall power of the trial could be as low as 81% (64%). One can increase the overall power by lowering the phase II cutoff (the magnitude of the observed improvement of the experimental treatment over the control treatment required for continuing to phase III). This, however, comes at the cost of more often going on to the full phase III sample size with negative phase III results. (If the phase II cutoff is lowered enough, the design

Table 4. Pros and Cons of Various Trial Design Strategies

Trial Design	Pros	Cons
Phase II followed by phase III, if phase II positive (phase II → III)	Only a 10% chance of accruing to the full phase III sample size when the experimental treatment is ineffective	Lengthiest trial design strategy for definitively demonstrating efficacy when the experimental treatment works
	Allows for a complete evaluation of the phase II data for starting a phase III trial, including consideration of other better experimental treatments that may have become available during the phase II accrual and evaluation	When the experimental treatment works, this design strategy is inefficient in that the phase II patients are not used in the phase III analysis
Phase II/III with no accrual suspension after phase II patients accrued	Only a 10% chance of accruing to the full phase III sample size when the experimental treatment is ineffective	Under no treatment effect, No. of accrued patients will be larger than the phase II → III design or the phase II/III design with accrual suspension
	Phase II patients used in phase III analysis	Phase II evaluation will be less thorough than in the phase II → III design or the phase II/III design with accrual suspension because of less complete follow-up for the phase II patients
Phase II/III with accrual suspension after phase II patients accrued	Only a 10% chance of accruing to the full phase III sample size when the experimental treatment is ineffective	When the experimental treatment works, it will take longer to get definitive results than with the phase II/III trial design with no accrual suspension
	Phase II patients used in phase III analysis	Possible delay in bringing accrual rates up to presuspenstion levels
Phase III (with interim inefficacy/futility monitoring)	Quickest trial design for definitively demonstrating efficacy when the experimental treatment works	Approximately 50% chance of accruing to the full sample size when the experimental treatment is ineffective

essentially becomes a phase III design with an inefficacy/futility interim analysis). Alternatively, the overall power can be increased by increasing the sample size of the phase II or III components. Instead, we recommend setting the power for both the phase II and III components at 90% and using the same type I errors and target alternatives that would have been used if the trials were performed separately. The power of this phase II/III approach is at least 81%. (This is approximately the same power that would be achieved by the sequential use of stand-alone phase II and III trials, each powered at 90%).

There are tradeoffs for suspending trial accrual while evaluating phase II data (Table 4). There can be a delay in bringing accrual rates up to presuspension levels after an accrual suspension. However, such a delay would presumably be considerably shorter than the time required to mount a separate phase III trial after a phase II trial and could be surmountable if investigators better appreciated the rationale for the trial design. We recommend suspending accrual while awaiting phase II results, especially when accrual is rapid, to avoid accruing many extra patients to a potentially negative phase II trial. In addition, an accrual suspension provides a minimum follow-up for the phase II patients, which may provide a more complete PFS evaluation (eg, when PFS curves first separate and then come back together). In the infrequent situations in which single-arm response rate analyses are acceptable, phase II data will mature quickly, and suspension of accrual is unnecessary.

The phase II/III design does have some disadvantages as compared with sequentially performing a phase III trial after a phase II trial (Table 4). With a phase II/III trial, one is committing to the possibility of a phase III trial, so the phase III infrastructure must be planned for at the beginning of the trial, which may result in a delay.[22] In addition, information about dosing/scheduling, supportive care, accrual difficulties, and follow-up issues that might be acquired from a phase II experience may be difficult to utilize in a phase II/III trial.[23] Finally, in some situations, there will be multiple treatments being tested in separate stand-alone phase II trials. This allows a sponsoring organization to pick the most promising treatment(s) to move forward into a phase III trial, whereas a phase II/III trial locks one into testing a particular treatment (unless one uses multiple experimental treatments in the design). In fact, if a novel promising agent becomes available during phase II/III testing of other treatments, it may be difficult to stop an ongoing trial, whereas this is not a problem if the phase II trials are performed separately. In many applications, however, the efficiency advantages of the phase II/III design will outweigh its disadvantages. As randomized phase II screening trials gain in popularity, one should consider the pros and cons of whether such trials would be better designed as phase II/III trials.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Thall PF, Simon R, Ellenberg SS: Two-stage selection and testing designs for comparative clinical trials. Biometrika 75:303-310, 1988

2. Schaid DJ, Wieand S, Therneau TM: Optimal two-stage screening designs for survival comparisons. Biometrika 77:507-513, 1990

3. Scher HI, Heller G: Picking the winners in a sea of plenty. Clin Cancer Res 8:400-404, 2002

4. Bretz F, Schmidli H, König F, et al: Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: General concepts. Biom J 48:623-634, 2006

5. Maca J, Bhattacharya S, Dragalin V, et al: Adaptive seamless phase II/III designs: Background, operational aspects, and examples. Drug Inf J 40:463-473, 2006

6. Parmar MK, Barthel FM, Sydes M, et al: Speeding up the evaluation of new agents in cancer. J Natl Cancer Inst 100:1204-1214, 2008

7. Hunsberger S, Zhao Y, Simon R: A comparison of phase II study strategies. Clin Cancer Res 15:5950-5955, 2009

8. Stallard N, Todd S: Seamless phase II/III designs. Stat Methods Med Res [epub ahead of print on August 19, 2010]

9. Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10:1-10, 1989

10. Korn EL, Arbuck SG, Pluda JM, et al: Clinical trial designs for cytostatic agents: Are new approaches needed? J Clin Oncol 19:265-272, 2001

11. Rubinstein LV, Korn EL, Freidlin B, et al: Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 23:7199-7206, 2005

12. Korn EL, Liu PY, Lee SJ, et al: Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials. J Clin Oncol 26:527-534, 2008

13. Simon R, Wittes RE, Ellenberg SS: Randomized phase II clinical trials. Cancer Treat Rep 69: 1375-1381, 1985

14. Conroy T, Desseigne F, Ychou M, et al: FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. N Engl J Med 364:1817-1825, 2011

15. Bookman MA, Brady MF, McGuire WP, et al: Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: A phase III trial of the Gynecologic Cancer InterGroup. J Clin Oncol 27:1419-1425, 2009

16. Freidlin B, Korn EL, Gray R, et al: Multi-arm clinical trials of new agents: Some design considerations. Clin Cancer Res 14:4368-4371, 2008

17. Royston P, Barthel FM, Parmar MK, et al: Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. Trials 12:81, 2011

18. Freidlin B, Korn EL: Monitoring for lack of benefit: A critical component of a randomized clinical trial. J Clin Oncol 27:629-633, 2009

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

20. Stadler WM, Phillips G, George DJ, et al: Bevacizumab and everolimus in renal cancer: A rational way forward. J Clin Oncol 33:e692-e693, 2010

21. Goldman B, Leblanc M, Crowley J: Interim futility analysis with intermediate endpoints. Clin Trials 5:14-22, 2008

22. Green S: Pitfalls in oncology clinical trial designs and analysis, in Kelly WK, Halabi S (eds): Oncology Clinical Trials. New York, NY, Demos Medical Publishing, 2010, pp 197-214

23. Emerson SS, Fleming TR: Adaptive methods: Telling "the rest of the story." J Biopharm Stat 20:1150-1165, 2010