VOLUME 23 · NUMBER 28 · OCTOBER 1 2005

JOURNAL OF CLINICAL ONCOLOGY

SPECIAL ARTICLE

Design Issues of Randomized Phase II Trials and a Proposal for Phase II Screening Trials

Lawrence V. Rubinstein, Edward L. Korn, Boris Freidlin, Sally Hunsberger, S. Percy Ivy, and Malcolm A. Smith

From the Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, MD.

Submitted September 1, 2004; accepted June 9, 2005.

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

Address reprint requests to Lawrence V. Rubinstein, PhD, Room 8130, 6130 Executive Plaza, Rockville, MD 20852; e-mail: rubinsteinl@ctep.nci.nih.gov.

0732-183X/05/2328-7199/$20.00

DOI: 10.1200/JCO.2005.01.149

ABSTRACT

Future progress in improving cancer therapy can be expedited by better prioritization of new treatments for phase III evaluation. Historically, phase II trials have been key components in the prioritization process. There has been a long-standing interest in using phase II trials with randomization against a standard-treatment control arm or an additional experimental arm to provide greater assurance than afforded by comparison to historic controls that the new agent or regimen is promising and warrants further evaluation. Relevant trial designs that have been developed and utilized include phase II selection designs, randomized phase II designs that include a reference standard-treatment control arm, and phase II/III designs. We present our own explorations into the possibilities of developing "phase II screening trials," in which preliminary and nondefinitive randomized comparisons of experimental regimens to standard treatments are made (preferably using an intermediate end point) by carefully adjusting the false-positive error rates (α or type I error) and false-negative error rates (β or type II error), so that the targeted treatment benefit may be appropriate while the sample size remains restricted. If the ability to conduct a definitive phase III trial can be protected, and if investigators feel that by judicious choice of false-positive probability and false-negative probability and magnitude of targeted treatment effect they can appropriately balance the conflicting demands of screening out useless regimens versus reliably detecting useful ones, the phase II screening trial design may be appropriate to apply.

J Clin Oncol 23:7199-7206.

INTRODUCTION

The large numbers of new anticancer agents under development has challenged clinical investigators seeking to design and implement clinical trials to improve outcome for persons with cancer. Many of these new agents are molecularly targeted and have distinctive toxicity profiles compared with conventional cytotoxic agents. Given the often non-overlapping toxicity profiles, it is frequently possible to combine these new agents with standard chemotherapy regimens in attempts to develop more effective treatments. Once the tolerability of such combinations has been demonstrated, an obvious clinical trial design for evaluating the contribution of the new agent is a phase III comparison of standard therapy to standard therapy plus the new agent. The straightforward solution of directly proceeding to phase III trials without phase II data might be reasonable if there were limited numbers of new agents available for evaluation and if there were unlimited resources with which to conduct these evaluations. However, resources are constrained and there are a multitude of new agents available for clinical evaluation, including antiangiogenic agents, growth factor receptor inhibitors, potentiators of apoptosis, modulators of gene expression, and modulators of signal transduction pathways. Recent disappointing clinical trial results for regimens proceeding directly from phase I to phase III testing further dampen enthusiasm for initiating large clinical trials of new agents when there are limited clinical data to support their potential for benefit.[1,2]

How should investigators prioritize from among the many options available the ones that most warrant evaluation in definitive phase III trials? If an agent's potential for clinical benefit is thought to be related to its cytotoxic activity as a single agent, then response data from conventional phase II trials should be useful for prioritization. Molecularly targeted agents may have substantial activity as single agents, as illustrated by the single-agent activity of imatinib for chronic myeloid leukemia and gastrointestinal stromal tumor, and by the single-agent activity of the epidermal growth factor–receptor inhibitor gefitinib against non–small-cell lung cancer with mutations in the epidermal growth factor receptor.[3-6] However, for other molecularly targeted agents, there may be little documented single-agent activity, and yet there may be reason to think that the agent could improve outcome when combined with standard chemotherapy. In this review of issues related to randomized phase II trial design, our particular focus is on trial designs applicable to the latter situation.

For two decades, there has been interest in utilizing phase II trials with randomization against a standard-treatment control arm to provide greater assurance than afforded by comparison to historical controls that the new agent or regimen is "promising" in comparison with what is currently available.[7,8] As noted by European Organisation for Research and Treatment of Cancer (EORTC) researchers, "there is much to recommend randomized comparison with standard treatment, even at the initial phase II stage" because this approach will "offer a protection against possible selection bias" as an explanation for promising results observed in the phase II setting.[9,10] Moreover, there are situations in which relevant clinical trials data from historical controls are not available. This may be particularly true for trials in which disease progression, rather than tumor response, is felt to be the most appropriate end point for a preliminary evaluation of efficacy. However, the inclusion of a randomized comparison control more than doubles the number of patients required to achieve comparable false-positive probability (α or type I error) or false-negative probability (β or type II error) for detection of a given improvement in outcome. Thus, despite the obvious attraction of including a randomized comparison population within phase II trials, this advantage is counterbalanced by the larger number of patients required to detect comparable differences in outcome.

In a recent editorial in the *Journal of Clinical Oncology*, Wieand[11] succinctly and cogently outlines attractions of the randomized phase II trial, as well as limitations of current designs. We review previously discussed randomized phase II study designs, specifically examining phase II selection designs, randomized phase II designs that include a reference standard-treatment control arm, and phase II/III designs. Finally, we present our own explorations into the possibilities of designing "phase II screening trials." These trials are to be used when evaluation of a new agent or regimen can best be made by preliminary and nondefinitive randomized comparison to a standard-treatment control, carefully adjusting the false-positive and false-negative error rates so that the targeted treatment benefit may be appropriate while the sample size remains restricted.

RANDOMIZED PHASE II SELECTION DESIGNS

Simon et al[7] introduced and explored the uses and characteristics of a randomized phase II selection design in which patients are randomized to two or more experimental agents or regimens, and the regimen that results in the highest observed response rate is selected for further study. Sample sizes are given that assure 90% probability to select the best study arm, so long as the true expected response rate exceeds that of any other arm by at least 15% (in absolute terms; eg, 35% *v* 20%). Appropriate uses of this design include selection among new agents administered singly as well as among new combination regimens, especially if the regimens all have a common core regimen to which various new agents are added. This design could also be used to select among different doses or schedules of the same agent, assuming that these different doses or schedules had roughly similar degrees of toxicity. If there were substantial differences in toxicity among the different doses or schedules, then a design could be considered in which the more toxic treatment required some minimum improvement to be selected.[12,13] In each of these situations, the experimental arm selected in the trial could then be subjected to a definitive phase III evaluation against the standard regimen.

When appropriate historical controls are available, it has become common to structure each individual experimental arm within the randomized phase II selection trial as if it were, by itself, a one-armed two-stage phase II study.[14,15] In this way, either arm may be terminated early because of discouraging results, and the response rate of each arm can be assessed separately against a historical control rate with definable α and β error probabilities. The control response rate may be treated as a precise value or as an estimate with inherent variability based on outcomes from historical control populations.[16] Phase II selection designs may also be adapted for use with progression-free survival or overall survival end points.[17] It is not appropriate to apply the phase II selection design to randomized comparisons of experimental agents or regimens with standard-treatment control arms. The selection design will choose the experimental treatment over the standard treatment with roughly 50% probability, even when expected outcome for the two treatments is the same, and "impressive" evidence of effectiveness may be observed even when there is no true treatment effect.[18] The selection design is appropriate for prioritizing between two experimental regimens when there is no a priori reason (eg, significant differences in toxicity or cost) to prefer one treatment over the other.

RANDOMIZED PHASE II DESIGNS INCLUDING A REFERENCE STANDARD-TREATMENT CONTROL ARM

Herson and Carter[19] proposed including a simultaneously randomized standard-treatment control arm in their trial. The EORTC has advocated use of this design and applied it in selected phase II trials.[9,20] In this design, the standard-treatment control arm is not directly compared with the experimental arms, due to the statistical constraints resulting from the small sample sizes. The standard-treatment control arm acts as a check for whether the historical control patients, against which the experimental arms are being judged, are comparable to the patients entering the phase II trial. If the standard-treatment control arm does substantially worse than expected, failure of an experimental arm to improve on historically based standards does not necessarily imply an actual lack of benefit. Conversely, if the standard-treatment control arm does significantly better than expected, apparent improvement for an experimental arm is called into question. The authors' response to either of these situations is that the trial be repeated. It is not clear why a follow-up trial would yield a more "representative" standard-treatment control and result in a more satisfactory outcome among the standard-treatment control patients.

PHASE II/III STUDY DESIGNS

Schaid et al[21] proposed embedding a randomized phase II study within a phase III study as follows. New patients would be randomly assigned to a control arm (C), an experimental arm (E) that is ready to be phase III tested, or additional arms ($A_1, \ldots, A_K$) for which phase II results are desired. Patients who progress on one of the arms ($A_1, \ldots, A_K$) would then be randomly assigned to arms E or C, and their survival data would be combined in a stratified fashion with data from patients directly assigned to arms C or E for the phase III assessment of arm E versus arm C. This design allows untreated patients to be enrolled onto the phase II study, while assuring them a more established therapy if they progress. It retains the advantages of the selection design for comparison of the activities of the phase II agents being studied, while providing a reference standard-treatment control arm at no additional cost in terms of the number of patients enrolled. However, with respect to actual use of the reference standard-treatment control arm, it shares the difficulties of the design by Herson and Carter[19] discussed in the Randomized Phase II Designs Including a Reference Standard-Treatment Control Arm section, namely, that the precise statistical use of the reference arm is unclear and problematic.[19] In addition, the appropriateness of the phase III patient population for phase II testing must be assured,[22] and the potential impact of the phase II therapy on the relative treatment effect of the regimens being tested in the phase III component must be considered.[21]

Storer[23] proposed a related design in which a randomized phase II study of an experimental regimen versus a standard-treatment control comprises the initial component of a phase III trial of the same regimens. At the completion of the phase II component, the experimental regimen would be evaluated against historical standards, as in standard one-armed phase II trials, to determine whether the trial would be completed as a phase III study or terminated as a negative phase II study. The results from the standard-treatment control arm would be ignored in making this "phase II" assessment. This approach has the same problems cited previously for standard phase II trials with comparisons to historical controls, and is efficient only if the phase II trial is expected to be positive with a reasonably high likelihood, because otherwise the extensive effort expended in developing the clinical trial as a phase III evaluation would often be wasted because of early termination during the phase II component.

Ellenberg and Eisenberger[24] proposed a phase III design in which the initial component would serve as a randomized phase II study, with the experimental regimen compared directly with the standard-treatment control. Their decision rule was that the phase III study would continue to completion as long as the response rate for the experimental arm was the higher of the two. They designed the initial (randomized phase II) component to be just large enough so that the probability of mistakenly terminating the study when the true response rate for the experimental arm was better than the standard-treatment arm by the targeted difference Δ was no more than .05. Thus, the loss of power for the phase III comparison would be low. The sample size for the initial component (approximately one third of the total phase III sample size) would roughly be double the size for a single arm phase II trial with the same targeted difference (and α and β equal to .1). A significant disadvantage of this approach is that the probability of mistakenly continuing the study into its phase III component when the true response rates for the experimental and standard-control arm are equal is approximately .5, limiting the utility of this approach as a screening tool for prioritizing treatments for phase III evaluation.

A similar phase II/III clinical trial design was proposed by Schaid et al,[25] and was elaborated on by Scher and Heller[26] for the specific setting of prostate cancer clinical trials. This design can accommodate multiple experimental regimens and uses survival as the primary end point for determining whether to proceed from the phase II to the phase III component. Like the proposal of Ellenberg and Eisenberger,[24] the goal of the first or screening phase of the study is to determine whether any of the experimental treatments show sufficient activity to warrant continuation into the confirmatory phase III stage. At the end of the first phase, a log-rank test statistic is computed for the pairwise survival time comparisons between each experimental regimen and the standard-therapy regimen. Those experimental regimens with a pairwise log-rank statistic exceeding a prespecified minimum value move forward to the phase III component, while the remaining regimens are not studied further. The sample size for the first stage is determined by a numerical optimization program designed to produce the smallest total sample size, under the hypothesis that the survival distributions for all tested treatments are identical. In the example provided by Schaid et al,[25] and in one clinical application of this design involving one experimental regimen, the sample size for the phase II component was one half of the total study accrual.[27]

The phase II/III designs of Ellenberg and Eisenbeberger and of Schaid,[24,25] as well as comparable designs using Bayesian methods,[28] offer some advantages over the alternative of separate phase II and phase III studies.[10,25] These designs allow phase II patient data to be used in the principal phase III trial analysis, and they minimize the delay between the completion of the phase II study and the start-up of the phase III study. They offer the flexibility of using either response rate or overall survival as measures of phase II success, and their use of a concurrent randomized control increases the validity of the phase II comparisons. However, the utility of phase II/III designs as screening tools is limited. The phase II component is sizable in comparison with that of conventional phase II studies and with the total study accrual, and development of clinical trials using these designs requires the infrastructure and commitment of the phase III component. However, because of the need to maintain power for the final phase III comparison, the phase II component may require a high α leading to an elevated rate (compared with standard phase II designs) of incorrectly proceeding to the phase III component for experimental arms that in reality are no more effective than the control arm. In this way, these phase II/III designs may best be considered as phase III trials with aggressive interim monitoring.

A practical limitation of studies using the phase II/III study design is that the end point for determining whether to proceed from the phase II to the phase III component should be ascertainable relatively quickly after treatment initiation. If this is not the case, then the study may need to close temporarily while awaiting sufficient numbers of events to occur among patients enrolled onto the phase II component.

RANDOMIZED PHASE II SCREENING DESIGNS

We now explore the possibilities of conducting a direct, but nondefinitive, "screening" comparison of the experimental regimen against a randomized standard-treatment control arm, within a trial with a moderate sample size. We build on descriptions of similar trial designs,[29,30] while providing additional discussion of design parameters appropriate for phase II screening trials.

Assume that we want to design a randomized phase II screening trial that will allow us to assess whether experimental regimen E is more promising than standard-treatment control regimen C, with respect to progression-free survival (PFS). We randomly assign patients to regimen E versus regimen C because we believe that comparing the PFS of regimen E with that of historical standard-treatment controls is prone to various sorts of bias. We wish to keep the two-armed trial to a sample size of approximately 50 to 100 patients. There are three parameters to vary. We allow α (the probability of concluding regimen E is superior when it actually offers no benefit) to be either 10% (the standard for phase II trials) or 20%. We allow β (the probability of a false-negative result, for the case of use of regimen E increasing median PFS by the target value Δ) to also be either 10% (the standard for phase II) or 20%. Note that the "power" of the screening trials is 1-β (ie, 90% or 80% for the values of β used). We allow the target multiplier value Δ to vary from a relatively modest median PFS ratio of 1.3 to a more optimistic ratio of 1.75. Table 1 lists the range of approximate required numbers of total observed failures because these three parameters are allowed to vary. Calculations are based on the assumption of exponentially distributed event times, but apply more generally to event times that satisfy the assumption of proportional hazards between treatment arms.[32] The relationship between the number of observed treatment failures and the number of enrolled patients depends on the event rate and follow-up time. If general screening is done in advanced disease with adequate follow-up, the number of patients is only slightly greater than the number of treatment failures. For other cases, the reader may approximate the number of patients needed if the average probability of treatment failure for a patient observed from randomization to trial end is known. Thus, one table encompasses a wide array of trial designs.

Table 1. Approximate Required Numbers of Observed (Total) Treatment Failures for Screening Trials With Progression-Free Survival End Points

	Hazard Ratios (Δ)			
Error Rates	Δ = 1.3	Δ = 1.4	Δ = 1.5	Δ = 1.75
(α, β) = (10%, 10%)	382	232	160	84
(α, β) = (10%, 20%) or (20%, 10%)	262	159	110	58
(α, β) = (20%, 20%)	165	100	69	36

NOTE. Calculations were carried out using nQueryAdvisor 5.0 software (Statistical Solutions, Saugus, MA), based on methods given in Collett[31] with one-sided α.

It is not possible, as listed in Table 1, to target $\Delta < 1.75$ while keeping the required number of patients at approximately 100 patients or fewer, and maintaining the standard phase II design values of α and β equal to 10%. However, if we are willing to let α or β be 20%, then a target of $\Delta = 1.5$ is attainable with approximately 100 patients, as is a target of $\Delta = 1.4$, if we let both α and β be 20%. A target of $\Delta = 1.3$ is attainable only if we let α or β be larger than 20%, which we consider inadvisable.

Table 2 lists the required sample sizes for phase II selection trials for a range of parameter values when the response rates associated with the standard-treatment control and experimental arms (P_1 and P_2, respectively) are the comparative end points. We vary α and β as before, and we let $P_1 = 20\%$ or 40%, defining $P_2 = P_1 + 15\%$ or $P_2 = P_1 + 20\%$. A 20% difference in response rates between the standard-treatment control and experimental arms cannot be detected with fewer than 100 patients using the standard α or standard β equal to 10%. However, if we are willing to let both α and β be 20%, then we can detect a 20% difference in absolute terms between the response rates, with fewer than 100 patients. We can detect a 15% difference between the rates only if we let α or β be larger than 20%, which we consider inadvisable. We note that in Table 2, as well as in Table 1, that letting $(\alpha, \beta) = (15\%, 15\%)$ yields almost identical sample size requirements as letting $(\alpha, \beta) = (10\%, 20\%)$ (calculations not shown).

A third possible end point for phase II screening trials is progression-free rate as a binary outcome at a specified time from treatment initiation. This end point may be preferred over PFS (using actual progression times) because phase II randomized trials are not usually blinded and there is a potential for bias in determining progression times because of a tendency for more frequent response assessments for case control patients. The potential for bias may be especially great when patients progressing on the control arm are crossed over to the experimental arm. Using progression-free rate at a predetermined time from treatment initiation as the primary end point (eg, all patients evaluated for progression at 4 months) minimizes the chance for bias related to earlier outcome determinations for patients on the control arm. The same methods used for determining the number of patients required for phase II screening studies, with response rate as the primary end point, can be applied to studies with progression-free rate at a specified time as the primary end point.

In designing a phase II screening trial, choice of the above parameter values must be made with great care. An overly large α reduces the screening ability of the study, as the rate of false-positives essentially nullifies the screening effect. If this ability is reduced too much, there is little value in conducting the phase II screening trial at all—one is, in effect, moving directly from phase I to phase III. In contrast, an overly large β runs the risk of terminating the study of a potentially useful regimen. Likewise, an overly optimistic Δ runs the risk of rejecting a regimen with a more limited, but still clinically significant, benefit. Clearly, balancing among these conflicting demands requires a level of compromise not generally necessary for the single-arm phase II trial, for which it is practical to require that α and β be no more than .1. We suggest α and β equal .20 and Δ equal 1.5 (or a target difference in response rate of 20%) as appropriate design parameters for consideration in phase II screening trials.

The overall operating characteristics of the phase II screening design, as applied to the development of a particular drug or a particular tumor type, depend on a number of factors. For example, if different end points are used for the screening trial(s) compared with the eventual phase III trial(s), then the overall sensitivity and specificity for the screening test in predicting for a positive phase III trial will depend on the relationship of these two end points. For example, if the screening trial uses a PFS end point and the phase III trial uses an overall survival end point, and if the PFS differences between the regimens are larger than the survival differences, then the sensitivity of the screening design would be increased. The number of screening trials conducted for a specific agent or tumor type also affects the operating characteristics of the screening design. Conducting a series of phase II screening trials of a particular agent against several tumor types will increase the likelihood of observing a false-positive result. For example, if investigators conducted three screening tests with $\alpha = \beta = .20$, then the probability of observing a false-positive result when the agent was truly inactive against all of the tumors tested would be approximately 50%. Thus, even if the agent added nothing to the standard therapy for the tumor types tested, the screening design would support proceeding

Table 2. Approximate Required Numbers of Total Patients for Screening Trials With Response Rate End Points

	Response Rates			
Error Rate	20% *v* 35%	20% *v* 40%	40% *v* 55%	40% *v* 60%
$(\alpha, \beta) = (10\%, 10\%)$	256	156	316	182
$(\alpha, \beta) = (10\%, 20\%)$ or $(20\%, 10\%)$	184	112	224	132
$(\alpha, \beta) = (20\%, 20\%)$	126	78	150	90

NOTE. Calculations were carried out using nQueryAdvisor 5.0 software (XXX, XXX), based on methods given in Fleiss et al[33] with one-sided α.

with further clinical evaluation of the agent in some setting about one half of the time. However, this could still represent an advantage over conducting phase III trials against all of the tumor types. Conversely, if the agent were truly active against all three tumor types, then the likelihood of dropping the agent for inactivity after conducting three screening trials would be less than 5%, despite $\beta = .20$ for each individual trial.

One potential outcome from a phase II screening trial is that substantial evidence for a treatment effect for the experimental arm will be observed and researchers will be tempted to view the trial results as conclusive, especially if the trial uses a definitive clinical end point like survival. A *P* value of .05 is not sufficient because, as discussed previously, the conduct of multiple phase II screening trials increases the overall false-positive rate for declaring an experimental treatment effective beyond the nominal *P* value observed for an individual trial. Moreover, in order to mount a phase III trial, sufficient preliminary evidence of clinical activity is typically required, whereas this evidence would not be required to initiate a phase II screening trial. In order to give a reliable conclusion, a phase II screening trial needs to provide a similar amount of evidence as would be expected from a standard phase II-phase III developmental sequence. We propose that the results from phase II screening trials be viewed in a manner similar to the interim results from phase III trials, and that a reasonable rule for considering results from these trials as convincing would be $P < .005$.[34] If this level of evidence favoring a beneficial effect for the experimental treatment is absent, a phase III study should be pursued in order to reliably define the treatment's contribution to the therapy of the cancer under study. In addition, researchers applying the phase II screening design should appreciate that due to the relatively small sample size of studies using this design, a significant result will correspond to a large estimate of treatment effect with wide CIs. Further clinical evaluation of the experimental regimen in the target population would be required to provide more precise estimates of outcome.

DISCUSSION

Future progress in improving cancer therapy can be expedited by better prioritization of new treatments for phase III evaluation. The phase II screening trial is presented as a possible approach to improve this prioritization process in an era in which there are multiple new agents available for study. Phase III trials, which provide definitive evidence for the benefit to patients of new treatment approaches, are time consuming and expensive. In the 1980s, when there was a relative paucity of new anticancer agents, a major concern was minimizing the false-negative rates of phase II designs.[35] However, as multiple research opportunities increased, the cost of false-positive results also increased. An adverse impact of false-positive results in phase II trials is the conduct of phase III trials that produce negative results. Negative phase III trials, though providing valuable information, represent a lost opportunity to improve outcome and may be considered a failure of the prioritization process. Next, we consider two drug development scenarios and describe how randomized phase II studies might be applied for prioritization purposes. In each of the scenarios, we assume that the new agents in question are unlikely to induce objective responses as single agents and/or that the agents are likely to modulate the activity of the standard therapy with which they are combined; implying that single-agent phase II data will be of little use in prioritization, and data from combination studies could be valuable.

First, consider the situation of investigators who are responsible for developing clinical trials for a specific type of cancer. There is a standard therapy (regimen A) for this cancer, and there are several drugs of potential utility (drugs x, y, and z). There is interest in eventually conducting a trial comparing regimen A to regimen A plus one of the new agents. The ideal design in this situation would be to conduct a randomized phase II selection design as described by Simon et al,[7] comparing regimens A + x, A + y, and A + z.[7] Applicability of the selection design implies that the regimens are essentially equitoxic, and that there is no a priori preference for one of the regimens over any of the others. The selection design would assure that the arm selected for a phase III comparison was unlikely to be substantially inferior to any of the alternative regimens. Although a selection design would be the optimal approach when several new agents are available for combination with a standard regimen, there are pragmatic factors that might limit its application. If the new drugs are not yet commercially available and are being developed by different pharmaceutical companies, pharmaceutical sponsors may be reticent to provide their agents for a selection design comparison. Also, the three new drugs may be at different stages of development and all may not be ready at the same time for the selection design phase II trial.

If the phase II selection design could not be applied, then a series of conventional phase II trials of regimen A plus each of the new drugs could be conducted. However, conventional phase II studies of drug combinations that include one or more active agents often provide little more than toxicity/feasibility data, because patient selection and small sample size severely limit assessments of the level of activity in comparison with the level of activity anticipated for the agents administered individually.[7] To overcome this problem, the phase II screening trial design could be considered, with each trial comparing regimen A to regimen A plus one of the new drugs. A primary factor in considering whether this approach is appropriate would be whether conducting the screening trial would limit the ability to conduct a definitive phase III comparison. Phase II trials often use different end points than phase III trials (objective response and disease progression for phase II and survival for phase III), which may minimize the impact of a positive phase II screening trial on the ability to conduct a subsequent phase III trial. The investigators would need to appreciate the relatively high false-positive rate when a series of screening tests are conducted. Another option would be to apply one of the phase II/III designs previously described.[24,25] However, these designs lack much of the potential benefit of a "screening test," because for each agent tested there must be the commitment to proceed to a phase III trial.

A second drug development scenario is that of a pharmaceutical sponsor attempting to decide whether to evaluate its new agent as therapy for several different types of cancer, with each diagnosis having its own standard therapy. The sponsor's eventual goal would be to demonstrate the utility of its agent in combination with standard therapy for one or more of these cancers. As noted previously, the approach of conducting conventional phase II studies of the new agent with each of the active standard regimens provides useful toxicity data, but provides little useful information concerning the contribution of the new agent to the standard regimen. The phase II screening design could be applied to test the agent plus standard chemotherapy for each of the tumor types. If the agent were active against more than one tumor type, the screening design would have a high probability of identifying this activity for at least one of the tumor types, and hence provide evidence to support further development of the agent. Conversely, the false-positive rate from a series of three or more phase II screening trials would approach or exceed 50%. However, this would still allow an early decision to limit the development of inactive agents in a substantial percentage of scenarios, while at the same time identifying target tumor types for the agent when activity truly exists.

There are specific situations in which the phase II screening design is either unnecessary or inappropriate. For those patient populations for which outcome with standard therapy is very well defined, conventional phase II clinical trials with design parameters based on the known response/outcome for standard therapy can be used for screening new treatments. For new agents that demonstrate substantial activity as single agents, and that can be safely combined with standard therapy, the single agent activity will likely provide sufficient rationale for proceeding to a phase III evaluation. When the comparisons of interest are between experimental regimens rather than with a standard therapy, selection phase II designs, rather than the phase II screening design, are appropriate. The most important caveat in using the phase II screening design is that it may compromise the ability to conduct definitive phase III trials. The screening design should not be applied unless investigators can be reasonably certain that a positive result in their small study (excepting an extreme result, as discussed in Randomized Phase II Screening Designs) will not be appreciated as definitive and will not preclude conduct of a definitive phase III test of the experimental regimen. Although phase III trials for adults with cancer often have survival as their primary end point, phase II trials commonly use end points such as response rate or disease progression that can be ascertained more quickly than survival. The use of nonsurvival end points in phase II screening trials might make it easier to conduct definitive phase III trials that use survival end points.

CONCLUSION

If the ability to conduct a definitive phase III trial can be protected, and if investigators feel that by judicious choice of false-positive probability (α or type I error) and false-negative probability (β or type II error) and magnitude of targeted treatment effect that they can appropriately balance the conflicting demands of screening out useless regimens while detecting useful ones, it may be appropriate to apply the phase screening trial design. The screening design may be especially useful as a preliminary test of regimens in which a new agent is added to a standard regimen.

Authors' Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Moore MJ, Hamm J, Dancey J, et al: Comparison of gemcitabine versus the matrix metalloproteinase inhibitor BAY 12-9566 in patients with advanced or metastatic adenocarcinoma of the pancreas: A phase III trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol 21:3296-3302, 2003

2. Van Cutsem E, van de Velde H, Karasek P, et al: Phase III trial of gemcitabine plus tipifarnib compared with gemcitabine plus placebo in advanced pancreatic cancer. J Clin Oncol 22:1430-1438, 2004

3. Druker BJ, Talpaz M, Resta DJ, et al: Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 344:1031-1037, 2001

4. van Oosterom AT, Judson I, Verweij J, et al: Safety and efficacy of imatinib (STI571) in metastatic gastrointestinal stromal tumours: A phase I study. Lancet 358:1421-1423, 2001

5. Paez JG, Janne PA, Lee JC, et al: EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. Science 304:1497-1500, 2004

6. Lynch TJ, Bell DW, Sordella R, et al: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350:2129-2139, 2004

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

8. Simon R, Thall PF, Ellenberg SS: New designs for the selection of treatments to be tested in randomized clinical trials. Stat Med 13:417-429, 1994

9. European Organisation for Research and Treatment of Cancer: Phase II trials in the EORTC: The Protocol Review Committee, the Data Center, the Research and Treatment Division, and the New Drug Development Office. Eur J Cancer 33:1361-1363, 1997

10. Van Glabbeke M, Steward W, Armand JP: Non-randomised phase II trials of drug combinations: Often meaningless, sometimes misleading—Are there alternative strategies? Eur J Cancer 38:635-638, 2002

11. Wieand HS: Randomized phase II trials: What does randomization gain? J Clin Oncol 23:1794-1795, 2005

12. Sargent DJ, Goldberg RM: A flexible design for multiple armed screening trials. Stat Med 20:1051-1060, 2001

13. Lui KJ: A flexible design for multiple armed screening trials by Daniel J. Sargent and Richard M. Goldberg, Statistics in Medicine 2001. Stat Med 21:625-628, 2002

14. Fleming TR: One-sample multiple testing procedure for phase II clinical trials. Biometrics 38:143-151, 1982

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

16. Thall PF, Simon R: Incorporating historical control data in planning phase II clinical trials. Stat Med 9:215-228, 1990

17. Liu PY, Dahlberg S, Crowley J: Selection designs for pilot studies based on survival. Biometrics 49:391-398, 1993

18. Liu PY, LeBlanc M, Desai M: False positive rates of randomized phase II designs. Control Clin Trials 20:343-352, 1999

19. Herson J, Carter SK: Calibrated phase II clinical trials in oncology. Stat Med 5:441-447, 1986

20. Fossa SD, Mickisch GH, de Mulder PH, et al: Interferon-alpha-2a with or without 13-cis retinoic acid in patients with progressive, measurable metastatic renal cell carcinoma. Cancer 101:533-540, 2004

21. Schaid DJ, Ingle JN, Wieand S, et al: A design for phase II testing of anticancer agents within a phase III clinical trial. Control Clin Trials 9:107-118, 1988

22. Smith MA, Anderson B: Phase II window studies: 10 years of experience and counting. J Pediatr Hematol Oncol 23:334-337, 2001

23. Storer BE: A sequential phase II/III trial for binary outcomes. Stat Med 9:229-235, 1990

24. Ellenberg SS, Eisenberger MA: An efficient design for phase III studies of combination chemotherapies. Cancer Treat Rep 69:1147-1154, 1985

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

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

27. Frasci G, Lorusso V, Panza N, et al: Gemcitabine plus vinorelbine versus vinorelbine alone in elderly patients with advanced non–small-cell lung cancer. J Clin Oncol 18:2529-2536, 2000

28. Inoue LY, Thall PF, Berry DA: Seamlessly expanding a randomized phase II trial to phase III. Biometrics 58:823-831, 2002

29. Simon RM, Steinberg SM, Hamilton M, et al: Clinical trial designs for the early clinical development of therapeutic cancer vaccines. J Clin Oncol 19:1848-1854, 2001

30. 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

31. Collett D: Modeling survival data in medical research. London, United Kingdom, Chapman and Hall, 1994

32. George SL, Desu MM: Planning the size and duration of a clinical trial studying the time to some critical event. J Chronic Dis 27:15-24, 1974

33. Fleiss JL, Tytun A, Ury HK: A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36:343-346, 1980

34. Freidlin B, Korn EL, George SL: Data monitoring committees and interim monitoring guidelines. Control Clin Trials 20:395-407, 1999

35. Carter SK: Clinical aspects in the design and conduct of phase II trials,. in Buyse ME, Staquet MJ, Sylvester RJ (eds): Cancer Clinical Trials Methods and Practice. New York, NY, Oxford University Press, 1984