

Phase II Clinical Trial Design: Methods in Translational Research from the Genitourinary Committee at the Eastern Cooperative Oncology Group

Robert Gray,¹ Judith Manola,² Scott Saxman,⁴ John Wright,⁴ Jan Dutcher,⁵ Michael Atkins,³ Michael Carducci,⁶ William See,⁸ Christopher Sweeney,⁹ Glenn Liu,¹⁰ Mark Stein,⁷ Robert Dreicer,¹¹ George Wilding,¹⁰ and Robert S. DiPaola⁷

Abstract Given the increase in novel agents and difficulty with planning and completing many phase III studies, various phase II trial design options should be considered to more effectively guide phase III trial plans. The need for novel phase II trial designs has increased, given the number of novel molecular targeted therapies now available for testing, an abundance of cytostatic agents, and limited resources to conduct phase III studies for all interesting agents or combinations. This review will focus on options for phase II trial designs. We review randomized phase II designs with placebo control, randomized selection designs, and randomized discontinuation designs. As agents become available for testing in the clinic, the strengths and weaknesses of different phase II trial designs should be considered to optimize a trial development plan that guides phase III trial decisions more effectively.

The goal of the National Cancer Institute to eliminate the death and suffering due to cancer can only be accomplished through effective strategies to test the best science in clinical trials and to adapt these strategies to a complex and changing time. The need for improved and more efficient clinical trial design continues to increase, given the development of potentially cytostatic agents and the need to select from many novel targeted agents and combinations.

The challenge for phase II trial design currently includes the need to identify drugs with sufficient activity for phase III testing. However, the number of novel molecular targeted therapies is increasing, and although phase III studies are critical to molecular targeted therapy testing, the study of all molecular targeted therapies in phase III studies is difficult. Additionally, although trial designs with single agents or molecular targeted therapies in combination with chemotherapy are likely candidates for experimental arms on phase III studies, greater effect may be obtained by targeting multiple pathways. Therefore, the problem is not just determining if a single drug should move on to a phase III study but which

doublet or triplet combinations may be best for phase III studies. Additional concerns over phase II trial limitations include the dilemma of defining the activity of a cytostatic agent when planning a phase III study since our current paradigm of response rate was based on evaluation of cytotoxic agents.

This review will focus on options for phase II trial designs that could be considered to further optimize trial development. We discuss the use of the randomized phase II design with placebo control and a time to progression end point (TTP). We also discuss the strengths and limitations of phase II trial designs that may help investigators select from among numerous novel agents and combinations for phase III testing. Finally, we address phase II trial designs of cytostatic agents, including a randomized discontinuation design as an enrichment strategy.

Standard Single-Arm Phase II Trials

The standard phase II design is a single-arm study that tests a null hypothesis of insufficient efficacy versus an alternative that the treatment has sufficient activity to merit further investigation. The type I error then defines the chance that an ineffective agent will be studied further, and the type II error the chance that an effective agent will not be studied further, which is usually regarded as the more serious error in phase II testing. In a study of monotherapy, an agent might be considered uninteresting if the true response rate is $\leq 5\%$ and interesting if the response rate is at least 20%; this will vary depending on study. These predetermined benchmarks are often tested with Simon's two-stage designs (1).

Determining appropriate end points and levels of activity for single-arm studies of cytostatic agents can be more problematic, but one option is to determine the proportion of patients with

Authors' Affiliations: ¹Eastern Cooperative Oncology Group, Harvard School of Public Health; ²Dana-Farber Cancer Institute; ³Beth Israel Deaconess Medical Center, Boston, Massachusetts; ⁴National Cancer Institute, Cancer Therapy Evaluation Program, Bethesda Maryland; ⁵Our Lady of Mercy, Bronx, New York; ⁶Johns Hopkins University, Baltimore, Maryland; ⁷The University of Medicine and Dentistry of New Jersey, The Cancer Institute of New Jersey, New Brunswick, New Jersey; ⁸Medical College of Wisconsin, Milwaukee, Wisconsin; ⁹Indiana University, Bloomington, Indiana; ¹⁰The University of Wisconsin Comprehensive Cancer Center, Madison, Wisconsin; and ¹¹Cleveland Clinic Foundation, Cleveland, Ohio
Received 5/25/05; revised 12/5/05; accepted 12/29/05.

Requests for reprints: Robert S. DiPaola, Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08901. Phone: 732-235-7469; E-mail: dipaolrs@umdnj.edu.

©2006 American Association for Cancer Research.
doi:10.1158/1078-0432.CCR-05-1136

improvement in disease stabilization (i.e., freedom from progression) at a defined time point (2). Additional end points may include multinomial testing for improvement in either response or disease stabilization, TTP, and progression-free survival. Typically, single-agent treatments need to be shown to have activity similar to other active single agents in the particular disease setting, and new agents in combination with standard therapies need to show a higher level of activity than the standard combination treatments alone. If the activity levels defining promising and inactive agents have not been selected appropriately for the population enrolled on a study, then the actual rate of errors in selecting agents for further study could be larger than the nominal type I and II error rates in the design. Despite the inadequacies of phase II trials with these inherent error rates, the selection of therapies for phase III testing is usually based on multiple phase II studies, reducing both the type I and type II error. Although a standard in trial design, these single-arm phase II studies may be inadequate for many current agents and combinations. As noted earlier in this review, the panoply of targeted agents and potential combinations creates a greater need to effectively decide on the correct phase III trial to move forward. Additionally, the end points for cytostatic agents are usually difficult to define in standard phase II trial designs. For example, the level of interest of a TTP end point of a cytostatic agent is usually difficult to define, given a greater variability of this variable historically than even response with a cytotoxic agent (2–5). Using such end points will often require more frequent CT scanning, centralized review of tumor measurements, measuring all tumor burden, and coming up with a formula that integrates tumor volume and time. For example, a study by Yang et al. (6) treated patients with renal cell carcinoma with bevacizumab and effectively assessed tumor shrinkage over time. Such activity was not adequately captured by either response rates or TTP analyses and seemed to be a more powerful discriminator of clinical activity to reduce the type II error.

Although a single-arm phase II study may be the best choice to quickly determine if a drug should move on for further development, limitations exist, and the strengths and weaknesses of other phase II designs should also be considered. Additional trial designs considered in this review include randomized selection phase II studies, comparative randomized control studies, and randomized discontinuation studies.

Randomized Selection Phase II Studies

A randomized phase II selection design allows multiple single-arm studies to be conducted in the same time frame and with the same entry criteria (3, 7). The advantages of a randomized study over separate studies include decreasing the effects of patient selection bias, population drift and stage migration, and the ability to ensure that uniform evaluation criteria are used. Although these studies are often designed to separately evaluate each arm, and there is generally not adequate power for formal tests comparing arms, a predetermined plan for selection of arms for future study can be made in such a design. Typically, this randomized selection design randomizes between two or more experimental arms without a control arm (3, 4, 7). Usually, there will be a test for activity of each arm, using standard criteria for single-arm studies, and a

rule for selecting the “best” of the active arms for further testing. This differs from the common inference design, which randomizes without preplanned selection rules. Often, the rule for a selection design study is to select as “best” the arm with the best efficacy level, no matter what the magnitude of the difference. As noted above, the strengths of this design include less selection bias due to changing natural history or outcome improvements over time in sequentially conducted phase II studies. A weakness of this design is the reduced likelihood of being able to select the best arm with increasing number of arms in the study or if there is a small difference in activity among arms. Table 1 gives selection probabilities for a two-arm study with a null hypothesis of a response rate of 20% and an alternative hypothesis of 40% for each arm. If neither arm meets the criteria for sufficient activity (at least 11 responses of 36 patients in this case), then neither is selected. The probability that the better arm has the most responses (and hence is the winner) is 90% if one arm is at the null and one at the alternative response rate. The probability of selecting the better arm decreases as the difference between the arms decreases, and there is a 14% chance of selecting an arm with a 30% response rate over one with a 40% rate, for example. As with any type of testing, there is a tradeoff between larger studies with smaller error rates versus smaller studies with more rapid accrual, evaluation, and selection. In this setting, if arms have nearly equal levels of activity, then selecting either for further study is reasonable. Viewed as tests of the null hypothesis of equal activity, selection designs have high type I error rates, however, and thus are not a substitute for phase III studies. The probability of selecting the best arm also declines as the number of arms increases. A design with 90% probability of correct selection with two arms, one at the null and one at the alternative, would have 72% probability of picking a winner in a six-arm study with the same number of patients per arm with five arms at the null and one at the alternative hypothesis. This could be more problematic if the difference between some arms is small. However, the design may still have value as an improvement in selection beyond multiple sequential phase II studies, due to reduced potential for bias, and the randomized phase II with more than two arms would

Table 1. Probability of arm selection based on meeting criteria for activity and observing more responses

Response		Probability arm selected as winner			
RX1	RX2	RX1	RX2	Tie	Neither*
0.20	0.20	0.08	0.08	0.00	0.83
0.20	0.25	0.07	0.26	0.01	0.66
0.20	0.30	0.05	0.51	0.01	0.42
0.20	0.35	0.03	0.75	0.01	0.21
0.20	0.40	0.01	0.90	0.01	0.08
0.25	0.40	0.06	0.85	0.03	0.07
0.30	0.40	0.14	0.76	0.06	0.04
0.35	0.40	0.28	0.62	0.08	0.02
0.40	0.40	0.45	0.45	0.09	0.01

*The probability that neither meets the minimum level of activity (11 responses in 36 patients).

often lead to elimination of some of the arms based on formal selection criteria. The feasibility of correlative studies can also be considered in these studies, but studies with correlative end points must often be larger than those looking exclusively at traditional end points due to increased interpatient and inpatient variability. The validation of molecular targets in larger studies would likely be required before the design of future trials restricting entry criteria based on target assessment.

Phase II Randomized Control Study Designs

Whereas a randomized selection design may have advantages when screening multiple treatments, the study of fewer comparisons could warrant a study designed to test a direct comparison, as a randomized controlled phase II study. A randomized control phase II study design typically compares an experimental regimen to a control arm (with or without a placebo). Comparison to a control arm is most useful when there is little prior information on expected efficacy rates in a population, and can also be useful for end points that can be heavily influenced by patient selection, such as TTP and progression-free survival (2–4). They are also useful when improvements in supportive care or other clinical factors are suspected to change the expected outcome in the population being studied. The efficacy end point could be a standard measure of disease status such as objective response or TTP, allowing the study to be completed with fewer patients than required in a phase III study of survival. Because the primary end point is not survival, a crossover to drug from the placebo arm when the end point is reached can also be considered and is a strength by attracting patient accrual, but a weakness in that it can dilute determination of a survival benefit. The type I and II error rates vary, but as in other phase II designs, the type I error can generally be larger than in phase III studies. Korn et al. (2) suggest considering one-sided type I error rates as large as 20%. The magnitude of the difference between the null and alternative hypotheses may also be larger than appropriate for a phase III study. For example, Yang et al. (6) studied the effect of two different doses of bevacizumab compared with placebo in a randomized phase II study. The primary end point of this study was to detect a 100% improvement in median TTP with a two-sided type I error of 0.05 and a type II error of 0.20. A second example, E5397, comparing cisplatin/C225 versus cisplatin/placebo for head and neck cancer, was designed to have 90% power to detect a doubling of median progression-free survival from 2 to 4 months, also using a two-sided type I error of 0.05 (3, 8, 9). Both studies allowed patients to crossover from placebo at progression, creating an advantage to offer all patients the option of therapy, but may have diluted a survival benefit. A potential weakness of a randomized phase II control design often includes the need for a second larger study. A decision as to whether a single-arm phase II study will be adequate to make a decision to go forward to a phase III study or if a randomized comparison with a progression end point will be required should be made prospectively.

Additional aspects to consider in a phase II randomized control study include the question of whether a placebo is needed in a control arm (versus an unblinded control). In settings where an effective therapy exists, the control arm might thus consist of effective standard therapy plus placebo. The

added expense of a placebo and administrative burden needs to be weighed against the potential for bias. The effectiveness of the placebo in blinding treatment should also be considered because many agents have significant side effects. For evaluating TTP in a study with crossover at progression, there is usually a serious risk of bias, especially when the control arm includes no active agents. However, in many phase II studies, the magnitude of the potential bias should be small relative to the effect that can be detected.

The study by Yang et al. (6) showed an improved TTP and therefore supported additional study, despite a 10% response rate and 4.8 months median TTP, and, therefore, supported continued interest in this agent and guided further studies. In contrast, E5397 did not find a significant difference in progression-free survival. Because of the large difference that was targeted, this study did not definitively establish whether there is a meaningful benefit from C225 therapy on progression-free survival or survival. Also, this design comparing a doublet with a single agent may be useful in a situation where we do not know if the combinations are additive, synergistic, or antagonistic, or may add toxicity. In summary, although useful in some settings, the phase II randomized control design typically will not be definitive, but will often aid decisions on whether to pursue further study of the treatment and guide the design of additional trials. A careful decision needs to be made to determine if the information from a single-arm phase II study will be enough to guide future studies. Moreover, if a randomized control phase II design is considered, care must be taken in deciding on the difference that will be considered a clinically meaningful positive result. Another option with multiple experimental arms is that randomized selection can be combined with a control arm in a phase II/III design that rolls directly into a phase III comparison of the best arm(s) to the control (7).

Randomized Discontinuation Design

As noted in the phase II randomized selection and control designs, cytostatic agents may be better evaluated with TTP end points in contrast to response end points. To better determine the activity of agents that may have limited response activity, an enrichment strategy may be appropriate. A randomized discontinuation design is an example of an enrichment strategy (5). The design selects a more homogeneous group of patients, and theoretically makes it more likely to derive clinical benefit for randomization. In a discontinuation design, patients are randomized between continuing drug or going on a placebo if free of progression at some defined time point. Patients are usually crossed over from placebo to treatment at progression or specified progression-free interval. Patients who are responding continue on therapy until progression. The design is complex with three registration points: initial registration, at the defined time of randomization during stable disease, and crossover at progression. This design is most attractive in settings where it is thought that only a subset of the population will benefit from the treatment, which might be the case for some targeted therapies. It has been shown that these designs can theoretically be effective in these settings (10, 11), but their effectiveness depends on the extent to which the initial run-in can select out the subgroup that is clearly benefiting on the basis of disease stabilization versus rapid progression. Other potential

weaknesses of this study design include the possibility of a carry-over effect from the run-in that could dilute differences between the randomized arms. Desired patient numbers also need to be considered carefully because these studies tend to be much larger than other forms of phase II studies. For example, one completed randomized discontinuation design in renal cancer by Stadler et al. (5, 12) enrolled 374 patients on a drug (carboxyaminoimidazole) that seemed to have little activity in this disease. In this study, only 17% of patients were eligible to be randomized as the enriched stable population. Although 64 patients were randomized, a Bayesian futility analysis was done in the first 49 patients randomized, which suggested that a positive conclusion in the best of assumptions with continuing to 100 randomized patients was <9% and the trial was appropriately halted. Although this analysis will be discussed later in this review, the incorporation of early stopping rules based on the nonrandomized portion of this trial was important and should be considered in these trial designs. An additional issue is the concern that patients randomized to placebo may be aware of the likelihood of placebo, having experienced effects of the drug in the nonrandomized phase of the study. Patients on the placebo may discontinue the study or be scored more easily as progression. In the study cited above by Stadler et al., a patient-completed survey was obtained to determine if patients may have withdrawn early from study when randomized to placebo. Similar assessments should be encouraged if this study design is considered. Therefore, it is also important for these studies to be done in a double-blind manner and to ensure that clear and strict rules of progression are followed.

Other Designs

Factorial designs, where subjects are randomized among all possible combinations of two or more drugs, can also be useful in some settings, such as in cancer vaccine development in which vaccine could be combined with different adjuvants (4). A strength of this design is that outcome of an agent, such as a vaccine with added adjuvants, could be analyzed combining all arms with a specific adjuvant compared with all arms without that adjuvant, reducing the numbers needed for comparison. Similarly, early stopping rules can be calculated with the combined patients from all arms with a specific agent and apply to all arms with this agent (4). A weakness, however, in

oncology is that it would usually be necessary to perform phase I studies of all of the individual combinations before phase II testing. Thus, all combinations may not simultaneously be ready for testing in a randomized factorial design.

Bayesian designs allow more flexibility in frequency of analysis and direct incorporation of information on historical controls (13). Response-adaptive designs and flexible selection rules can also easily be formulated. As noted above, in the study by Stadler et al., the use of Bayesian analyses can help to determine early stopping rules. Additionally, in multiarm randomized selection studies, a Bayesian design could be formulated where frequent analyses would be done and arms that were unlikely to have sufficient activity or to be as effective as other arms could be dropped, with randomization continuing among the remaining arms. The practical usefulness of designs based on frequent interim analysis and response-adaptive randomization may be somewhat limited, especially in multi-institution studies, because complete data on consecutive sets of cases enrolled on a study are needed to avoid bias in interim results, which usually means that either accrual must be suspended or that the interim analyses will lag well behind patient enrollment.

Conclusion

The demand for additional phase II trial designs has increased due to the emergence of many promising agents. Concerns with standard phase II design includes the inability to determine activity with a cytostatic agent and the inability to decide among many agents and combinations for future phase III testing. Trial designs considered in this review included randomized phase II selection studies, comparative randomized control studies, and randomized discontinuation studies to guide decisions on cytostatic agents. These phase II trial designs may require greater accrual capacity, multidisciplinary interactions, agents from multiple pharmaceutical companies, and comprehensive laboratory expertise for expanded correlate studies. As agents become available for testing in the clinic, the strengths and weaknesses of many different phase II trial designs should be considered to optimize a trial development plan that guides phase III trial decisions more effectively. The efficient development and completion of phase II and III studies will be critical for the advancement of clinical research.

References

- Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989;10:1–10.
- Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *J Clin Oncol* 2001;19:265–72.
- Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep* 1985;69:1375–81.
- Simon RM, Steinberg SM, Hamilton M, et al. Clinical trial designs for the early clinical development of therapeutic cancer vaccines. *J Clin Oncol* 2001;19:1848–54.
- Rosner GL, Stadler W, Ratain MJ. Randomized discontinuation design: application to cytostatic antineoplastic agents. *J Clin Oncol* 2002;20:4478–84.
- Yang JC, Haworth L, Sherry RM, et al. A randomized trial of bevacizumab, an anti-vascular endothelial growth factor antibody, for metastatic renal cancer. *N Engl J Med* 2003;349:427–34.
- Scher HI, Heller G. Picking the winners in a sea of plenty. *Clin Cancer Res* 2002;8:400–4.
- Burtneess B, Goldwasser MA, Flood W, et al. Phase III randomized trial of cisplatin plus placebo compared with cisplatin plus cetuximab in metastatic/recurrent head and neck cancer: an Eastern Cooperative Oncology Group study. *J Clin Oncol* 2005;23:8646–54.
- Burtneess B. The role of cetuximab in the treatment of squamous cell cancer of the head and neck. *Exper Opin Biol Ther* 2005;5:1085–93.
- Temple RJ. Enrichment designs: efficiency in development of cancer treatments. *J Clin Oncol* 2005;23:4838–9.
- Freidlin B, Simon R. Evaluation of randomized discontinuation design. *J Clin Oncol* 2005;23:5094–8.
- Stadler WM, Rosner G, Small E, et al. Successful implementation of the randomized discontinuation trial design: an application to the study of the putative anti-angiogenic agent carboxyaminoimidazole in renal cell carcinoma-CALGB 69901. *J Clin Oncol* 2005;23:3726–32.
- Thall PF, Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med* 1998;17:1563–80.