Published in final edited form as:

Clin Trials. 2012 December ; 9(6): 741–747. doi:10.1177/1740774512464724.

Integrated phase II/III clinical trials in oncology : A case study

Meihua Wang^a, James J Dignam^{a,b}, Qiang E Zhang^a, John F DeGroot^c, Minesh P Mehta^d, Sally Hunsberger^e

^aDepartment of Statistics, Radiation Therapy Oncology Group, American College of Radiology, Philadelphia, PA, USA,

^bDepartment of Health Studies, University of Chicago, Chicago, IL, USA,

^cDepartment of Neuro-Oncology, Division of Cancer Medicine, MD Anderson Cancer Center, The University of Texas, Houston, TX, USA,

^dDepartment of Radiation Oncology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA,

^eDivision of Cancer Treatment and Diagnosis, Biometric Research Branch, National Cancer Institute, Bethesda, MD, USA

Abstract

Background—Integrated phase II/III trial designs implement the phase II and phase III aspects of oncology studies into a single trial. Despite a body of literature discussing the merits of integrated phase II/III clinical trial designs within the past two decades, implementation of this design has been limited in oncology studies.

Purpose—We provide a brief discussion of the potential advantages and disadvantages of integrated phase II/III clinical trial designs in oncology and provide an example of the operating characteristics of a Radiation Therapy Oncology Group (RTOG) trial.

Methods—We review the differences among proposed integrated phase II/III designs. Then, we illustrate the use of the design in a brain tumor trial to be conducted by the RTOG and examine the impact of association between endpoints on design performance in terms of type I error, power, study duration, and expected sample size.

Results—Although integrated phase II/III designs should not be used in all situations, under appropriate conditions, significant gains can be achieved when using integrated phase II/III designs, including smaller sample size, time and resources savings, and shorter study duration.

Reprints and permission: <http://www.sagepub.co.uk/journalsPermissions.nav> 10.1177/1740774512464724

Author for correspondence: Meihua Wang, Department of Statistics, Radiation Therapy Oncology Group, American College of Radiology, 1818 Market Street, 16th Floor, Philadelphia, PA 19103, USA. sumswang@yahoo.com.

Conflict of interest

Conflict of Interest Notification: Minesh Mehta has or has had the following roles in the last 2 years (2010–2011): **Consultant:** Adnexus, Bayer, Bristol-Meyers-Squibb, Elekta (non-reimbursed), Merck, Novartis, Quark, and Tomotherapy; **Stock Options:** Accuray, Colby, Pharmacyclics, Procetus, and Stemina; **Data Safety Monitoring Boards:** Apogenix; **Board of Directors:** Pharmacyclics; **Medical Advisory Boards:** Colby, Stemina, and Procetus; **Speaker:** GRACE Foundation, MCM, Merck, prime Oncology, Strategic Edge, and WebMD; **Patents:** WARF/Procetus; **Royalties:** DEMOS Publishers. This publication's contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute nor the Pennsylvania Department of Health.

Limitations—Data submission without delay and sufficient evaluation of intermediate endpoints are assumed.

Conclusions—Although there are potential benefits in using phase II/III designs, there also may be disadvantages. We recommend running design simulations incorporating theoretical and practical issues before implementing an integrated phase II/III design.

Background

Integrated phase II/III clinical trial designs are those in which the phase II and III components are combined in a single trial and the data based on the phase II component are used to test the study hypothesis of the phase III component. This approach was first proposed and then formalized in terms of controlling overall size and power and optimizing overall sample size at least two decades ago [1, 2]. Until recently, implementation of this design in cancer clinical trials has been relatively limited, but there is renewed interest in this design with increasing demands in terms of improving the entire clinical trials conduct process. The main reason for the previous limited use of these designs may be that small single-arm phase II studies were considered adequate to obtain informative activity data for new experimental therapies and thus served ‘adequately’ to screen treatments for advancement to phase III testing. Currently, randomized phase II designs are favored (and very often needed because of the recognition that ‘historic’ phase II controls are increasingly unreliable) to reliably evaluate the activity of new agents in combination with known active agents or in combination with other new agents [3,4]. Randomized trials require many more patients than single-arm studies, which would be considered a slight disadvantage, but often the arms evaluated in the phase II study will be identical to those evaluated in the phase III study, and clever designs might permit the use of phase II data for the phase III evaluation.

In recent years, there have been several reviews of phase II/III designs. Rubinstein *et al.* [3] reviewed phase II/III trials in the context of a proposal for wider use of randomized phase II trials. In his review, Thall [5] concluded that a properly designed phase II/III design utilizes resources more efficiently and yields more reliable results. Bretz *et al.* [6] and Schmidli *et al.* [7] each reviewed phase II/III clinical trials where one uses interim analysis data to guide hypothesis selection, and Jennison and Turnbull [8,9] have provided additional considerations and examples of this particular type of design.

In this article, we first briefly review the integrated phase II/III designs that have been proposed, discussing the advantages and limitations of each. We then present an application of the design using an example of a brain tumor trial to be conducted by the Radiation Therapy Oncology Group (RTOG). We also present a simulation study to further evaluate the operating properties of this design.

A brief review of phase II/III designs and their application to cancer clinical trials

Phase II/III designs using the same endpoint for each stage

In the classical phase II/III design setting as proposed by Thall *et al.* [10], Schaid *et al.* [11], and Stallard and Todd [12], the same endpoint is used for both the phases II and III components of the trial. These designs can effectively be viewed as a phase III study with aggressive group sequential interim futility analyses. The futility boundary corresponds to the marginal or minimal improvement of the new regimen, and the investigator would like to observe before continuing to the phase III component. The type I error, power, and effect size for the phase III study are used to determine the sample size, critical value, and the decision rule for proceeding to phase III, which are prespecified in the protocol. Generally, the criteria for continuing to the end of the study correspond to a very modest observed benefit in the primary endpoint, and this is necessary to maintain the overall power of the study. The use of these designs with an endpoint such as overall survival (OS) may be problematic when median OS is relatively long and accrual is rapid. In order to maintain power for the overall trial, the phase II analysis must therefore be carried out when it is sufficiently informative, that is, an adequate number of events has been reached, and frequently, this could occur after accrual is complete, thereby losing a considerable part of the benefit of performing the interim analysis.

Adaptive designs

Adaptive designs proposed by Bauer and Köhne [13], Proschan and Hunsberger [14], Bauer and Kieser [15], and other authors also use the same endpoint for the interim analysis as for the final analysis. Adaptive designs differ in that along with stopping or continuing after the phase II (or interim analysis), the study may continue with a new sample size goal for the overall study. The new sample size for the phase III study can be recalculated based on the phase II data from the study or other external information. Data from both the phases are combined for final statistical inference. The approaches adjust the critical values or use a combination test with p values from the different stages to avoid inflation of the type I error. The adaptive designs are less efficient than the phase II/III designs but offer more flexibility.

Phase II/III designs with different endpoints for each phase

In order to address the issue of minimal savings in time and patient numbers when using endpoints with longer time horizons, such as OS for both the phases II and III portions of a trial, Hunsberger *et al.* [4] and Royston *et al.* [16,17] have proposed using an endpoint that is reached earlier for the phase II component and could putatively be reflective (and not necessarily as a true surrogate endpoint) of clinical benefit, thereby providing confidence to proceed to Phase III testing utilizing a more definitive endpoint. Examples of such an 'early' endpoint include response rate, progression-free survival (PFS), or any other endpoint that could provide confidence that a benefit in a more robust endpoint such as OS could be expected. In the setting of phase II/III trials, we note as others have, that the phase II endpoint need not satisfy traditional surrogacy criteria, but rather should provide evidence of potential activity on the ultimate endpoint [18,19]. For example, the intermediate endpoint

response may be a ‘necessary but not sufficient’ condition for benefit on the ultimate endpoint.

Multiple treatment arms in phase II with possible reduction of arms in phase III

One or more experimental arms could be tested in integrated phase II/III designs. Simon *et al.* [20], Whitehead [21], Thall *et al.* [10], Schaid *et al.* [11], and Royston *et al.* [16] proposed the ideas of testing several experimental arms at the same time instead of testing one experimental treatment each time. Only the arms crossing the prespecified efficacy boundaries in the phase II stage would be carried forward to the phase III stage. From the patient perspective, the multiple arm designs increase the chance of being assigned to a promising new treatment. When several sponsors are involved, a single trial may be complex to develop and carry out because of conflicting proprietary interests and regulatory concerns, but these barriers can be overcome [18]. Furthermore, there are many instances when multiple agents are commercially available, or different dosing schedules of the same agent are of interest to study. In all of these situations, evaluating several treatments simultaneously instead of one at a time would speed up the discovery of new beneficial treatments.

Example: a clinical trial for anaplastic glioma

The RTOG is interested in testing whether the antiangiogenesis agent bevacizumab combined with lomustine, an established chemotherapy agent, is more beneficial than lomustine alone in treating recurrent temozolomide-resistant anaplastic glioma (AG, Grade 3 glioma). Bevacizumab has been shown to be beneficial in other types of cancer, including recurrent glioblastoma (Grade 4 glioma). Prior smaller studies in AG and glioblastoma multiforme (GBM) suggest lomustine and bevacizumab may be synergistic. Therefore, there is significant confidence by the RTOG that if a randomized phase II study of the combination showed promise over single agent lomustine, a phase III study of the same two arms would be of top priority.

Response rate is not a good indicator of activity for brain tumors due to lack of evidence of association between response and OS, the relatively low overall rate of response, and issues surrounding pseudoresponse with the use of antiangiogenic agents. PFS is thought to be a better early indicator of activity, and several analyses have demonstrated an association between PFS and OS in recurrent malignant glioma [22]. However, PFS remains highly variable among different malignant glioma patient cohorts, and therefore, data from a single-arm trial would be difficult to interpret as providing pilot evidence in favor of this treatment. Thus, a randomized phase II study would be the pilot activity design of choice, with positive results supporting continuation to a phase III study with OS as the more definitive and robust endpoint.

We now describe in detail what a phase II/III study for the trial above would look like. The primary endpoint would be PFS for the phase II portion and OS for the phase III portion. Based on the published data in this patient population [23–26], the median PFS and OS with standard treatments are estimated as 5.1 and 14.6 months, respectively. Hunsberger *et al.*’s integrated II/III design [4] has been proposed for this study. The calculations are based on a

conservative sample size approach that assumes that the PFS and OS endpoints are independent. There would be no accrual suspension between the phases II and III components. For the PFS endpoint, we consider an increase in median PFS to 8.4 months to warrant a phase III study. Therefore, we will power the phase II portion of the study to be able to detect a hazard ratio of 0.61. For the OS endpoint, we will consider a 25% relative reduction in mortality hazard to be clinically relevant, which implies an increase in median OS to 19.4 months.

The total sample size for the II/III study is 675 patients, with definitive analysis taking place when 532 deaths have occurred. This sample size is based on the following assumptions and parameters: exponential survival time distributions (and the standard logrank formula), accrual rate of 22.5 patients per month (30 month accrual period), a final test performed at the one-sided 0.025 level, and a minimum follow-up of 19 months after accrual closure. This design will provide 90% power to detect a hazard ratio of 0.75. The phase II futility analysis will take place after 102 PFS events have occurred and is similarly based on the logrank test. If the one-sided p-value is less than 0.2, the study will continue to the full sample size. Otherwise, the trial will be terminated for insufficient evidence of activity with respect to PFS. The phase II sample size (PFS events) of 225 (102) has high power for large PFS hazard reduction, with 95% power to detect a hazard ratio of 0.61.

Several important aspects of this design are now presented. The overall statistical power for the study or the probability of correctly concluding a difference in OS is 0.85 (inclusion of the futility analysis reduces the power from 0.9 to 0.85). The overall significance level is no larger than 0.025, and no adjustment is needed since the study will not conclude efficacy based on the PFS endpoint. The results for the PFS analysis will be reported only to the RTOG Data Monitoring Committee (DMC), so that no clinical accrual bias is introduced as the phase III portion is accruing. Aside from the phase II/III aspects of the design, standard components of phase III study will be included. For example, two interim analyses for efficacy will be conducted using the O'Brien–Fleming method [27] with one-sided significance level of 0.0015 and 0.0092, respectively. A futility analysis will also be performed based on the OS endpoint, testing whether the alternative hypothesis can be rejected at the 0.005 level [28].

In contrast, this trial could also be designed in the traditional way, such as a randomized screening phase II, followed by a randomized phase III trial if applicable. We investigate the possible benefits and drawbacks of using an integrated phase II/III design through the following perspectives.

Sample size

A randomized screening phase II design with types I and II errors of 0.2 and 0.05, respectively, would require 225 patients (102 events) to detect the PFS difference. It is important to note that the above power specification is not typical for randomized phase II trials, but we use this so that we can make comparisons of these two designs. If the phase II results support further testing through a phase III trial, in which the design parameters are kept the same as those in the above integrated phase II/III design, 675 analyzable patients are needed for the phase III trial. Therefore, we can see that the integrated phase II/III design

leads to a relative 25% reduction in the number of required patients. In addition, the expected sample size under the global null hypothesis (no benefit for either endpoint), equaling $n_1 + n_2\alpha_1$, where n_1 is the sample size from phase II, n_2 is the additional sample size required for phase III, and α_1 is the type I error for the phase II study, is 315 for the integrated phase II/III design and 360 for the traditional approach.

Development time and resources

In the phase II/III design, a single protocol is written rather than two separate protocols, reducing redundancies in effort by sharing much of the structure of the protocol and other materials such as case report forms. There is also potential for considerable savings in the author and sponsor reviews and approvals, as well as approval by the Institutional Review Boards. A single protocol, although more complicated, thus reduces the administrative work associated with sequential phase II and III trials. Although it is important to note that if there is not a certainty that a phase III study would be launched, the administrative work may actually be increased, since the phase III protocol is often more complex to develop than a phase II study. In contrast, for the traditional design, there exists a gap between completion of phase II and launch of phase III for administrative time, and an additional gap due to minimum follow-up for data to mature from the phase II trial (if applicable). Therefore, considerable time and resources can be saved using a phase II/III design.

Study duration

The entire study duration is between 4 and 4.5 years for the integrated phase II/III design, while it is projected to run to between 7 and 8 years for the traditional design due to the larger sample size and gap between the phases II and III components. Therefore, the integrated phase II/III design reduces the study duration by an estimated period of 3 years.

Patient accrual and event rates

In the protocol, it would be more appropriate to base the phase II analysis on the number of events so that the accrual rate and length of follow-up will not impact the operating characteristics. If the analysis time is stated in terms of patient numbers, then if the accrual rate is slower than expected, the futility analysis will be delayed but potentially fewer patients would be needed. If the accrual rate is faster than expected, more patients may need to be accrued in order to observe the number of events needed, but the futility analysis may be able to occur earlier.

Intermediate endpoint evaluation

There remain challenges in clinically distinguishing pseudoprogression from true progression and response from pseudoresponse in brain tumor patients. In the context of our trial described earlier, pseudoprogression would become relevant if patients treated with radiotherapy and temozolomide for their newly diagnosed AG experiencing pseudoprogression, who were labeled as having 'progression', were enrolled on the recurrent disease study. This issue is relatively easily corrected by requiring a minimum timeframe from completion of chemoradiotherapy as an eligibility requirement (e.g., 6 months), since the bulk of pseudoprogression events occur early. The second issue, that is,

response versus pseudoresponse, has some bearing on the PFS endpoint, since it is well known that very early after administration of antiangiogenic agents, the contrast-enhancing portion of the tumor will regress rapidly, thereby creating a radiographic response, but in time, the surrounding nonenhancing fluid attenuated inversion recovery (FLAIR) signal abnormality continues to worsen, suggesting continued growth of the nonenhancing tumor component. This issue is also correctable, especially with the use of the new response assessment in neuro-oncology (RANO) criteria [29]. Therefore, the study protocol must contain detailed and rigorous eligibility criteria, and response and PFS assessments. In addition, delays in data submission must be avoided because timely submission of intermediate endpoint data is also critical to the phase II analysis. Finally, to provide objectivity for the futility decision, the study statistician will provide at the appropriate time the PFS analysis with recommendation to continue or terminate to the DMC, who will then deliberate based on the review of treatment-blinded findings.

Simulation: the impact of correlation between endpoints on operating characteristics

The phase II/III seamless design aims to add logistical efficiency and maximize the contribution of patient information. A critical aspect of the statistical efficiency of this design is the degree to which the effect of treatment on the phase II endpoint can reliably predict the treatment effect on the phase III endpoint. This question actually involves two components: (1) the correlation between endpoints in the absence of treatment and (2) similarity of the treatment effect on the two endpoints. For simplicity here, we assume that if the two endpoints are correlated, then the treatment effect is manifest in both endpoints. We examined the impact of correlation between endpoints for the two trial stages on type I error, power, study duration, and expected sample size. We considered the endpoints of PFS for phase II and OS for phase III and assumed that individual OS and PFS follow a bivariate exponential distribution [30] indexed by three parameters (θ_x , θ_y , and δ)

$$S(X, Y) = \Pr(X > x, Y > y) = \exp\left(-\left(\left(\frac{x}{\theta_x}\right)^{1/\delta} + \left(\frac{y}{\theta_y}\right)^{1/\delta}\right)^\delta\right)$$

where θ_x and θ_y denote the scale parameter for X and Y , respectively; δ introduces a nonnegative correlation between X and Y , which can be expressed in a function of θ_x , θ_y , and δ ; here, X denotes OS and Y denotes PFS. For our purposes, if $X > Y$, OS = X and PFS = Y ; otherwise, OS = PFS = X .

To evaluate the impact of correlation between endpoints under (1) no treatment effect on either PFS or OS (global null); (2) treatment effect on both PFS and OS (global alternative); (3) treatment effect on PFS, but not on OS (PFS-only alternative); and (4) treatment effect on OS, but not on PFS (OS-only alternative), θ_x and θ_y (inverse of hazard rates) were adopted from the AG trial described in section ‘Example: a clinical trial for anaplastic glioma’. Four values of association parameters were studied: $\delta = 0.2, 0.45, 0.7$, and 1. Accounting for the influence of replacing PFS with OS when the generated PFS was greater than the generated OS, and the impact of censoring on the correlation between PFS and OS,

the four selected values approximately represented the Pearson correlation (ρ) of 0.9, 0.7, 0.54, and 0.33, respectively. To measure the operating characteristics of the design under the above four scenarios with varying δ , these four trial outcome summaries were obtained based on 10,000 simulations per scenario: probability of going to phase III, probability of claiming positive phase III, study duration, and expected sample size. The first stage accrual of 10 months for the phase II component, additional accrual of 20 months for the phase III component, monthly accrual of 22.5 patients, and minimal follow-up of 24 months for patients in the stage III accrual were used for the simulations. The simulation results were presented in Table 1.

The simulation results demonstrated the following: (1) under the global null, as expected, the significance level was maintained; (2) under the global alternative, a high correlation led to slightly higher probability of going to phase III and slightly higher probability of positive phase III claims; (3) there was little influence on study duration and expected sample size with varying correlations under either global hypothesis; (4) under the PFS-only alternative, a high correlation led to a higher probability of going to the phase III, longer study duration, and larger expected sample size, but the overall significance level was maintained; and (5) under the OS-only alternative, a high correlation led to a lower probability of going to the phase III and lower probability of positive phase III claims, shorter study duration, and smaller expected sample size.

One motivation for considering the joint distribution of PFS and survival is to be able to assess the association between endpoints on the trial operating characteristics. The effect of association between toxicity and response was studied in the phase II frequentist design context by Bryant and Day [31]. In that case, correlation between toxicity events and response events had surprisingly little effect on the rejection region. In the Bayesian context studied by Wang and Day [32], the correlation also demonstrated little effect on dose assignments and overall operating characteristics. The effect of correlation between two treatment effects (the correlation of hazard ratios in terms of two endpoints) was studied in phase II/III design context by Royston *et al.* [16]. They demonstrated that correlation did not influence the operating characteristics of each separate trial stage, only the overall type I and power. A high correlation substantially increased the overall type I error probability while only slightly increasing the power of the trial. Our results seem to echo those results, with little effect of correlation on study duration and expected sample size under either global hypothesis, while a high correlation led to slightly increased power.

Discussion

In summary, we have presented the integrated phase II/III designs as a strategic way to accelerate assessment of new treatment regimens and deliver options to cancer patients more quickly. The integrated phase II/III design may be appropriate as an alternative to sequential phase II/III designs when (1) a rapidly obtained phase II endpoint, correlated with the primary endpoint (and not necessarily as a direct surrogate) in the phase III component, is available; and (2) positive phase II results would provide sufficient motivation for the launch of the phase III component, and other needs such as budgetary support, patient accrual, and drug distribution are in place. Although the integrated phase II/III designs deliver many

advantages, it would likely be inappropriate in the following situations, due to the requirements of integration of the phase III component: (1) insufficient evidence to warrant the implementation of phase III component after the positive phase II results; (2) a phase II endpoint that requires a lengthy period of time to observe failures with the phase III endpoint observed soon after the phase II endpoint, and thus, little savings in patient numbers or duration of study would be observed; and/or (3) insufficient resources to implement a phase III trial after positive phase II results at this time. We recommend running design simulations incorporating theoretical and practical issues before implementing an integrated phase II/III design. These general principles for the use of the integrated design have guided the RTOG in proposing this design for a trial investigating bevacizumab in recurrent temozolomide-resistant AG. In this instance, this design leads to a relative 25% reduction of required sample size compared to separate randomized phase II and III trials, in addition to anticipated time and resource savings. Investigators should consider the phase II/III design when appropriate, as it may serve as a means to both increase the efficiency of therapy discovery and development process and aid in making better choices in terms of which approaches to take forward to phase III evaluation.

Funding

This project was supported by RTOG grant U10 CA21661 and CCOP grant U10 CA37422 from the National Cancer Institute (NCI) and by Pennsylvania Department of Health 2009 Formula Grant 4100050889.

References

1. Ellenberg S, Eisenberger M. An efficient design for phase III studies of combination chemotherapies (with discussion). *Cancer Treat Rep* 1985; 69: 1147–54. [PubMed: 4042093]
2. Thall PF, Simon R, Ellenberg SS, Shrager R. Optimal two-stage designs for clinical trials with binary response. *Stat Med* 1988; 7(5): 571–79. [PubMed: 3387716]
3. Rubinstein L, Korn E, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol* 2005; 23: 7199–206. [PubMed: 16192604]
4. Hunsberger S, Zhao Y, Simon R. A comparison of phase II study strategies. *Clin Cancer Res* 2009; 15: 5950–55. [PubMed: 19789306]
5. Thall P A review of phase 2–3 clinical trial designs. *Lifetime Data Anal* 2008; 14(1): 37–53. [PubMed: 17763973]
6. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts (with discussion). *Biom J* 2006; 48(4): 623–34. [PubMed: 16972714]
7. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biom J* 2006; 48(4): 635–43. [PubMed: 16972715]
8. Jennison C, Turnbull BW. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. *Biom J* 2006; 48(4): 650–55. [PubMed: 16972717]
9. Jennison C, Turnbull BW. Adaptive seamless designs: Selection and prospective testing of hypotheses. *J Biopharm Stat* 2007; 17(6): 1135–61. [PubMed: 18027222]
10. Thall P, Simon R, Ellenberg S. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; 75: 303–10.
11. Schaid D, Wieand S, Therneau T. Optimal two-stage screening designs for survival comparisons. *Biometrika* 1990; 77(3): 507–13.

12. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med* 2003; 22: 689–703. [PubMed: 12587100]
13. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analysis. *Biometrics* 1994; 50: 1029–41. [PubMed: 7786985]
14. Proschan MA, Hunsberger S. Designed extension of studies based on conditional power. *Biometrics* 1995; 51: 1315–24. [PubMed: 8589224]
15. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med* 1999; 18: 1833–49. [PubMed: 10407255]
16. Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes, with an application in ovarian cancer. *Stat Med* 2003; 22(14): 2239–56. [PubMed: 12854091]
17. Royston P, Barthel FMS, Parmar MKB, Choodari-Oskooei B, Isham V. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials* 2011; 12(81).
18. Parmar MK, Barthel FM, Sydes M, et al. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst* 2008; 100(17): 1204–14. [PubMed: 18728279]
19. Sydes MR, Parmar MK, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: The MRC STAMPEDE trial. *Trials* 2009; 10: 39. [PubMed: 19519885]
20. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep* 1985; 69: 1375–81. [PubMed: 4075313]
21. Whitehead J Sample sizes for phase II and phase III clinical trials: An integrated approach. *Stat Med* 1986; 5(5): 459–64. [PubMed: 3787000]
22. Ballman KV, Buckner JC, Brown PD, et al. The relationship between six-month progression-free survival and 12-month overall survival end points for phase II trials in patients with glioblastoma multiforme. *Neuro-Oncol* 2007; 9(1): 29–38. [PubMed: 17108063]
23. Norden A, Drappatz J, Muzikansky A, et al. An exploratory survival analysis of anti-angiogenic therapy for recurrent malignant glioma. *J Neurooncol* 2009; 92: 149–55. [PubMed: 19043778]
24. Norden A, Young G, Setayesh K, et al. Bevacizumab for recurrent malignant gliomas: Efficacy, toxicity, and patterns of recurrence. *Neurology* 2008; 70: 779–87. [PubMed: 18316689]
25. Desjardins A, Reardon DA, Herndon JE, et al. Bevacizumab plus irinotecan in recurrent WHO grade 3 malignant gliomas. *Clin Cancer Res* 2008; 14(21): 7068–73. [PubMed: 18981004]
26. Chamberlain M, Johnston S. Bevacizumab for recurrent alkylator-refractory anaplastic oligodendroglioma. *Cancer* 2009; 115(8): 1734–43. [PubMed: 19197992]
27. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35: 549–56. [PubMed: 497341]
28. Freidlin B, Korn EL. A comment on futility monitoring. *Control Clin Trials* 2002; 23: 355–66. [PubMed: 12161079]
29. Wen PY, Macdonald DR, Reardon DA, et al. Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *J Clin Oncol* 2010; 28(11): 1963–72. [PubMed: 20231676]
30. Hougaard P A class of multivariate failure time distributions. *Biometrika* 1986; 73: 671–78.
31. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995; 51: 1372–83. [PubMed: 8589229]
32. Wang M, Day R. An adaptive Bayesian approach to jointly modeling response and toxicity in phase I dose-finding trials. *J Biopharm Stat* 2010; 20(1): 125–44. [PubMed: 20077253]

Simulation results showing the impact of correlation between endpoints on phase II/III design performance PFS: progression-free survival; OS: overall survival.

Table 1.

	$\delta\rho$	Probability of going to phase III	Probability of claiming positive phase III	Expected sample size ^a	Study duration ^b (months)
Global null	0.2 (0.9)	0.202	0.014	316	18.9
	0.45 (0.7)	0.198	0.013	314	18.7
	0.7 (0.54)	0.202	0.011	316	18.9
Global alternative	1 (0.33)	0.207	0.012	318	19.1
	0.2 (0.9)	0.923	0.849	640	50.6
	0.45 (0.7)	0.914	0.838	636	50.2
PFS-only alternative	0.7 (0.54)	0.913	0.833	636	50.2
	1 (0.33)	0.916	0.837	637	50.3
	0.2 (0.9)	0.917	0.042	638	50.4
OS-only alternative	0.45 (0.7)	0.871	0.044	617	48.3
	0.7 (0.54)	0.836	0.042	601	46.8
	1 (0.33)	0.820	0.040	594	46.1
	0.2 (0.9)	0.203	0.199	316	18.9
	0.45 (0.7)	0.227	0.220	327	20.0
	0.7 (0.54)	0.269	0.254	346	21.8
	1 (0.33)	0.324	0.301	371	24.2

Note: δ denotes the association parameter and ρ denotes the Pearson correlation parameter between the two endpoints.

^aExpected sample size is $n_1 + n_2p_1$, where n_1 is the sample size from phase II, n_2 is the additional sample size required for phase III, and p_1 is the probability of going forward to phase III.

^bStudy duration is $t_1 + (t_2 + t_2)p_1$, where t_1 is the accrual time from phase II, t_2 is the additional accrual time required for phase III, t_2 is the minimal follow-up time for patients enrolled to phase III, and p_1 is the probability of going forward to phase III.