EUROPEAN JOURNAL OF CANCER 45 (2009) 275–280

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Optimising the design of phase II oncology trials: The importance of randomisation

Mark J. Ratain[a,*], *Daniel J. Sargent*[b]

[a]*Section of Hematology/Oncology, Department of Medicine, Committee on Clinical Pharmacology and Pharmacogenomics, The University of Chicago, Chicago, IL, United States*
[b]*Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, United States*

ARTICLE INFO

Article history:
Received 17 October 2008
Accepted 29 October 2008
Available online 6 December 2008

Keywords:
Phase II
Clinical trial design
Randomization

ABSTRACT

Oncology trial end-points continue to receive considerable attention, as illustrated by the development and revisions to the RECIST criteria. In this article, we focus the reader away from the issue of end-points for phase II trials and towards what we believe to be an even more important issue, the fundamental need for randomisation in phase II oncology trials, ideally with blinding and dose-ranging. We present arguments to support the proposition that randomisation will enable greater clarity in the interpretation of the phase II trial results, as well as allowing for more precise estimates of the effect size and sample size requirements for definitive phase III trials. Randomisation will also reduce potential bias resulting from inter-trial variability, which inflates both type I and II errors if historical controls are utilised. In the context of a randomised blinded trial, the exact choice of end-point is less critical, although we favour end-points such as the change in tumour size or progression status at a fixed early time point (i.e. 8–12 weeks after randomisation). Although end-points based on RECIST criteria can and should be utilised in randomised phase II trials, we do not believe that revision of the RECIST criteria will result in a fundamental improvement in drug development decisions in the absence of randomised clinical trials at the phase II stage of drug development.

1. Introduction

The phase II clinical trial plays a central role in oncology drug development. After a phase I trial has determined a tolerable dose for a new agent or combination, a well-designed phase II trial should provide the information required to make a go/no-go decision regarding subsequent phase III testing. As phase III trials require several years, hundreds or thousands of patients and often tens or hundreds of millions of dollars, the information that a quality phase II trial can provide is essential to a decision regarding the potential investment in a larger trial. In this paper, we present a rationale for the expanded use of randomisation in phase II oncology trials, in order to better inform this decision-making process.

2. Why do we do phase II trials?

Phase II trials should be most appropriately viewed as proof of concept trials, used for the purpose of determining whether a particular agent (or combination) should be studied further.[1] In this sense, they serve a critical filtering mechanism, in which a negative trial should lead to the discontinuation of development of a new agent for the selected indication. Optimising the filtering process is the critical issue: too tight

* *Corresponding author:* Address: 5841 S, Maryland Ave., MC 2115, Chicago, IL 60637, United States. Tel.: +1 (773) 702 4400; fax: +1 (773) 702 3969.
E-mail address: mratain@medicine.bsd.uchicago.edu (M.J. Ratain).

a filter will terminate promising agents improperly, but a too porous filter will result in an excessive number of costly negative phase III trials. In the previous era of oncology drug development, there were very few drugs available for the study, and as such a porous filter was very appropriate to minimise the possibility that a false-negative result would result in the discarding of a promising agent.

In the current era, there are hundreds of investigational oncology drugs available for the study. As such, we believe that it is more appropriate (at least at a societal level) to use a tighter filter, one that only advances to phase III development of those drugs for which there is a high probability of success in the phase III trial. We recognise that this strategy may be problematic for companies with a single drug in development. However, from a societal view, minimising the number of phase III trials of ineffective agents is important, as patients enrolled on a negative phase III trial may have lost the opportunity to participate in the trials of active investigational agents, and financial and intellectual resources spent on a negative phase III trial similarly would be better spent on the development of new agents. Given the patient and financial resources required for a phase III trial in the current environment, we feel that it is more important to minimise the risk of failing to demonstrate the efficacy in phase III trial than to be concerned about the lost opportunities subsequent to not proceeding to phase III testing.

We propose that attempting to ensure that phase III trials attain success on their specified end-point is best determined through the explicit modelling of the relationship of dose to both efficacy and toxicity. If an identifiable and reliable relationship between dose and the chosen efficacy end-point can be established, considerable evidence is provided in support of the success of the eventual phase III trial, particularly if that efficacy signal is apparent at doses with an acceptable toxicity. On the other hand, if there is no relationship between dose and efficacy, the drug may either be inactive or be active with a wide therapeutic index (Table 1).

Table 1 – Impact of randomisation on strength of inference in phase II trial.

Phase II strategy	Result	Strength of inference
Single-arm trial	Positive signal	Unknown if due to true efficacy or bias
	Negative signal	Unknown if due to lack of efficacy, bias or wrong end-point
Two-arm randomised trial with no control	Positive signal	Some sign that one agent more promising than another, but overall efficacy unclear
	Negative signal	Unknown if due to lack of efficacy, bias or wrong end-point
Two-arm randomised trial with control	Positive signal	Strong efficacy signal, but dosing may not be optimised
	Negative signal	Dose ineffective
Multiple-arm randomised trial with control	Positive signal	Strong efficacy signal, dosing able to be optimised

The information required to enable the best determination of dose–efficacy and dose–toxicity relationships is only obtainable from a randomised comparative phase II trial, as a single-arm trial only provides information for a single dose. Furthermore, unless the efficacy end-point is a response rate, it may not even be clear that the drug is efficacious at all, as demonstration of a prolonged period of stable disease may be due to an inadvertent selection of patients with a favourable natural history, independent of any pharmacological drug effect.

A further factor influencing our endorsement of randomised phase II trials is the reality that a decision to move a drug from phase II to III is not a simple 'thumbs up' or 'thumbs down' decision. The standard phase II single-arm design is based on the promise that a difference of a single success (typically tumour response) determines the phase III go/no-go decision. While even proponents of a single-arm trial likely realise that this is an artificial construct, the lack of randomisation in a single-arm trial hinders the ability to judge the toxicity relative to control, to consider alternate end-points that may be more sensitive to the treatment effect, and to estimate the effect size relative to control that may be expected in the envisioned phase III trial.[2] These effect sizes are the primary determinant of the sample size (and probability of success) in the phase III trial, and may also impact phase III end-point selection. Drugs may be active but fail in phase III because a phase III trial is too small, or because the chosen dose is too low (or high). Furthermore, it is important to have a fairly robust estimate of the effect size, and the reliability of that estimate is determined by the sample size in phase II. Every one of these factors (effect size, relative toxicity, sensitive end-points and dose–efficacy) is better estimated through a direct randomised comparison at the phase II testing stage.

3. What are the options for design of phase II trials?

Historically, phase II oncology trial designs can be divided into three categories: non-randomised (single-arm) trials compared to historical controls, randomised trials with multiple-experimental regimens compared to a historical control, and randomised trials including a prospective control arm. In this section, we provide a brief overview of each of these design strategies.

The single-arm trial has been the most frequently used approach to the efficacy evaluation at the phase II level in oncology, although it should be emphasised that this approach is rarely used in other therapeutic areas. Although there are multiple approaches to the implementation of the single-arm design, the overall principle has changed little since the design of Gehan.[3] Patients are typically enrolled in two stages. If sufficient activity is observed at the end of the first stage, accrual continues to a second stage, and the number of successes (typically tumour responses) at the trial's conclusion determines whether the agent is 'recommended' for further

interim analysis time points, additional stages of accrual, considerations for multiple end-points, Bayesian approaches and approaches allowing three possible outcomes have been proposed.[4–8] In an era when most new anticancer agents had very limited efficacy and thus any evidence of tumour activity was sufficient to warrant future testing, the single-arm design provided a simple, direct and rapid method to assess an agent's activity. The fundamental assumptions underlying the single-arm phase II trial are that data from the previous studies provide an adequately robust estimate of the experience the patients from the current trial would have had were they not to have received the experimental therapy, and that the end-point selected for the trial represents a definitive measure of agent activity. Both these assumptions are suspect in the current era. Rapidly changing standards of care in many diseases in therapy, imaging and supportive care imply that historical data may not be reflective of the results observed in the current clinical practice. This is compounded by the fact that advances in tumour biology are segregating disease into marker-specific subtypes (e.g. based on Her-2 or K-ras status), for which historical data may be totally absent. In addition, new agents with novel mechanisms of action imply that standard end-points, such as tumour response, may not accurately capture an agent's overall patient benefit. The combination of these factors, in our opinion, questions the appropriateness of the single-arm trial at a fundamental level.

A second class of phase II trials includes a randomisation to two or more experimental arms, but no prospective control arm. The goal of such a trial, often referred to as a selection design, is to provide more robust data than are available through single-arm trials to select the most promising experimental regimen to compare to a prospective control in the phase III setting. Numerous potential options for such trials have been proposed,[9–11] all of which rely on an implicit comparison to historical controls in terms of the level of activity required to justify moving forward the most promising experimental regimen to the phase III setting. As these phase II trials are not designed to establish a new standard of care, little concern is paid to the true type I error (the false-positive rate), which generally does not account for the multiplicity of testing (through multiple-experimental arms), as well as to an error in the implicit historical control. While these selection designs do clearly provide much stronger comparative data to select a promising regimen amongst many possible phase III trials, the implicit comparison to historical controls confers the same disadvantages previously specified for the single-arm phase II trial.

The third class of phase II trials includes a randomisation to one or more experimental arms and a prospective control arm. This is the standard approach to drug development in other therapeutic areas, in which the phase II programme is often divided into phase IIa (a two-arm trial for proof of concept) followed by phase IIb (a dose-ranging trial to optimise the phase III design). Two sub-classes of these randomised controlled designs have been used: those that formally use the control arm in the efficacy determination, and those that use the control informally. In the latter design, the control arm data are assessed as to its similarity to historical controls; if the control arm produces results that are 'similar' to arm are considered valid. In our opinion, the inclusion of an 'informal' control arm serves a false purpose. These trials are inherently comparative, and if it is recognised that the go/no-go decision for phase III testing will be based on many factors and will not be dependent on traditional statistical significance ($p < 0.05$), then to ignore the power of randomisation to allow direct between-arm comparisons serves no valid scientific purpose. If the control arm is used in a formal efficacy comparison, the method of comparison is similar to that of a phase III trial, such as using a chi-squared comparison of response rates, or a log-rank comparison of progression-free survival (PFS) times. The formal use of the control group in the efficacy comparisons has been discussed by Rubinstein et al.,[12] Fleming and Richardson[13] and others.

4. When is a non-randomised phase II trial sufficient?

We believe non-randomised phase II trials in oncology should be the exception, not the rule. With that caveat, we acknowledge that such trials may be appropriate for trials in which the desired outcome (e.g. a partial response) will not occur in the absence of the investigational agent, and the rate of that outcome for existing agents or regimens is historically highly reliable. In this context, a positive phase II trial, defined as a response rate greater than some predefined threshold, would be evidence for activity, and a negative phase II trial would lead to the discontinuation of development for the specific indication.

Importantly, these criteria would never be met by many end-points often used in oncology clinical trials, such as 'clinical benefit or survival beyond some threshold (unless that threshold was inconsistent with the natural history of the disease). In addition, combination trials would generally not meet these criteria, unless the non-investigational components of the combination were known to be inactive in the target patient population. Finally, given the currently rapidly changing standards of care in many disease, once a new 'standard' has been defined in any given disease, the historical data obtained to date are invalid, due both to the availability of the new standard, and the fact that a new standard may change the treatment paradigm in that disease (e.g. poor prognosis patients who were previously untreated, thus not included in the historical control rates, may now become part of the treatment population).

Additionally, non-randomised trials should never be used if there is an uncertainty regarding the optimal dose. The maximally tolerated dose has been traditionally utilised for phase II evaluation, despite the clear evidence against the general validity of the 'more is better' paradigm.[14] For example, the randomised dose-ranging phase II trial of temsirolimus in renal cell cancer showed that doses of 25, 75 and 250 mg weekly were all equally effective, leading to a successful phase III trial at the lowest dose.[15,16]

We acknowledge that many experienced investigators continue to advocate the utility response rate as assessed in single-arm phase II trials. El-Maraghi and Eisenhauer[17] reviewed 89 trials of 19 agents, and concluded that the observa-

Given that the drugs with a 0% response rate in a single-arm phase II trial would not proceed to phase III testing, it is not surprising that no such agent achieved FDA approval. On the other hand, four of the six agents with the response rates under 10% were approved, including two agents with the response rates of less than 5% as single agents (cetuximab and sorafenib). The authors concluded that 'even low levels of response may be interesting' and acknowledged that larger phase II trials are required. However, the authors do not discuss an important implication of their recommendation, which is likely the further increase in the number of phase III trials that will fail to achieve their desired end-point if all the drugs with the response rates of 3–4% (as observed with cetuximab and sorafenib) underwent phase III testing. We feel that PFS provides a superior signal from which to assess the efficacy for such agents, which is clearly assessed optimally through a randomised trial.

Some authors have also suggested that the utility of single-arm trials using PFS or overall survival rates at a predefined landmark could be improved through the use of a model-based, comparison to historical controls. One example is a recent study by Korn and colleagues from five cooperative groups[18] where the authors reported a multivariate prognostic model for metastatic melanoma, which they proposed could be utilised in single-arm phase II trials, as an alternative to randomised trials. They suggest that only 63 patients would be required in a single-arm trial in this setting to have the same type I and II errors as a 220 patient randomised trial. While this model-based approach likely does provide some improvement compared to an unadjusted comparison to historical controls, it does not change our fundamental belief in the value of randomization: unless the model is perfect (which it never is), it is impossible to fully adjust the type I and II error for inter-trial variability, which in our opinion is the primary reason that single-arm phase II results are not replicated in phase III trials, as discussed further in the next section.

5. When does a non-randomised phase II trial lead to a phase III trial that fails to achieve success on its primary end-point?

As stated previously, the fundamental assumptions that could justify a single-arm phase II trial are that data from the previous studies provide an adequately robust estimate of the experience of the treated patients from the current trial were they not to have received the experimental therapy, and that the end-point selected for the trial represents a definitive measure of agent activity. Not surprisingly, when these assumptions are violated (which is common), phase III trials have a high rate of failing to demonstrate a significant improvement in the primary end-point from the experimental therapy.

From our perspective, the single greatest reason that uncontrolled phase II trials generate 'false' hope is the inter-trial variability in end-point success rates. Statistically, the failure to acknowledge the variability in the historical success rate inflates the nominal type I error level in the uncontrolled point, rapid advances in treatment outside the study protocol, supportive care, imaging modalities (to assess disease status) and other factors imply that patient outcomes will likely improve over time regardless of any new therapy. A likely even more dominant factor in the overall patient level success rate for a trial is the institutional composition of enrolling physicians and sites. Patient outcome has been repeatedly shown to associate with physician volume, and the patient mix at academic medical centres (which conduct the majority of phase II trials) differs widely from that in community practice. These three factors (statistical underestimation, patient outcome drift and patient selection) combine to make the true type I error rate for the primary end-point in uncontrolled phase II trials completely unknown (and unknowable).

Multiple additional factors contribute to difficulty in predicting successful phase III trials based on single-arm phase II trials. In single-arm trials of drug combinations, it is very difficult to distinguish the relative contribution of the standard component from the experimental agent. The choice of a different end-point in a phase II trial (such as tumour response) compared to phase III (where overall survival or PFS is likely to be primary) is problematic, as response rate has been repeatedly shown to be a poor surrogate for these time-to-event end-points.

The use of a control arm in a two-arm randomised phase II trial, comparing a single experimental regimen to a control, alleviates many of the issues above; however, several problematic issues remain. As regimen doses and schedules are typically established through small phase I trials, with few (typically <12) patients treated at the MTD, limited information is available in most cases on the dose–efficacy and dose–toxicity relationships for the new agent. One potential cause of the failure in phase III trials is the error in the dose – either too high resulting in an unacceptable toxicity or too low resulting in an inadequate efficacy. This risk can be minimised with a multiple-arm randomised phase II trial, where several dose levels are explored initially, and adaptively removed from the trial based on prespecified criteria.[20]

6. What are the considerations in the design of randomised phase II trials?

Although there has been a debate about the value of formal statistical comparisons in phase II trials, we feel strongly that such comparisons are appropriate, with the caveat that phase II trials do not necessarily need to provide reliable definitive comparisons at a traditional two-sided type I error of 0.05. Given that the purpose of any phase II trial is to determine whether to proceed with further agent development, there is only one outcome of interest, superiority of one or more experimental arms to the control. In this scenario, we believe a one-sided testing framework is appropriate. Given the need for phase II trials to be as small as possible and that a phase III trial will be required to confirm the efficacy in most cases, standard type I error rate control at the 0.05 level is not necessary. We and others[12] propose that a one-sided test of the null hypothesis that the true primary outcome is no different between treatment and control with a false-positive rate of

trials were required to be justified by a randomised phase II trial with $p < 0.20$, this would lead to a phase III success rate of 80%, a significant improvement from the status quo. In contrast, in the uncontrolled phase II setting, the true false-positive rate is unknown.

The use of this higher type I error remedies one common criticism of randomised phase II trials, that of very low statistical power.[21] In the case of a multiple-experimental arm phase II randomised trial with a control, appropriate type I error control is necessary to maintain an overall type I error rate of 20%, either through adjusting the alpha level for each comparison of control versus treatment, or by comparing the experimental treatments to each other to identify the best, then comparing that regimen to control. Sample size considerations typically also dictate a type II error rate of 0.20, for a modest effect size (e.g. a hazard ratio of 1.5 in a time-to-event end-point, or a difference in response rates of 15–20%), although a lower type I and II error rates (e.g. 0.10) should also be considered if resources allow.[12]

The choice of primary end-point in a randomised phase II trial depends on multiple factors. Ideally, a phase II screening trial should be a similar as possible to the subsequent phase III trial, including choice of primary end-point. Given the movement in advanced disease trials towards PFS as a phase III primary end-point,[22] this is increasingly possible. Although overall survival can also be considered as the primary end-point for phase II and/or III testing, this would not permit the use of crossover designs, especially useful in the phase II setting.

In our opinion, the most promising end-points for randomised phase II trials involve a comparison of a primary outcome measure at a single time point between the treatment and control groups. This outcome measure could be a tumour assessment at a fixed time point (e.g. change in absolute tumour size from baseline at 8–12 weeks following treatment initiation), or a rate of overall survival or PFS at a fixed time point (for example, 3 or 6 months). Such end-points facilitate independent radiologic review, eliminate subtle differences in scanning frequencies, simply patient scheduling and can be chosen to represent clinically meaningful time points. Ongoing research in multiple disease areas is examining these and other end-points to identify which end-points optimally predict phase III success. In the future, newer methods such as functional imaging or assessment of circulating tumour cells may allow even earlier assessment of disease status.

The validity of almost all end-points is improved by blinding, a tool infrequently used in oncology trials, particularly phase II trials. This is less important when overall survival is the primary end-point, but critical for end-points such as PFS, since independent radiological review can address bias in measurement but not bias in timing of radiological studies. When a trial is blinded, one can consider the use of novel end-points that incorporate patient-reported outcomes (e.g. time to symptomatic progression). In addition, blinding will result in a more robust analysis of toxicity attributable to the investigational agent, as it controls for the 'placebo effect'. Blinding can be logistically complicated, as it requires the availability of either an indistinguishable placebo (for oral agents) or [illegible] (for intravenous or subcutaneous

for all randomised comparative phase II trials, we acknowledge that the final decision must consider all of these issues.

The use of blinding also permits the use of the randomised discontinuation (or withdrawal) design, in which all eligible patients initially receive active treatment for a predefined run-in period. In this design, patients with stable disease at the end of a run-in period proceed to the randomised portion of the study (the primary analysis), in which they are randomised to continued active treatment or placebo. To minimise the risk to patients randomised to the placebo arm, all patients are offered crossover at progression, and those patients who had been randomised to placebo are restarted on open-label active treatment (those patients who had been randomised to active treatment are withdrawn). This design is particularly useful for the trials of the new agents as a single agent where it is hypothesised that the response rate is low, but that the drug has a clinically significant antitumour effect. We strongly feel that this design is superior in this setting to the use of a large single-arm trial, as a positive trial demonstrating a low response rate does not provide sufficient information to inform a robust decision to proceed to phase III testing.

7. Conclusions

Randomisation is greatly underutilised in early clinical trials in oncology. The use of randomised trials allows complete flexibility in the choice of end-points, particularly if blinding can be incorporated. This technique is the most powerful and reliable technique for distinguishing the effect of a drug from a placebo, a necessary predicate for a successful phase III trial. Given the large number of agents now available for clinical testing, the costs of phase III trials, and the limited patient and financial resources, available for such testing, the increased use of randomised phase II trials provides a rational way to move new agents forward. Although end-points based on RECIST criteria can and should be utilised in randomised phase II trials, we do not believe that revision of the RECIST criteria will lead to any fundamental impact on improving drug development decisions in the absence of randomised clinical trials at the phase II stage of the drug development.

Conflict of interest statement

None declared.

REFERENCES

1. Sheiner LB. Learning versus confirming in clinical drug development. *Clin Pharmacol Ther* 1997;**61**(3):275–91.
2. De Ridder F. Predicting the outcome of phase III trials using phase II data: a case study of clinical trial simulation in late stage drug development. *Basic Clin Pharmacol Toxicol* 2005;**96**(3):235–41.
3. Gehan EA. The determination of the number of patients [illegible]

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.