DOI: 10.1093/jnci/djr218
Advance Access publication on June 27, 2011.

REVIEW

Randomized Phase II Trials: A Long-term Investment With Promising Returns

Manish R. Sharma, Walter M. Stadler, Mark J. Ratain

Manuscript received August 10, 2010; revised May 11, 2011; accepted May 17, 2011.

Correspondence to: Mark J. Ratain, MD, Department of Medicine, University of Chicago, 5841 S. Maryland Ave, MC 2115, Chicago, IL 60637-1470 (e-mail: mratain@medicine.bsd.uchicago.edu).

Given the multitude of novel anticancer drugs and the limited resources available to study them, phase II trials should identify drugs with the highest probability of succeeding in subsequent phase III trials. Currently, single-arm phase II trial results are interpreted relative to historical control subjects, introducing selection bias and confounding that may limit the validity of the conclusions. The rate of success (defined as a statistically significant difference between arms) in phase III oncology trials is only 40%, suggesting that current phase II trials are insufficiently informative. However, simulation studies suggest that randomized phase II trials would have lower error rates and greater predictive power for phase III results. Randomized phase II trials may also be more informative than single-arm phase II trials because of the hypotheses being tested, the variety of possible endpoints, and the opportunities for biomarker discovery. There are a wide variety of randomized phase II designs that can be used, including the randomized discontinuation design, the delayed-start design, adaptive (Bayesian) designs, selection designs, and phase II/III designs. The barriers to widespread adoption of randomized phase II trials include time to completion, sample size considerations, and ethical concerns, but none are insurmountable. We conclude that randomized phase II trials are a worthy investment considering finite patient and financial resources and should be the rule rather than the exception for evaluating novel therapies in oncology.

J Natl Cancer Inst 2011;103:1093–1100

Phase II trials are undertaken to determine whether a novel drug (or combination) is promising enough to justify a definitive phase III study for efficacy, often referred to as making a go/no-go decision. The threshold for "promising enough" is not easily defined, but the decision about where to set the threshold is likely to involve several factors. One such factor is the disease, because oncologists are more likely to accept a lower threshold in rare cancers with few or no treatment alternatives. Another factor is the patient population, because oncologists are more likely to accept a lower threshold in the case of patients with incurable disease and limited life expectancy, as compared with relatively healthy patients receiving adjuvant therapy. A final factor, and perhaps the most important one, is the allocation of finite patient, investigator, and financial resources that are available for oncology drug development.

The growing number of new oncology drugs, coupled with the increasing cost of drug development and the high rate of failure in phase III trials, suggest that current phase II trials are not sufficiently informative. Since 1995, the Food and Drug Administration (FDA) has approved 63 new anticancer drugs (http://www.centerwatch.com/drug-information/fda-approvals/drug-areas.aspx?AreaID=12), and a 2009 report by the Pharmaceutical and Research Manufacturers of America estimated that there were 861 additional anticancer drugs in clinical trials or under FDA review (http://www.healthinfoispower.files.wordpress.com/2009/04/phrmacancer.pdf). The majority of drugs in development are molecularly targeted (ie, developed based on activity against a specific target), in contrast to cytotoxic chemotherapy drugs (the majority of approved agents), and multiple sponsors are often developing drugs against the same (or similar) targets. Based on a public database of all drugs that went into clinical trials for the first time between 1989 and 2002, Adams and Brantner (1) found that cancer drugs (n = 681) had a total expected capitalized cost per new drug of 1.042 billion dollars (in 2000 US dollars), which included spending on regulatory submissions and marketing and not just on clinical trials. Furthermore, an average of more than 8 years in phase I–III trials was spent in development of those drugs (1). Although some degree of failure is to be expected at each stage of drug development, it is sobering that only 5% of oncology drugs make it from first-in-human trials through registration, including a 60% failure rate in phase III trials (2). The main driver of failure in phase III oncology trials is lack of efficacy compared with placebo or an existing standard of care (3), suggesting that we are making inappropriate go/no-go decisions at the end of phase II. Maitland et al. (4) analyzed all phase II combination chemotherapy trials published in 2001 and 2002 and found that despite 72% of them having been reported as positive, the likelihood of a subsequent trial showing an improvement in standard of care within 5 years was only 3.8% (4). Given the finite resources available to study the large number of drugs in the pipeline, it is increasingly clear that the status quo is not sustainable in the long term.

It is therefore reasonable to consider efforts to improve the predictive value of phase II trials. Compared with other specialties,

phase II oncology trials are less likely to use control subjects (5), and there are abundant examples to suggest that this may be a causative factor in the lower observed success rate in phase III trials (6). The role of randomization in phase II oncology trials remains controversial, with some experienced investigators supporting the continued use of single-arm phase II trials, in which all participants receive the study drug or regimen (7,8). El-Maraghi and Eisenhauer (9) reviewed 89 phase II trials that studied 19 targeted drugs, only three (3.4%) of which used randomization between a control arm (placebo or standard therapy) and an experimental arm. In this review, we will summarize the scientific evidence and theoretical principles in favor of randomized phase II trials, as they pertain to the goals of increasing the efficiency and success rate of the drug development process. We will discuss the various types of randomized phase II designs and explore the barriers to widespread adoption of randomized phase II trials. For the purposes of this review, "negative" phase III trials are those that fail to show a statistically significant difference between arms with respect to the primary endpoint, whereas "positive" phase III trials are those that show a statistically significant difference between arms with respect to the primary endpoint.

The Evidence in Favor of Randomized Phase II Trials

The results of single-arm trials are typically interpreted relative to data from historical control subjects, past study participants with similar characteristics to the study population in question. The appropriateness of using historical control subjects is highly dependent on the endpoint that is being used (ie, the metric of success), as well as the patient population that is being studied. For example, a single-arm phase II trial may be appropriate in a disease setting for which there are no active therapies and for which the metric of success is a high rate of marked tumor shrinkage (ie, response rate [RR] according to Response Evaluation Criteria in Solid Tumors [RECIST]). In other words, a drug meeting this endpoint may be worthy of further investigation, because there is sufficient confidence that the historical control subjects had essentially no RECIST responses with observation alone.

In most cases, however, we are comparing a new drug to historical control subjects who were treated with some type of active therapy or we are using endpoints with much greater historical variability, such as the proportion of patients who are progression free at an arbitrary time point or time-to-event endpoints (progression-free survival [PFS] or overall survival). In such cases, the validity of conclusions from single-arm trials based on historical control subjects is limited by two classic epidemiological factors, selection bias and confounding. Selection bias refers to the phenomenon that current study participants may be different from historical control subjects in ways that affect the outcome of interest. Differences that might bias toward a positive result include baseline patient factors, such as younger age or better performance status; baseline disease factors, such as smaller tumor burden or less aggressive tumor biology; or provider factors, such as size and other characteristics of the treating centers. The same factors, if they are different in the opposite direction, would bias toward a negative result. Korn et al. (10) validated this concept by identifying a number of patient-specific and trial-specific prognostic variables (performance status, visceral metastases, sex, and exclusion for brain metastases) that influenced the 1-year overall survival rate in phase II trials of metastatic melanoma and showing that between-trial variability could be essentially eliminated by controlling for these variables.

Confounding refers to the phenomenon that current study participants may have a different (better or worse) outcome than historical control subjects because of factors during the treatment period that are not related to the quality of the intervention. For example, if drug X is actually no better than the standard of care but patients receive better supportive care during the treatment period, the results with respect to the primary endpoint may appear better than those of historical control subjects. The impact of supportive care should not be underestimated, because it was recently shown that overall survival was statistically significantly longer among patients receiving early palliative care along with chemotherapy for non–small cell lung cancer (11). Another potential confounder is the availability of subsequent effective treatments, as exemplified by the recent success with v-raf murine sarcoma viral oncogene homolog B1 (BRAF) inhibitors in BRAF-mutated melanoma. Patients with BRAF mutations who are initially treated with therapies other than BRAF inhibitors would be expected to have improved overall survival than prior historical control subjects because of the subsequent benefit from the BRAF inhibitor. Patients without BRAF mutations are obviously not representative of the overall population, and the survival model that has been proposed by Korn et al. (10) for screening new agents did not evaluate this important covariate, making this model invalid in the era of targeted therapy for melanoma. The impacts of selection bias and confounding are impossible to quantify when making comparisons with historical control subjects and can even be difficult to identify because most published reports of single-arm trials do not clearly specify the historical data that were used to formulate the null hypotheses (12). Acknowledging the shortcomings of historical control subjects, some investigators are now conducting single-arm phase II trials that include a simultaneous but smaller control arm, such that both arms are compared with historical control subjects but not to each other (because of inadequate statistical power). Although this is somewhat reassuring if the control arm and historical control subjects have similar outcomes, the precision is not sufficient to ensure comparability (13), and it is unclear what to do if they have markedly different outcomes.

The use of historical control subjects leads to a high risk of "false positives," single-arm phase II trials that appear promising but are followed by negative randomized phase III trials. Although examples can be found in all tumor types, there is perhaps no better collective example than in the field of advanced pancreatic cancer. In the last decade, eight drugs were studied in combination with gemcitabine in single-arm phase II trials and found to be promising compared with historical control subjects who received gemcitabine alone (14–21). Definitive phase III trials, however, have been disappointingly negative for every one of these combinations (22–29), leading to no appreciable change in the standard of care over this period. The only exception, erlotinib, was never studied in combination with gemcitabine in a dedicated phase II

trial, and the benefit of the combination in the phase III trial (0.33 months increase in overall survival, compared with gemcitabine alone) is of dubious clinical relevance and has never been replicated (30). These nine phase III trials in pancreatic cancer also demonstrate the problem with variability of historical control subjects, because the median overall survival for patients receiving gemcitabine alone ranged between 5.4 and 7.2 months, despite the trials being conducted in very similar patient populations. Walter et al. (31) have pointed out that acute myeloid leukemia is another disease in which progress has been impeded by a high false-positive rate in phase II trials and have proposed that randomization is a key to solving this problem.

Because single-arm and randomized phase II trials are rarely conducted simultaneously or sequentially for the same drug or combination, investigators have used simulation techniques to compare the two designs. Tang et al. (32) simulated and compared error rates in single-arm vs randomized phase II trials, using both statistical models and individual patient data from a large phase III trial in colorectal cancer. For single-arm trials, they found that random and systematic variation in historical control data could increase the type I (false positive) error rates by two- to fourfold. They also found that the statistical power of single-arm trials was sensitive to unanticipated factors, such as the selection of patients from high-volume vs mid- or low-volume treatment centers. In a similar type of study, our group (33) resampled data from a large phase III trial in renal cell cancer to simulate and compare various phase II designs and endpoints based on a computed tomography scan at 6 weeks. We found that randomized phase II designs with a continuous change in tumor size endpoint had greater predictive power than a conventional single-arm design for the known phase III result. Stewart et al. (34) used survival times from patients with non–small cell lung cancer to simulate randomized trials with hypothetical novel therapies that quintupled or doubled survival in only 10% of patients who express a specific target, with no effect on the remaining 90% of patients. They found that randomized trials with a large number of unselected patients would incorrectly conclude that the drug had no benefit, whereas randomized trials with a small number of patients selected for the target would correctly conclude that the drug had a benefit in those patients. Although Stewart et al. (34) did not simulate any single-arm designs, one implication of their work is that single-arm phase II trials could easily detect drug benefit if the patient population is preselected for the drug's target, as illustrated by examples (Table 1) of first-in-class targeted drugs that have shown remarkably high RRs in phase Ib or phase II studies (35–38). Even compared with historical control subjects, the promising nature of such drugs cannot be called into question, and single-arm phase II studies would adequately demonstrate their benefit in these biomarker-defined populations or subpopulations. We would caution, however, that we often do not know the relevant target of a drug or do not have a reliable predictive biomarker when it is being developed and studied in clinical trials (eg, sorafenib, originally developed as a raf inhibitor). There are many examples of drugs for which selection of patients based on a therapeutic target has not resulted in dramatic RRs suggestive of clinical efficacy, for example, fms-related tyrosine kinase 3 (FLT3) inhibitors for acute myeloid leukemia patients harboring activating FLT3 mutations (39).

Theoretical Advantages of Randomized Phase II Trials

As Ratain and Karrison (40) point out, single-arm and randomized phase II trials with a true comparator arm are fundamentally testing different hypotheses. In a conventional, two-stage single-arm phase II trial (41), the endpoint is an objective RR, which is defined as the proportion of patients responding to the drug according to RECIST (42). The null hypothesis is that RR is less than a certain threshold based on historical data, say 5%, whereas the alternative hypothesis is that RR is somewhat higher, say 20%. Rejection of the null hypothesis means that RR was greater than 5%, but generally does not establish that RR is 20% or higher. With regard to the alternative hypothesis, what can be concluded is that the observed data are not inconsistent with such a value. Alternatively, under the usual one-sided hypothesis-testing framework used in randomized phase II trials (ie, H_0: $\delta \leq 0$ vs H_A: $\delta > 0$, where δ is the true difference), the null (H_0) and alternative (H_A) hypotheses represent the only possible truths. If we reject the null hypothesis that the new drug is no better than the existing standard of care, then we accept the alternative hypothesis that the drug is potentially better than the existing standard of care (ie, the probability is sufficiently high to warrant further testing). In other words, single-arm trials only screen out very ineffective drugs (ie, those with an RR less than a certain threshold), whereas randomized designs screen in potentially effective drugs (ie, those that may surpass the existing standard of care, without establishing superiority in a scientifically rigorous fashion).

Another important difference between single-arm and randomized phase II trials involves the types of endpoints that are typically used. Single-arm trials generally use binary endpoints, such as RR or the rate of PFS at a certain time point. Alternatively, randomized trials typically involve time-to-event endpoints (PFS, time to progression, and overall survival), although binary endpoints can

Table 1. First-in-class targeted drugs that have resulted in high response rates in phase Ib or phase II trials

Drug	Disease	Target*	Response rate in phase Ib/II, %	Reference
PLX4032	Melanoma	V600E BRAF mutant	81	36
Crizotinib	Non–small cell lung cancer	EML4-ALK	57	37
Imatinib	Chronic myeloid leukemia—chronic phase	BCR-ABL	95 (hematologic)	38
Imatinib	Gastrointestinal stromal tumor	KIT	54	39

* BCR-ABL = breakpoint cluster region fusion with V-abl Abelson murine leukemia viral oncogene homolog 1; BRAF = v-raf murine sarcoma viral oncogene homolog B1; EML4-ALK = echinoderm microtubule associated protein like 4 fusion with anaplastic lymphoma receptor tyrosine kinase in non–small cell lung cancer; KIT = v-kit Hardy–Zuckerman 4 feline sarcoma viral oncogene homolog.

also be used in this setting. In the review by El-Maraghi and Eisenhauer (9), four targeted drugs that were eventually approved by the FDA had RR less than 10%, and two of them had RR less than 5%. As we continue to develop and study growth inhibitory drugs, it is potentially important to use time-to-event endpoints, so that active drugs with a low RR are not overlooked. Furthermore, the selection and justification of null and alternative hypotheses for single-arm trials becomes increasingly difficult as we test more combination therapies and therapies in diseases in which the standard of care is moderately successful. Randomized phase II trials also encourage the development and use of alternative endpoints, such as the continuous endpoint of change in tumor size that was proposed by Karrison et al. (43). Continuous response endpoints do not require a predefined threshold for an objective response and can be normalized by analysis on a logarithmic scale, but they are not feasible in single-arm trials because of the absence of continuous historical data. Of course, continuous response and other alternative endpoints are only useful once they have been validated as surrogates for the primary outcome of interest.

Yet, another advantage of randomized phase II trials over single-arm trials is the ability to study biomarkers in a scientifically rigorous fashion. In randomized phase II trials, blood and/or tissue samples can be collected and studied for biomarkers that may reflect pharmacodynamic effects or correlate with findings regarding tumor response and time-to-event endpoints, compared with samples from patients who did not receive the drug. Such studies are scientifically less valid in single-arm trials because there is no population of control subjects who did not receive the investigational drug. Biomarker studies in phase II trials are valuable because they can confirm the drug's relevant target, and they may be validated or used to guide the selection of patients in the definitive phase III trial (44).

Types of Randomized Phase II Trials

A wide variety of designs have been used for randomized phase II trials, and selection of the most appropriate design involves considering the drug, the disease, the patient population, the endpoint, and the overall objectives of the study. The simplest design is a two-arm trial with up-front randomization (Figure 1,A) to the new drug or combination vs the existing standard of care (or placebo). Minor variations include the use of multiple arms to allow for dose ranging of the investigational drug or the use of an unbalanced randomization (2 : 1 or 3 : 1) to boost accrual with only a modest reduction in statistical power. We will discuss five alternative types of randomized phase II designs as follows: the randomized discontinuation design, the delayed-start design, adaptive (Bayesian) designs, selection designs, and phase II/III designs (Figure 1, B–F).

In the randomized discontinuation design (45), all patients receive the study drug for a run-in period (Figure 1, B). At the end

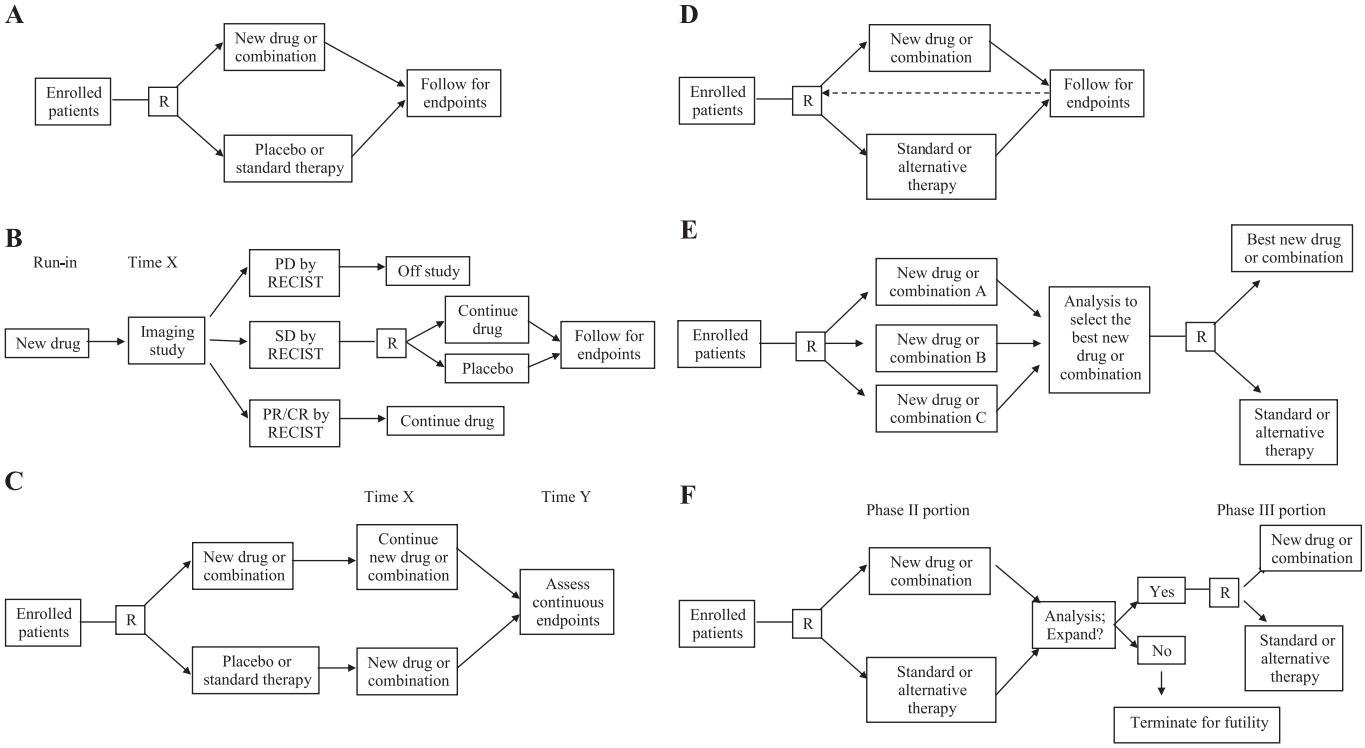

Figure 1. Schematic illustrations of various randomized phase II trial designs. **A**) Up-front randomized design. **B**) Randomized discontinuation design. Time X is the prespecified end of the run-in period, which is typically in the range of 8–12 weeks for most drugs and tumor types. **C**) Delayed-start design. Time X is the prespecified end of the placebo period (for those randomly assigned to placebo). Time Y is the pre specified point at which continuous endpoints for the two arms are compared. The difference is the disease-modifying effect. **D**) Adaptive (Bayesian) randomized design. The design is the same as the up-front randomized design, but the **dashed line** indicates that available information regarding outcomes is used to adjust the randomization scheme in real time. **E**) Selection ("pick the winner") design. **F**) Phase II/III design. The analysis of the phase III trial includes patients from both the phase II and phase III portions. CR = complete response; PD = progressive disease; PR = partial response; R = randomization; RECIST = response evaluation criteria in solid tumors; SD = stable disease.

of the run-in period, an imaging study is performed and patients are categorized as having stable disease, progressive disease, or a partial/complete response according to RECIST. Those with progressive disease come off the study, whereas those with partial/complete responses continue with the study drug. Those with stable disease are randomly assigned to continue the study drug or switch to a placebo in a blinded fashion. Patients who progress after randomization can be unblinded and offered an opportunity to cross over and resume the study drug if they had been taking placebo. The primary outcome is typically PFS among the randomly assigned patients. This type of phase II design successfully demonstrated that sorafenib was active in renal cell cancer (46), leading to the definitive phase III trial (47) and FDA approval of the drug. Alternatively, a single-arm phase II design with a RECIST RR endpoint would likely have concluded that the drug was inactive, based on the observed RR (by independent review) of 2% in the phase III trial (47). The randomized discontinuation design (Figure 1, B) is especially useful for phase II trials of slow-growing cancers, or in cases in which the drug is expected to have an effect in a subset of the population, but no validated tool is available to select these patients in advance. Early stopping rules for both efficacy (a high RR during the run-in period) and futility (randomization rate below a certain threshold) should be used. The disadvantages are that the total sample size required can be very large if the randomization rate is low, and the results cannot be readily compared with a standard randomized controlled trial.

The delayed-start design (Figure 1,C), which was used in a recent study of Parkinson disease (48) and described as widely applicable to chronic progressive diseases in the accompanying editorial (49), has the potential to be used in phase II oncology trials. In the first phase of this design, patients are randomly assigned to the study drug or placebo and are followed for a period of time with regular assessments. In the second phase of this design, patients who were initially randomly assigned to placebo switch to the study drug, whereas patients who were randomly assigned to the study drug continue taking it. After some period of time in this second phase, the outcome of interest is measured and compared between the patients who were initially assigned to the study drug and those who were initially assigned to placebo. The difference between the two groups is the disease-modifying effect of the drug. Although this design was developed to follow the effect of the study drug on symptom progression in a neurological disease, it could easily be used to follow the effect of an anticancer drug on tumor progression (using longitudinal data on tumor size from serial computed tomography scans). Advantages include the ability to detect a benefit for drugs that slow down tumor growth without causing tumor regression and the statistical power achieved by having longitudinal data on a subset of patients before and after starting the active drug. Disadvantages include the need to start a subset of patients on placebo, which might bias toward selection of patients with relatively indolent disease, and the inability to use conventional endpoints such as RR or PFS.

An interesting recent trend in phase II trial design has been the increasing use of adaptive (Bayesian) designs (Figure 1, D). There are many statistical approaches for these designs, but they generally use interim data available at the time of each new enrollment to assess outcomes and alter the randomization probabilities in favor of the treatment that is resulting in a comparatively better outcome (hereafter, "adaptive randomization"). Cheung et al. (50) described how adaptive randomization could be done for a phase II trial with a time-to-event (eg, PFS) outcome, using a method that also accounts for baseline prognostic covariates. Biswas et al. (51) described how adaptive designs have been used in a variety of phase II trials at the MD Anderson Cancer Center, including an example of adaptive randomization in a phase II trial and an example of continuous monitoring for efficacy and toxicity in a combined phase I/II trial. Lee et al. (52) studied adaptive randomization in the context of targeted drug development and concluded that this design is the most suitable for trials of multiple targeted drugs with multiple biomarkers of interest. The principle advantage is the ability to randomly assign as many patients as possible to the most promising treatment, whereas the principle disadvantage is the large investment of statistical resources in performing so many interim analyses. There is also a potential for bias if the nature of the patient populations shifts over time in a way that is not captured by the covariates.

In selection designs (Figure 1, E) often referred to as "pick the winner" designs and initially described by Simon et al. (53), the best of several experimental therapies is selected for further comparison to the standard of care. In the first stage, patients are randomly assigned to one of several experimental arms. Following an interim analysis, the experimental arm with the greatest efficacy based on the primary endpoint (without needing to be statistically superior to others) is selected for head-to-head comparison with the control arm in the second stage, provided that it exceeds a predefined threshold based on historical data. Liu et al. (54,55) proposed statistical methods that would allow overall survival (or PFS) to be used as an endpoint for such a design and demonstrated that these could be done with reasonable sample sizes and adequate statistical power for detecting a difference between the control arm and the best experimental arm. The principle advantage of this design is that it maximizes efficiency by testing several experimental therapies simultaneously, whereas the disadvantage is that sponsors may be reluctant to participate in a trial that compares their agent to that of a competitor.

Finally, there is growing support for the concept of conducting combined randomized phase II and phase III trials, which can be referred to as a phase II/III design (Figure 1, F). In this design, an interim analysis is performed after the randomized phase II portion of the trial, and the decision to expand to a phase III trial is made on the basis of these results. This approach was first described for binary outcomes by Storer (56), who proposed that only the experimental arm would be compared with a historical benchmark (rather than to the control arm) after a prespecified number of patients to determine whether or not to expand. A sequential Bayesian approach to the phase II/III design was later proposed by Inoue et al. (57), in which the decision to stop early, continue, or expand to a phase III trial is continually assessed during the phase II portion. The primary endpoint for the phase II portion could be binary or continuous (eg, time to event) and would not necessarily have to be the same as the primary phase III endpoint. Since targeted anticancer therapies may benefit identifiable subpopulations, the expanded phase III portion could involve the whole population, a subpopulation, or both the whole population and subpopulation as coprimary populations (58). The advantages of

this design are that the time gap between the end of a promising phase II trial and the beginning of a definitive phase III trial is eliminated and that the patients in the phase II portion count toward the accrual goal for the phase III trial. A disadvantage is that coordination of multiple centers that will be involved in the phase III trial is necessary at the start of phase II. Another disadvantage is that this approach does not allow time for dose ranging, identification of which tumor types are of greatest interest for phase III development, or identification of biomarkers by correlative analysis of phase II samples.

Barriers to Widespread Adoption of Randomized Phase II Trials

Barriers to conducting randomized phase II trials are present for all parties involved in drug development such as industry and government sponsors, investigators, institutional review boards, and patients. Industry sponsors and academic investigators are motivated to develop drugs quickly, which may bias them toward designing single-arm phase II trials that can be completed sooner. Although ultimately trying to develop profitable drugs, industry sponsors are subject to the temperament of stockholders and private investors who may reward the company for a promising result in a single-arm phase II trial. Academic investigators are motivated by professional advancement, which is inherently linked to the presentation and publication of completed trials. All parties, however, have substantially more to gain in the long term from well-designed and informative randomized phase II trials. A randomized phase II trial that is negative when its single-arm counterpart may have been promising (compared with historical control subjects) saves industry sponsors large amounts of money that might otherwise be spent on negative phase III trials. Likewise, a randomized phase II trial that provides valuable information to the oncology community will bring greater publicity to the investigators than a single-arm trial, opening up opportunities to conduct similar studies for other drugs and definitive phase III trials for promising drugs.

The major criticism of randomized phase II trials is that they involve much larger sample sizes than their single-arm counterparts, thereby requiring more time to complete and more resources invested. Although this is true, it is a classic example of the old saying that "you get what you pay for." As pointed out previously, the hypotheses being tested are different in the two designs, with the randomized design allowing for a more meaningful conclusion about the drug. Furthermore, the higher false-positive rate associated with single-arm trials leads to negative phase III trials that could have been avoided, a much greater investment of time and resources than the randomized phase II trial would have required. Finally, as Rubinstein et al. (13) point out, the sample size for randomized phase II trials can be limited by liberalizing the statistical parameters for type I and type II error. A one-sided type I error rate (α) of 0.10 and type II error rate (β) of 0.15 (statistical power of 85%) are reasonable statistical parameters for an exploratory phase II study. What would these assumptions mean for the subsequent phase III success rate? Suppose that out of 1000 drugs tested, 200 are truly effective in the target population and 800 are not. Randomized phase II trials with $\alpha = 0.10$ and 85% statistical power would yield $0.10 \times 800 = 80$ false positives and $0.85 \times 200 = 170$ true positives. Thus, out of the 250 drugs brought to phase III with the assumption of 90% statistical power in the phase III trial, $170 \times 0.90 = 153$ (61%) out of 250 would succeed, a substantial improvement over the current phase III success rate of 40%. Of the 200 truly effective drugs, 47 would fail in either phase II or phase III, a false-negative rate of 24%. Using the same $\alpha = 0.10$ and statistical power of 85%, and assuming a nonadaptive design and 1 : 1 randomization, a randomized phase II trial with a PFS endpoint targeting a hazard ratio of 1.75 (eg, an increase in PFS from 4 to 7 months) would require 69 randomly assigned patients with events across both treatment arms, whereas the same trial with a hazard ratio of 1.5 (eg, an increase in PFS from 4 to 6 months) would require 131 randomly assigned patients with events. Although these sample sizes are not trivial, they are feasible for phase II trials conducted through consortia and/or multiple centers.

Another criticism of randomized phase II trials is that they have a substantial risk of false negatives, which may result in the exclusion of potentially effective drugs from further testing. This risk is highest in cases in which the drug's true target is only present in a small subpopulation of patients while the study is being conducted in an unselected patient population. Although this is a real concern, experienced investigators would identify a group of patients with exceptional outcomes and seek to design additional studies aimed at identifying biomarkers that correlate with these outcomes. Moreover, nonrandomized studies in unselected patient populations are also susceptible to false negatives resulting from a very small subset of highly responsive patients.

Investigators may also be reluctant to conduct randomized phase II trials because of the greater complexity and choice in design. The sample size for a single-arm trial can easily be calculated online. In contrast, randomized phase II designs require more initial effort and, in the case of adaptive designs, a great deal of flexibility, which may be intimidating to some investigators.

From the institutional review board and patient perspective, there have been growing concerns that it may not be ethical and/or desirable to withhold promising treatments from patients by forcing them to enroll in randomized trials. A recent article in the *New York Times* (59) brought this issue to light by highlighting two cousins with BRAF-mutated melanoma in a randomized trial for PLX4032; one cousin received the drug, and the other received a standard therapy. In response, we would argue that the definition of equipoise must be based on all available data. If the data suggest that a drug's RR far exceeds that of available therapies, as with PLX4032 and other drugs (Table 1), then a randomized phase II trial may be unnecessary. In most cases, however, there is true equipoise regarding whether or not the drug is better than existing therapies in the population being tested, and methods for selecting patients more likely to respond are not available. Expectations from the sponsor, the investigator, or the patient regarding the potential promise of the drug should not cloud the judgment about trial design that should be based on scientific facts. In fact, it would be reasonable to suggest that the investigator has an ethical obligation to avoid enrolling patients in a negative phase III trial that could have been made unnecessary by a randomized phase II trial. Finally, there are a number of elements of trial design that can minimize patient exposure to a less active treatment. One option is to use unbalanced randomization when feasible, although this will

increase the total sample size. In addition, randomized trials should always include early stopping rules for futility as well as efficacy and should allow crossover to the active treatment arm when PFS is the primary endpoint. Randomized discontinuation designs allow all patients to start out on the investigational treatment and minimize the number who are randomly assigned to placebo by restricting randomization to the stable disease group. Trials with adaptive features are also appealing, because they use real-time data to randomly assign more patients to the "winning" arm without compromising the statistical power of the trial, if performed in the manner described by Cheung et al. (50). Selection designs and phase II/III designs maximize the efficiency of the drug development process for all patients, and the latter may alleviate investigator concerns about equipoise in phase II because the patients enrolled would, if the investigational therapy is promising, contribute to a definitive phase III result.

Conclusions

Given the evidence and theoretical advantages in favor of randomized phase II trials, they should be the rule rather than the exception when it comes to evaluating new drugs in oncology. The weight of opinion among experts is moving in this direction, as suggested by the recent recommendations of the Investigational Drug Steering Committee regarding trials with time-to-event endpoints (60) and by an editorial by Cannistra in the *Journal of Clinical Oncology* (61). The variety of randomized phase II designs gives investigators flexibility to choose the best fit for a certain drug, disease, and patient population. As the number of drugs in development continues to grow, the use of randomized phase II trials will reduce the rate of future negative phase III trials (assuming that truly effective agents or combinations are available) and, in doing so, optimize the use of limited patient and financial resources. Single-arm phase II trials will continue to play a role in drug development, but their use should be limited to the following situations: 1) monotherapy trials for diseases in which there is no standard therapy, in cases in which early data suggest an RR that is dramatically higher than those with available therapies (as for the drugs listed in Table 1), or in cases in which the predefined goal of the trial is to identify a subset of patients with a profound tumor response (eg, >50% tumor reduction); or 2) patient populations or subpopulations for which there are no effective therapies (including investigational agents) and for which robust and contemporaneously validated historical control subjects exist (We are not aware of any current examples that meet this latter criterion.).

The routine use of randomized phase II trials will be a cultural shift for oncologists, sponsors, and patients requiring that we prioritize the pursuit of long-term gains (ie, approved and available drugs) over the appeal of short-term gains (ie, promising drugs that later fail to improve outcomes in phase III). If we make the investment, randomized phase II trials have the potential to usher in an era of unprecedented success in oncology drug development.

References

1. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff (Millwood).* 2006;25(2):420–428.
2. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov.* 2004;3(8):711–715.
3. Elias T, Gordian M, Singh N, et al. Why products fail in phase III. *In Vivo.* 2006;24:49–54.
4. Maitland ML, Hudoba C, Snider KL, Ratain MJ. Analysis of the yield of phase II combination therapy trials in medical oncology. *Clin Cancer Res.* 2010;16(21):5296–5302.
5. Michaelis LC, Ratain MJ. Phase II trials published in 2002: a cross-specialty comparison showing significant design differences between oncology trials and other medical specialties. *Clin Cancer Res.* 2007;13(8):2400–2405.
6. Sharma MR, Maitland ML, Ratain MJ. Other paradigms: better treatments are identified by better trials: the value of randomized phase II studies. *Cancer J.* 2009;15(5):426–430.
7. Gan HK, Grothey A, Pond GR, et al. Randomized phase II trials: inevitable or inadvisable? *J Clin Oncol.* 2010;28(15):2641–2647.
8. Stewart DJ. Randomized phase II trials: misleading and unreliable. *J Clin Oncol.* 2010;28(31):649–650.
9. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *J Clin Oncol.* 2008;26(8):1346–1354.
10. Korn EL, Liu PY, Lee SJ, et al. Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials. *J Clin Oncol.* 2008;26(4):527–534.
11. Temel JS, Greer JA, Muzikansky A, et al. Early palliative care for patients with metastatic non-small-cell lung cancer. *N Engl J Med.* 2010;363(8):733–742.
12. Vickers AJ, Ballen V, Scher HI. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. *Clin Cancer Res.* 2007;13(3):972–976.
13. Rubinstein L, Crowley J, Ivy P, et al. Randomized phase II designs. *Clin Cancer Res.* 2009;15(6):1883–1890.
14. Kindler HL, Friberg G, Singh DA, et al. Phase II trial of bevacizumab plus gemcitabine in patients with advanced pancreatic cancer. *J Clin Oncol.* 2005;23(31):8033–8040.
15. Xiong HQ, Rosenberg A, LoBuglio A, et al. Cetuximab, a monoclonal antibody targeting the epidermal growth factor receptor, in combination with gemcitabine for advanced pancreatic cancer: a multicenter phase II trial. *J Clin Oncol.* 2004;22(13):2610–2616.
16. Stathopoulos GP, Syrigos K, Polyzos A, et al. Front-line treatment of inoperable or metastatic pancreatic cancer with gemcitabine and capecitabine: an intergroup, multicenter, phase II study. *Ann Oncol.* 2004;15(2):224–229.
17. Philip PA, Zalupski MM, Vaitkevicius VK, et al. Phase II study of gemcitabine and cisplatin in the treatment of patients with advanced pancreatic carcinoma. *Cancer.* 2001;92(3):569–577.
18. Kindler HL. The pemetrexed/gemcitabine combination in pancreatic cancer. *Cancer.* 2002;95(4 suppl):928–932.
19. Louvet C, André T, Lledo G, et al. Gemcitabine combined with oxaliplatin in advanced pancreatic adenocarcinoma: final results of a GERCOR multicenter phase II study. *J Clin Oncol.* 2002;20(6):1512–1518.
20. Rocha Lima CM, Savarese D, Bruckner H, et al. Irinotecan plus gemcitabine induces both radiographic and CA 19-9 tumor marker responses in patients with previously untreated advanced pancreatic cancer. *J Clin Oncol.* 2002;20(5):1182–1191.
21. Berlin JD, Adak S, Vaughn DJ, et al. A phase II study of gemcitabine and 5-fluorouracil in metastatic pancreatic cancer: an Eastern Cooperative Oncology Group Study (E3296). *Oncology.* 2000;58(3):215–218.
22. Kindler HL, Niedzwiecki D, Hollis D, et al. Gemcitabine plus bevacizumab compared with gemcitabine plus placebo in patients with advanced pancreatic cancer: phase III trial of the Cancer and Leukemia Group B (CALGB 80303). *J Clin Oncol.* 2010;28(22):3617–3622.
23. Philip PA, Benedetti J, Corless CL, et al. Phase III study comparing gemcitabine plus cetuximab versus gemcitabine in patients with advanced pancreatic adenocarcinoma: Southwest Oncology Group-directed intergroup trial S0205. *J Clin Oncol.* 2010;28(22):3605–3610.

24. Herrmann R, Bodoky G, Ruhstaller T, et al. Gemcitabine plus capecitabine compared with gemcitabine alone in advanced pancreatic cancer: a randomized, multicenter, phase III trial of the Swiss Group for Clinical Cancer Research and the Central European Cooperative Oncology Group. *J Clin Oncol.* 2007;25(16):2212–2217.
25. Heinemann V, Quietzsch D, Gieseler F, et al. Randomized phase III trial of gemcitabine plus cisplatin compared with gemcitabine alone in advanced pancreatic cancer. *J Clin Oncol.* 2006;24(24):3946–3952.
26. Oettle H, Richards D, Ramanathan RK, et al. A phase III trial of pemetrexed plus gemcitabine versus gemcitabine in patients with unresectable or metastatic pancreatic cancer. *Ann Oncol.* 2005;16(10):1639–1645.
27. Louvet C, Labianca R, Hammel P, et al. Gemcitabine in combination with oxaliplatin compared with gemcitabine alone in locally advanced or metastatic pancreatic cancer: results of a GERCOR and GISCAD phase III trial. *J Clin Oncol.* 2005;23(15):3509–3516.
28. Rocha Lima CM, Green MR, Rotche R, et al. Irinotecan plus gemcitabine results in no survival advantage compared with gemcitabine monotherapy in patients with locally advanced or metastatic pancreatic cancer despite increased tumor response rate. *J Clin Oncol.* 2004;22(18):3776–3783.
29. Berlin JD, Catalano P, Thomas JP, et al. Phase III study of gemcitabine in combination with fluorouracil versus gemcitabine alone in patients with advanced pancreatic carcinoma: Eastern Cooperative Oncology Group Trial E2297. *J Clin Oncol.* 2002;20(15):3270–3275.
30. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *J Clin Oncol.* 2007;25(15):1960–1966.
31. Walter RB, Appelbaum FR, Tallman MS, et al. Shortcomings in the clinical evaluation of new drugs: acute myeloid leukemia as paradigm. *Blood.* 2010;116(14):2420–2428.
32. Tang H, Foster NR, Grothey A, et al. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol.* 2010; 28(11):1936–1941.
33. Sharma M, Karrison T, Maitland ML, et al. VEGF pathway therapy: resampling positive phase III data to assess phase II trial designs and endpoints. *J Clin Oncol.* 2010;28(suppl):15s. Abstract 2520.
34. Stewart DJ, Whitney SN, Kurzrock R. Equipoise lost: ethics, costs, and the regulation of cancer clinical research. *J Clin Oncol.* 2010;28(17): 2925–2935.
35. Flaherty KT, Puzanov I, Kim KB, et al. Inhibition of mutated, activated BRAF in metastatic melanoma. *N Engl J Med.* 2010;363(9):809–819.
36. Kwak EL, Bang YJ, Camidge DR, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med.* 2010;363(18): 1693–1703.
37. Kantarjian H, Sawyers C, Hochhaus A, et al. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N Engl J Med.* 2002;346(9):645–652.
38. Demetri GD, von Mehren M, Blanke CD, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *N Engl J Med.* 2002;347(7):472–480.
39. Kindler T, Lipka DB, Fischer T. FLT3 as a therapeutic target in AML: still challenging after all these years. *Blood.* 2010;116(24):5089–5102.
40. Ratain MJ, Karrison TG. Testing the wrong hypothesis in phase II oncology trials: there is a better alternative. *Clin Cancer Res.* 2007;13(3):781–782.
41. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 1989;10(1):1–10.
42. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45(2):228–247.
43. Karrison TG, Maitland ML, Stadler WM, et al. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst.* 2007;99(19):1455–1461.
44. Dancey JE, Dobbin KK, Groshen S, et al.; Biomarkers Task Force of the NCI Investigational Drug Steering Committee. Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. *Clin Cancer Res.* 2010;16(6):1745–1755.
45. Stadler W. Other paradigms: randomized discontinuation trial design. *Cancer J.* 2009;15(5):431–434.
46. Ratain MJ, Eisen T, Stadler WM, et al. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. *J Clin Oncol.* 2006;24(16):2505–2512.
47. Escudier B, Eisen T, Stadler WM, et al.; Target Study Group. Sorafenib in advanced clear-cell renal-cell carcinoma. *N Engl J Med.* 2007;356(2): 125–134.
48. Olanow CW, Rascol O, Hauser R, et al.; Adagio Study Investigators. A double-blind, delayed-start trial of rasagiline in Parkinson's disease. *N Engl J Med.* 2009;361(13):1268–1278.
49. D'Agostino RB. The delayed-start study design. *N Engl J Med.* 2009; 361(13):1304–1306.
50. Cheung YK, Inoue LY, Wathen JK, et al. Continuous Bayesian adaptive randomization based on event times with covariates. *Stat Med.* 2006;25(1): 55–70.
51. Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clin Trials.* 2009; 6(3):205–216.
52. Lee JJ, Xuemin Gu, Suyu Liu. Bayesian adaptive randomization designs for targeted agent development. *Clin Trials.* 2010;7(5):584–596.
53. Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics.* 1989;45(2):537–547.
54. Liu PY, Dahlberg S, Crowley J. Selection designs for pilot studies based on survival. *Biometrics.* 1993;49(2):391–398.
55. Liu PY, Dahlberg S. Design and analysis of multiarm clinical trials with survival endpoints. *Control Clin Trials.* 1995;16(2):119–130.
56. Storer BE. A sequential phase II/III trial for binary outcomes. *Stat Med.* 1990;9(3):229–235.
57. Inoue LY, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics.* 2002;58(4):823–831.
58. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints [published online ahead of print]. *Pharm Stat.* 2010.
59. Harmon A. New drugs stir debate on rules of clinical trials. *New York Times.* September 19, 2010:A1.
60. Seymour L, Ivy SP, Sargent D, et al. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin Cancer Res.* 2010;16(6): 1764–1769.
61. Cannistra SA. Phase II trials in journal of clinical oncology. *J Clin Oncol.* 2009;27(19):3073–3076.

Funding

This work was supported by a training grant from the National Institutes of Health for Clinical Therapeutics (T32GM007019 to M.R.S.).

Notes

The funders did not have any involvement in the design of the study; the collection, analysis, and interpretation of the data; the writing of the article; or the decision to submit the article for publication.

Affiliations of authors: Department of Medicine (MRS, WMS, MJR), Cancer Research Center (WMS, MJR), and Committee on Clinical Pharmacology and Pharmacogenomics (MRS, MJR), University of Chicago, Chicago, IL.